Springer Texts in Business and Economics

Thomas Cleff

Applied Statistics and Multivariate Data Analysis for Business and Economics

A Modern Approach Using R, SPSS, Stata, and Excel

Second Edition



Springer

Springer Texts in Business and Economics

Springer Texts in Business and Economics (STBE) delivers high-quality instructional content for undergraduates and graduates in all areas of Business/ Management Science and Economics. The series is comprised of self-contained books with a broad and comprehensive coverage that are suitable for class as well as for individual self-study. All texts are authored by established experts in their fields and offer a solid methodological background, often accompanied by problems and exercises.

Thomas Cleff

Applied Statistics and Multivariate Data Analysis for Business and Economics

A Modern Approach Using R, SPSS, Stata, and Excel

Second Edition

Springer

Thomas Cleff
Pforzheim Business School
Pforzheim University of Applied Sciences
Pforzheim, Baden-Württemberg, Germany

ISSN 2192-4333 ISSN 2192-4341 (electronic)
Springer Texts in Business and Economics
ISBN 978-3-031-78069-1 ISBN 978-3-031-78070-7 (eBook)
https://doi.org/10.1007/978-3-031-78070-7

1st edition: © Springer Nature Switzerland AG 2014, 2019, corrected publication 2019
2nd edition: © The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

The second edition of this textbook, *Applied Statistics and Multivariate Data Analysis for Business and Economics: A Modern Approach Using R, SPSS, Stata, and Excel*, aims to familiarize students of business and economics and all other students of social sciences and humanities, as well as practitioners in firms with the basic principles, techniques, and applications of applied statistics and applied data analysis. Drawing on practical examples from business settings, it demonstrates the techniques of statistical testing, univariate, bivariate, and multivariate statistical analyses. The textbook covers a range of subject matter, from scaling, sampling, and data preparation to advanced analytic procedures for assessing multivariate relationships. Techniques covered include univariate analyses (e.g. measures of central tendencies, frequency tables, univariate charts, dispersion parameters), bivariate analyses (e.g. contingency tables, correlation), parametric and nonparametric tests (e.g. t-tests, Wilcoxon signed-rank test, U test, H test), and multivariate analyses (e.g. analysis of variance, linear regression, logistic regression, cluster analysis, and factor analysis). In addition, the book covers issues such as time series and indices, classical measurement theory, point estimation, and interval estimation. Each chapter concludes with a set of exercises. In this way, it addresses all of the topics typically covered in university courses on statistics and advanced applied multivariate analysis.

In writing this book, I have consistently endeavoured to provide readers with an understanding of the thinking processes underlying complex methods of data analysis. I believe this approach will be particularly valuable to those who might otherwise have difficulty with the formal method of presentation used by many other textbooks in statistics. In numerous instances, I have tried to avoid unnecessary formulas, attempting instead to provide the reader with an intuitive grasp of a concept before deriving or introducing the associated mathematics. Nevertheless, a book about statistics and data analysis that omits formulas would be neither possible nor desirable. Whenever ordinary language reaches its limits, the mathematical formula has always been the best tool to express meaning. To provide further depth, I have included practice problems and solutions at the end of each chapter, which are intended to make it easier for students to pursue effective self-study.

The broad availability of statistical software now makes it possible to learn and to teach statistics in new ways. Indeed, students today have access to a range of powerful computer and online applications, from Excel to various professional statistical packages. Accordingly, this textbook does not confine itself to presenting statistical methods, but also addresses the use of programmes such as *R, SPSS, Stata, and Excel*. To aid the learning process, datasets have been made available on the springer website, along with other supplemental materials, allowing all of the examples and practice problems to be recalculated and reviewed.

I want to take this opportunity to thank all those who have collaborated in making this book possible. Well-deserved gratitude for their critical review of the manuscript and valuable suggestions goes to Uli Föhl, Wolfgang Gohout, Bernd Kuppinger, Bettina Müller, Bettina Peters, Wolfgang Schäfer (†), Lucais Sewell, and Kirsten Wüst, as well as many other unnamed individuals. Any errors or shortcomings that remain are entirely my own. Finally, this book could not have been possible without the ongoing support of my family. They deserve my very special gratitude.

Please do not hesitate to contact me directly with feedback or any suggestions you may have for improvements (thomas.cleff@hs-pforzheim.de).

Pforzheim, Germany Thomas Cleff
October 2024

Contents

1 Statistics and Empirical Research 1
1.1 Do Statistics Lie? 1
1.2 Different Types of Statistics 3
1.3 The Generation of Knowledge Through Statistics 6
1.4 The Phases of Empirical Research 8
1.4.1 From Exploration to Theory 8
1.4.2 From Theories to Models 9
1.4.3 From Models to Business Intelligence 13
References 14

2 From Disarray to Dataset 17
2.1 Get Your Software Ready 17
2.2 Data Collection 20
2.3 Level of Measurement 22
2.4 Scaling and Coding 25
2.5 Missing Values 27
2.6 Outliers and Obviously Incorrect Values 29
2.7 Chapter Exercises 29
2.8 Exercise Solutions 30
References 30

3 Univariate Data Analysis 33
3.1 First Step in Data Analysis 33
3.2 Measures of Central Tendency 40
3.2.1 Mode or Modal Value 41
3.2.2 Mean 41
3.2.3 Geometric Mean 46
3.2.4 Harmonic Mean 47
3.2.5 The Median 50
3.2.6 Quartile and Percentile 52
3.3 The Boxplot: A First Look at Distributions 54
3.4 Dispersion Parameters 56

3.4.1 Standard Deviation and Variance 57
3.4.2 The Coefficient of Variation 59
3.5 Skewness and Kurtosis 60
3.6 Robustness of Parameters 63
3.7 Measures of Concentration 64
3.8 Using the Computer to Calculate Univariate Parameters 67
3.8.1 Calculating Univariate Parameters with R Commander 67
3.8.2 Calculating Univariate Parameters with SPSS 69
3.8.3 Calculating Univariate Parameters with Stata 69
3.8.4 Calculating Univariate Parameters with Excel 71
3.9 Chapter Exercises 72
3.10 Exercise Solutions 74
References 78

4 Bivariate Association 79
4.1 Bivariate Scale Combinations 79
4.2 Association Between Two Nominal Variables 80
4.2.1 Contingency Tables 80
4.2.2 Chi-Square Calculations 82
4.2.3 The Phi Coefficient 86
4.2.4 The Contingency Coefficient 89
4.2.5 Cramer's V 91
4.2.6 Nominal Associations with SPSS 91
4.2.7 Nominal Associations with Stata 95
4.2.8 Nominal Associations with R and R Commander 96
4.2.9 Nominal Associations with Excel 99
4.3 Association Between Two Metric Variables 100
4.3.1 The Scatterplot 101
4.3.2 The Bravais–Pearson Correlation Coefficient 103
4.4 Relationships Between Ordinal Variables 107
4.4.1 Spearman's Rank Correlation Coefficient (Spearman's Rho) 109
4.4.2 Kendall's Tau (τ) 114
4.5 Measuring the Association Between Two Variables with Different Scales 119
4.5.1 Measuring the Association Between Nominal and Metric Variables 119
4.5.2 Measuring the Association Between Nominal and Ordinal Variables 122
4.5.3 Association Between Ordinal and Metric Variables 122
4.6 Calculating Correlation with a Computer 123
4.6.1 Calculating Correlation with R and R Commander 123
4.6.2 Calculating Correlation with SPSS 124

4.6.3 Calculating Correlation with Stata 125
4.6.4 Calculating Correlation with Excel 126
4.7 Spurious Correlations 127
4.7.1 Partial Correlation 130
4.7.2 Partial Correlations with R and R Commander 131
4.7.3 Partial Correlations with SPSS 131
4.7.4 Partial Correlations with Stata 133
4.7.5 Partial Correlation with Excel 133
4.8 Chapter Exercises 134
4.9 Exercise Solutions 140
References 145

5 Classical Measurement Theory 147
5.1 Sources of Sampling Errors 149
5.2 Sources of Nonsampling Errors 151
References 154

6 Calculating Probability 155
6.1 Key Terms for Calculating Probability 156
6.2 Probability Definitions 157
6.3 Foundations of Probability Calculus 161
6.3.1 Probability Tree 161
6.3.2 Combinatorics 163
6.3.3 The Inclusion–Exclusion Principle for Disjoint Events 167
6.3.4 Inclusion–Exclusion Principle for Nondisjoint Events 169
6.3.5 Conditional Probability 169
6.3.6 Independent Events and Law of Multiplication 170
6.3.7 Law of Total Probability 171
6.3.8 Bayes' Theorem 172
6.3.9 Postscript: The Monty Hall Problem 173
6.4 Chapter Exercises 176
6.5 Exercise Solutions 179
References 186

7 Random Variables and Probability Distributions 189
7.1 Discrete Distributions 191
7.1.1 Binomial Distribution 191
7.1.1.1 Calculating Binomial Distributions Using Excel 194
7.1.1.2 Calculating Binomial Distributions Using Stata 195
7.1.2 Hypergeometric Distribution 195
7.1.2.1 Calculating Hypergeometric Distributions Using Excel 199

7.1.2.2 Calculating the Hypergeometric Distribution Using Stata 199
7.1.3 The Poisson Distribution 200
7.1.3.1 Calculating the Poisson Distribution Using Excel 202
7.1.3.2 Calculating the Poisson Distribution Using Stata 202
7.2 Continuous Distributions 203
7.2.1 The Continuous Uniform Distribution 206
7.2.2 The Normal Distribution 208
7.2.2.1 Calculating the Normal Distribution Using Excel 216
7.2.2.2 Calculating the Normal Distribution Using Stata 217
7.3 Important Distributions for Testing 217
7.3.1 The Chi-Squared Distribution 218
7.3.1.1 Calculating the Chi-Squared Distribution Using Excel 219
7.3.1.2 Calculating the Chi-Squared Distribution Using Stata 219
7.3.2 The *t*-Distribution 221
7.3.2.1 Calculating the *t*-Distribution Using Excel 223
7.3.2.2 Calculating the *t*-Distribution Using Stata 224
7.3.3 The *F*-Distribution 225
7.3.3.1 Calculating the *F*-Distribution Using Excel 225
7.3.3.2 Calculating the *F*-Distribution Using Stata 225
7.4 Chapter Exercises 227
7.5 Exercise Solutions 231
References 242

8 Parameter Estimation 243
8.1 Point Estimation 244
8.2 Interval Estimation 250
8.2.1 The Confidence Interval for the Mean of a Population (μ) 250
8.2.2 Planning the Sample Size for Mean Estimation 256
8.2.3 Confidence Intervals for Proportions 258
8.2.4 Planning Sample Sizes for Proportions 261
8.2.5 The Confidence Interval for Variances 262
8.2.6 Calculating Confidence Intervals with the Computer 263

8.2.6.1 Calculating Confidence Intervals with Excel 263
8.2.6.2 Calculating Confidence Intervals with SPSS 265
8.2.6.3 Calculating Confidence Intervals with Stata 266
8.2.6.4 Calculating Confidence Intervals with R Commander 270
8.3 Chapter Exercises 271
8.4 Exercise Solutions 273
References 277

9 Hypothesis Testing 279
9.1 Fundamentals of Hypothesis Testing 279
9.2 One-Sample Tests 285
9.2.1 One-Sample *Z*-Test (When σ Is Known) 285
9.2.2 One-Sample *t*-Test (When σ Is Not Known) 287
9.2.3 Probability Value (*p*-Value) 290
9.2.4 One-Sample *t*-Test with SPSS, Stata, R, and Excel 291
9.3 Tests for Two Dependent Samples 294
9.3.1 The *t*-Test for Dependent Samples 294
9.3.1.1 The Paired *t*-Test with SPSS 299
9.3.1.2 The Paired *t*-Test with Stata 299
9.3.1.3 The Paired *t*-Test with R Commander 301
9.3.1.4 The Paired *t*-Test with Excel 302
9.3.2 The Wilcoxon Signed-Rank Test 303
9.3.2.1 The Wilcoxon Signed-Rank Test with SPSS 308
9.3.2.2 The Wilcoxon Signed-Rank Test with Stata 310
9.3.2.3 The Wilcoxon Signed-Rank Test with R Commander 311
9.3.2.4 The Wilcoxon Signed-Rank Test with Excel 311
9.4 Tests for Two Independent Samples 312
9.4.1 The *t*-Test of Two Independent Samples 312
9.4.1.1 The *t*-Test for Two Independent Samples with SPSS 315
9.4.1.2 The *t*-Test for Two Independent Samples with Stata 316
9.4.1.3 The *t*-Test for Two Independent Samples with R Commander 317
9.4.1.4 The *t*-Test for Two Independent Samples with Excel 319

9.4.2 The Mann–Whitney *U* Test (Wilcoxon Rank-Sum Test) 320
9.4.2.1 The Mann–Whitney *U* Test with SPSS 324
9.4.2.2 The Mann–Whitney *U* Test with Stata 324
9.4.2.3 The Mann–Whitney *U* Test with R Commander 326
9.5 Tests for *k* Independent Samples 327
9.5.1 Analysis of Variance (ANOVA) 327
9.5.1.1 One-Way Analysis of Variance (ANOVA) 327
9.5.1.2 Two-Way Analysis of Variance (ANOVA) 331
9.5.1.3 Analysis of Covariance (ANCOVA) 335
9.5.1.4 ANOVA/ANCOVA with SPSS 338
9.5.1.5 ANOVA/ANCOVA with Stata 338
9.5.1.6 ANOVA/ANCOVA with R Commander 338
9.5.1.7 ANOVA with Excel 339
9.5.2 Kruskal–Wallis Test (*H* Test) 341
9.5.2.1 Kruskal–Wallis *H* Test with SPSS 346
9.5.2.2 Kruskal–Wallis *H* Test with Stata 347
9.5.2.3 Kruskal–Wallis *H* Test with R Commander 347
9.6 Other Tests 348
9.6.1 Chi-Square Test of Independence 348
9.6.1.1 Chi-Square Test of Independence with SPSS 351
9.6.1.2 Chi-Square Test of Independence with Stata 351
9.6.1.3 Chi-Square Test of Independence with R Commander 353
9.6.1.4 Chi-Square Test of Independence with Excel 353
9.6.2 Tests for Normal Distribution 354
9.6.2.1 Testing for Normal Distribution with SPSS 355
9.6.2.2 Testing for Normal Distribution with Stata 357
9.6.2.3 Testing for Normal Distribution with R Commander 358
9.7 Chapter Exercises 359
9.8 Exercise Solutions 368
References 381

10 Regression Analysis 383
10.1 First Steps in Regression Analysis 383
10.2 Coefficients of Bivariate Regression 385
10.3 Multiple Regression Coefficients 389
10.4 The Goodness of Fit of Regression Lines 391
10.5 Regression Calculations with the Computer 393
10.5.1 Regression Calculations with SPSS 394
10.5.2 Regression Calculations with Stata 394
10.5.3 Regression Calculations with R Commander 395
10.5.4 Regression Calculations with Excel 396
10.6 Goodness of Fit of Multiple Regressions 397
10.7 Regression with an Independent Dummy Variable 398
10.8 Leverage Effects of Data Points 400
10.9 Non-linear Regressions 401
10.9.1 Approaches to Regression Diagnostics 404
10.9.2 Chapter Exercises 411
10.9.3 Exercise Solutions 415
References 419

11 Logistic Regression 421
11.1 Odds, Log-Odds, and Maximum Likelihood Estimation 425
11.2 Estimating a Logistic Regression 429
11.3 Testing for Significance 432
11.3.1 Overall Model Significance (Likelihood Ratio Test) 432
11.3.2 Significance of Individual Coefficients (Wald Test) 433
11.4 Pseudo-R^2 434
11.5 Goodness of Fit of a Logistic Regression 436
11.6 Classification Performance 438
11.7 Impact of Heteroskedasticity 440
11.8 Impact of Multicollinearity 440
11.9 Another Example: Logistic Regression with the Computer 441
11.9.1 Logistic Regression Calculations with SPSS 444
11.9.2 Logistic Regression Calculations with Stata 444
11.9.3 Logistic Regression Calculations with R Commander 446
11.10 Chapter Exercises 449
11.11 Exercise Solutions 450
References 452

12 Time Series and Indices 453
12.1 Price Indices 454
12.2 Quantity Indices 461
12.3 Value Indices (Sales Indices) 462
12.4 Deflating Time Series by Price Indices 463
12.5 Shifting Bases and Chaining Indices 464
12.6 Chapter Exercises 466
12.7 Exercise Solutions 467
References 469

13 Cluster Analysis 471
13.1 Hierarchical Cluster Analysis 472
13.2 K-Means Cluster Analysis 485
13.3 Cluster Analysis Using SPSS, Stata, and R 488
13.3.1 Cluster Analysis with SPSS 488
13.3.2 Cluster Analysis with Stata 489
13.3.3 Cluster Analysis with R Commander 489
13.4 Chapter Exercises 490
13.5 Exercise Solutions 494
References 498

14 Factor Analysis 501
14.1 Factor Analysis: Foundations, Methods, Interpretations 501
14.2 Factor Analysis with SPSS 510
14.3 Factor Analysis with Stata 510
14.4 Factor Analysis with R Commander 510
14.5 Chapter Exercises 510
14.6 Exercise Solutions 515
References 515

List of Formulas 517

Appendix A: The Standard Normal Distribution 527

Appendix B: The Chi-Squared Distribution 529

Appendix C: The Student's *t*-Distribution 533

Appendix D: Critical Values for the Wilcoxon Signed-Rank Test 535

Appendix E: Getting Started with SPSS 537

Appendix F: Getting Started with Stata 541

Appendix G: Getting Started with R and the R Commander 545

Index 549

List of Figures

Fig. 1.1 Data beget information, which in turn begets knowledge 5
Fig. 1.2 Techniques for multivariate analysis 6
Fig. 1.3 Price and demand function for sensitive toothpaste 7
Fig. 1.4 Phases of empirical research 8
Fig. 1.5 A systematic overview of model variants 10
Fig. 1.6 What is certain? © Marco Padberg 12
Fig. 1.7 Intelligence cycle. *Source* Own graphic, adapted from Harkleroad (1996, p. 45) 14
Fig. 2.1 Retail questionnaire 22
Fig. 2.2 Statistical units/traits/trait values/level of measurement 23
Fig. 2.3 Label book 26
Fig. 3.1 Survey data entered in the data editor 34
Fig. 3.2 Frequency table for selection ratings 35
Fig. 3.3 Bar chart/frequency distribution for the selection variable 36
Fig. 3.4 Distribution function for the selection variable 36
Fig. 3.5 Different representations of the same data (1) 37
Fig. 3.6 Different representations of the same data (2) 37
Fig. 3.7 Using a histogram to classify data 38
Fig. 3.8 Distorting interval selection with a distribution function 39
Fig. 3.9 Grade averages for two final exams 40
Fig. 3.10 Mean expressed as a balanced scale 42
Fig. 3.11 Mean or trimmed mean using the zoo example 42
Fig. 3.12 Calculating the mean from classed data 44
Fig. 3.13 An example of geometric mean 46
Fig. 3.14 The median: the central value of unclassed data 51
Fig. 3.15 The median: the middle value of classed data 52
Fig. 3.16 Calculating quantiles with five weights 53
Fig. 3.17 Boxplot of weekly sales 55
Fig. 3.18 Interpretation of different boxplot types 56
Fig. 3.19 Coefficient of variation 60

Fig. 3.20 Skewness. The numbers in the boxes represent ages. The mean is indicated by the arrow. Like a balance scale, the deviations to the left and right of the mean are in equilibrium 61
Fig. 3.21 Third central moment. The numbers in the boxes represent ages. The mean is indicated by the triangle. Like a balance scale, the cubed deviations to the left and right of the mean are in disequilibrium 62
Fig. 3.22 Kurtosis distributions 63
Fig. 3.23 Robustness of parameters. Many studies use mean, variance, skewness, and kurtosis with ordinal scales as well. Section 2.3 described the conditions necessary for this to be possible 64
Fig. 3.24 Measure of concentration 66
Fig. 3.25 Lorenz curve 66
Fig. 3.26 Univariate parameters with R and R Commander (1) 68
Fig. 3.27 Univariate parameters with R and R Commander (2) 68
Fig. 3.28 Univariate parameters with SPSS 70
Fig. 3.29 Univariate parameters with Stata 70
Fig. 3.30 Univariate parameters with Excel 71
Fig. 3.31 Market research study 72
Fig. 3.32 Bar graph and histogram 77
Fig. 4.1 Contingency table (crosstab) 81
Fig. 4.2 Contingency tables (crosstabs) (1st) 82
Fig. 4.3 Contingency table (crosstab) (2nd) 83
Fig. 4.4 Calculation of expected counts in contingency tables 85
Fig. 4.5 Chi-square values based on different sets of observations 87
Fig. 4.6 The phi coefficient in tables with various numbers of rows and columns 88
Fig. 4.7 The contingency coefficient in tables with various numbers of rows and columns 90
Fig. 4.8 Crosstabs and nominal associations with SPSS (Titanic) 93
Fig. 4.9 From raw data to computer-calculated crosstab (Titanic) 94
Fig. 4.10 Computer printout of chi-square and nominal measures of association 95
Fig. 4.11 Crosstabs and nominal measures of association with Stata (Titanic) 95
Fig. 4.12 Convert variables to factors with R Commander (Titanic) 97
Fig. 4.13 Crosstabs and nominal measures of association with R Commander (Titanic) 98
Fig. 4.14 Crosstabs and nominal measures of association using R syntax (Titanic) 99
Fig. 4.15 Crosstabs and nominal measures of association with Excel (Titanic) 100
Fig. 4.16 The scatterplot 101

Fig. 4.17 Aspects of association expressed by the scatterplot 102
Fig. 4.18 Different representations of the same data (3) 103
Fig. 4.19 Relationship of heights in married couples 104
Fig. 4.20 Four-quadrant system 105
Fig. 4.21 Pearson's correlation coefficient with outliers 107
Fig. 4.22 Wine bottle design survey 108
Fig. 4.23 Non-linear relationship between two variables 108
Fig. 4.24 Data for survey on wine bottle design 110
Fig. 4.25 Rankings from the wine bottle design survey 111
Fig. 4.26 Kendall's τ and a perfect positive monotonic association 114
Fig. 4.27 Kendall's τ for a non-existent monotonic association 116
Fig. 4.28 Kendall's τ for tied ranks 117
Fig. 4.29 Deriving Kendall's τ_b from a contingency table 118
Fig. 4.30 Point-biserial correlation 120
Fig. 4.31 Association between two ordinal and metric variables 123
Fig. 4.32 Calculating correlation with R and R Commander 124
Fig. 4.33 Calculating correlation with SPSS 125
Fig. 4.34 Calculating correlation with Stata (Kendall's τ) 126
Fig. 4.35 Spearman's correlation with Excel 128
Fig. 4.36 Reasons for spurious correlations 129
Fig. 4.37 High-octane fuel and market share: an example of spurious correlation 131
Fig. 4.38 Partial correlation with R and R Commander (high-octane petrol) 132
Fig. 4.39 Partial correlation with SPSS (high-octane petrol) 132
Fig. 4.40 Partial correlation with Stata (high-octane petrol) 133
Fig. 4.41 Partial correlation with Excel (high-octane petrol) 134
Fig. 5.1 Empirical sampling methods 151
Fig. 5.2 Distortions caused by nonsampling errors. *Source* Based on Malhotra et al. (2017, p. 410). Figure compiled by the author 152
Fig. 6.1 Sample space and combined events when tossing a die 156
Fig. 6.2 Intersection of events and complementary events 157
Fig. 6.3 Event tree for a sequence of three coin tosses 158
Fig. 6.4 Relative frequency for a coin toss 160
Fig. 6.5 Approaches to probability theory 161
Fig. 6.6 Probability tree for a sequence of three coin tosses 162
Fig. 6.7 Combination and variation. *Source* Wewel and Blatter (2019) figure modified slightly 164
Fig. 6.8 Event tree for winner combinations and variations with four players and two games 166
Fig. 6.9 Event tree for winning variations without repetition for four players and two rounds 166

Fig. 6.10 Deciding between permutation, combination, and variation. *Source* Bourier (2018, p. 80). Compiled by the author 167
Fig. 6.11 Probability tree of the Monty Hall problem. This probability tree assumes that the host does not open the door with the main prize or the first door selected. It also assumes that contestants can choose any door. That is, even if contestants pick a door other than #1, the probability of winning stays the same. The winning scenarios are in grey 175
Fig. 6.12 Probability tree for statistics exam and holiday 183
Fig. 6.13 Probability tree for test market 184
Fig. 6.14 Probability tree for defective products 185
Fig. 6.15 Paint shop 186
Fig. 7.1 Probability function and distribution function of a die roll experiment 190
Fig. 7.2 Binomial distribution 193
Fig. 7.3 Binomial distribution of faces $x = 6$ with n throws of an unloaded die 194
Fig. 7.4 Calculating binomial distributions with Excel 195
Fig. 7.5 Calculating binomial distributions using Stata 196
Fig. 7.6 Hypergeometric distribution 197
Fig. 7.7 Calculating hypergeometric distributions with Excel 199
Fig. 7.8 Calculating hypergeometric distributions using Stata 200
Fig. 7.9 Poisson distribution 201
Fig. 7.10 Calculating the Poisson distribution with Excel 203
Fig. 7.11 Calculating the Poisson distribution using Stata 203
Fig. 7.12 Density functions 204
Fig. 7.13 Uniform distribution 207
Fig. 7.14 Production times 207
Fig. 7.15 Ideal density of a normal distribution 209
Fig. 7.16 Positions of normal distributions 210
Fig. 7.17 Different spreads of normal distributions 210
Fig. 7.18 Shelf life of yoghurt (1) 212
Fig. 7.19 Shelf life of yoghurt (2) 213
Fig. 7.20 Calculating the probability of a z-transformed random variable 214
Fig. 7.21 Calculating probabilities using the standard normal distribution 215
Fig. 7.22 Calculating the normal distribution using Excel 216
Fig. 7.23 Calculating the normal distribution using Stata 217
Fig. 7.24 Density function of a chi-squared distribution with different degrees of freedom (df) 218
Fig. 7.25 Calculating the chi-squared distribution with Excel 220
Fig. 7.26 Calculating the chi-squared distribution with Stata 220

Fig. 7.27 t-distribution with varying degrees of freedom 222
Fig. 7.28 Calculating the t-distribution using Excel 223
Fig. 7.29 Calculating the t-distribution using Stata 224
Fig. 7.30 F-distributions 226
Fig. 7.31 Calculating the F-distribution using Excel 226
Fig. 7.32 Calculating the F-distribution using Stata 227
Fig. 8.1 Distribution of sample means in a normally distributed population 245
Fig. 8.2 Generating samples using Excel: 1000 samples with a size of $n = 5$ from a population with a distribution of $N(\mu = 35; \sigma = 10)$ 246
Fig. 8.3 Distribution of mean with $n = 2$ throws of an unloaded die 247
Fig. 8.4 Distribution of the mean with $n = 4$ throws of an unloaded die 248
Fig. 8.5 Sample mean distribution of a bimodal and a left-skewed population for 30,000 samples of sizes $n = 2$ and $n = 5$ 249
Fig. 8.6 Confidence interval in the price example 252
Fig. 8.7 Calculating confidence intervals for means 253
Fig. 8.8 Length of a two-sided confidence interval for means 257
Fig. 8.9 Length of a one-sided confidence interval up to a restricted limit 258
Fig. 8.10 Calculating confidence intervals for proportions 260
Fig. 8.11 Length of a two-sided confidence interval for a proportion 261
Fig. 8.12 One-sided and two-sided confidence intervals for means with Excel 264
Fig. 8.13 One-side and two-sided confidence intervals for proportions with Excel 265
Fig. 8.14 One-side and two-sided confidence intervals for variance with Excel 266
Fig. 8.15 One-sided and two-sided confidence intervals with SPSS 267
Fig. 8.16 Confidence interval calculation using the Stata CI Calculator 268
Fig. 8.17 One-sided and two-sided confidence intervals for means with Stata 269
Fig. 8.18 One-sided and two-sided confidence intervals for a proportion value with Stata 270
Fig. 8.19 One-sided and two-sided confidence intervals for means with R 271
Fig. 9.1 Probabilities of error for hypotheses testing 281
Fig. 9.2 Error probabilities for diagnosing a disease 282
Fig. 9.3 The data structure of independent and dependent samples 283
Fig. 9.4 Tests for comparing the parameters of central tendency 284
Fig. 9.5 Rejection regions for H_0 287
Fig. 9.6 The one-sample Z-test and the one-sample t-test 288

Fig. 9.7 The one-sample *t*-test with SPSS 292
Fig. 9.8 The one-sample *t*-test with Stata 293
Fig. 9.9 The one-sample *t*-test with R Commander 294
Fig. 9.10 The one-sample *t*-test with Excel 295
Fig. 9.11 Prices of two coffee brands in 32 test markets 297
Fig. 9.12 The paired *t*-test with SPSS 300
Fig. 9.13 The paired *t*-test with Stata 301
Fig. 9.14 The paired *t*-test with R Commander 302
Fig. 9.15 The paired *t*-test with Excel 304
Fig. 9.16 Data for the Wilcoxon signed-rank test 305
Fig. 9.17 Rejection area of the Wilcoxon signed-rank test 309
Fig. 9.18 The Wilcoxon signed-rank test with SPSS 309
Fig. 9.19 The Wilcoxon signed-rank test with Stata 310
Fig. 9.20 The Wilcoxon signed-rank test with R Commander 311
Fig. 9.21 The *t*-test for two independent samples with SPSS 315
Fig. 9.22 The *t*-test for two independent samples with Stata 317
Fig. 9.23 The *t*-test for two independent samples with R Commander 318
Fig. 9.24 Testing for equality of variance with Excel 319
Fig. 9.25 The *t*-test for two independent samples with Excel 320
Fig. 9.26 Mann–Whitney *U* test 321
Fig. 9.27 The Mann–Whitney *U* test in SPSS 325
Fig. 9.28 The Mann–Whitney *U* test with Stata 325
Fig. 9.29 The Mann–Whitney *U* test with R Commander 326
Fig. 9.30 Overview of ANOVA 328
Fig. 9.31 ANOVA descriptive statistics 328
Fig. 9.32 Graphic visualization of a one-way ANOVA 329
Fig. 9.33 ANOVA tests of between-subjects effects (SPSS) 330
Fig. 9.34 ANOVA tests of between-subjects effects and descriptive statistics 332
Fig. 9.35 Interaction effects with multiple-factor ANOVA 333
Fig. 9.36 Estimated marginal means of unit sales 334
Fig. 9.37 Multiple comparisons with Scheffé's method 335
Fig. 9.38 ANCOVA tests of between-subjects effects 336
Fig. 9.39 Estimated marginal means for sales (ANCOVA) 337
Fig. 9.40 ANOVA/ANCOVA with SPSS 339
Fig. 9.41 Analysis of variance (ANOVA) with Stata 340
Fig. 9.42 Analysis of variance (ANOVA) with R Commander 340
Fig. 9.43 Analysis of variance in Excel 341
Fig. 9.44 Kruskal–Wallis test (*H* test) 342
Fig. 9.45 Kruskal–Wallis *H* test with SPSS 346
Fig. 9.46 Kruskal–Wallis *H* test with Stata 347
Fig. 9.47 Kruskal–Wallis *H* test with R Commander 348
Fig. 9.48 Nominal associations and chi-square test of independence 349

Fig. 9.49 Nominal associations and chi-square test of independence with SPSS 352
Fig. 9.50 Nominal associations and chi-square test of independence with Stata 353
Fig. 9.51 Nominal associations and chi-square test of independence with Excel 354
Fig. 9.52 Two histograms and their normal distribution curves 356
Fig. 9.53 Testing for normal distribution with SPSS 357
Fig. 9.54 Testing for normal distribution with R Commander 358
Fig. 9.55 Questionnaire for owners of a particular car 360
Fig. 9.56 Effect of three advertising strategies 362
Fig. 9.57 Effect of two advertising strategies 362
Fig. 9.58 Results of a market research study 363
Fig. 9.59 Product preference 365
Fig. 9.60 Price preference 1 366
Fig. 9.61 Price preference 2 366
Fig. 9.62 One-sample *t*-test 367
Fig. 9.63 ANOVA for Solution 1 (SPSS) 376
Fig. 9.64 ANOVA of Solution 2 (SPSS) 378
Fig. 9.65 ANOVA of Solution 3 (SPSS) 380
Fig. 10.1 Demand forecast using equivalence 384
Fig. 10.2 Demand forecast using image size 385
Fig. 10.3 Calculating residuals 387
Fig. 10.4 Lines of best fit with a minimum sum of deviations 388
Fig. 10.5 The concept of multivariate analysis 392
Fig. 10.6 Output from the regression function (for SPSS) 394
Fig. 10.7 Regression with SPSS 395
Fig. 10.8 Regression with Stata 395
Fig. 10.9 Regression with R commander 396
Fig. 10.10 Regression with excel 397
Fig. 10.11 Regression output with dummy variables 399
Fig. 10.12 The effects of dummy variables shown graphically 400
Fig. 10.13 Leverage effect 401
Fig. 10.14 Variables with non-linear distributions 403
Fig. 10.15 Regression with non-linear variables (1) 404
Fig. 10.16 Regression with non-linear variables (2) 405
Fig. 10.17 Autocorrelated and non-autocorrelated distributions of error terms 406
Fig. 10.18 Homoscedasticity and heteroscedasticity 407
Fig. 10.19 Solution for perfect multicollinearity 408
Fig. 10.20 Solution for imperfect multicollinearity 409
Fig. 10.21 Regression results (1) 412
Fig. 10.22 Regression results (2) 413
Fig. 10.23 Regression toothpaste 414
Fig. 10.24 Regression results burger slim 416

Fig. 10.25 Scatterplot 417
Fig. 11.1 Number of damaged O-rings for flights where anomalies occurred during previous shuttle flights. *Source* Own graphic, data from Dalal et al. (1989, p. 949) 423
Fig. 11.2 Number of damaged O-rings during previous shuttle flights. *Source* Own graphic, data from Dalal et al. (1989, p. 949) 423
Fig. 11.3 Probability of a damaged O-ring during previous shuttle flights. *Source* Own graphic, data from Dalal et al. (1989, p. 949) 424
Fig. 11.4 Odds and log-odds. *Source* Own graphic, data from Dalal et al. (1989, p. 949) 426
Fig. 11.5 Different behaviour of logistic functions. *Source* Own graphic 428
Fig. 11.6 Logistic regression and the shape of the S-curve regarding O-ring failures. *Source* Own graphic, data from Dalal et al. (1989, p. 949) 429
Fig. 11.7 Logistic regression results. *Source* Own graphic 430
Fig. 11.8 Hosmer–Lemeshow test and classification table. *Source* Own graphic 437
Fig. 11.9 Logistic regression results for online retail of dresses. *Source* Own graphic 442
Fig. 11.10 Logistic regression with SPSS. *Source* Own graphic 445
Fig. 11.11 Logistic regression with Stata. *Source* Own graphic 447
Fig. 11.12 Logistic regression with R Commander. *Source* Own graphic 448
Fig. 11.13 Logistic regression Exercise 2. *Source* Own graphic 451
Fig. 12.1 Diesel fuel prices by year, 2001–2007 454
Fig. 12.2 Fuel prices over time 455
Fig. 13.1 Beer dataset. *Source* Bühl (2019, p. 636) 473
Fig. 13.2 Distance calculation 1 474
Fig. 13.3 Distance calculation 2 475
Fig. 13.4 Distance and similarity measures 476
Fig. 13.5 Distance matrix (squared Euclidean distance) 477
Fig. 13.6 Sequence of steps in the linkage process 478
Fig. 13.7 Agglomeration schedule 479
Fig. 13.8 Linkage methods 480
Fig. 13.9 Dendrogram 482
Fig. 13.10 Scree plot identifying heterogeneity jumps 483
Fig. 13.11 *F*-value assessments for cluster solutions 2 to 5 483
Fig. 13.12 Cluster solution and discriminant analysis 484
Fig. 13.13 Cluster interpretations 485
Fig. 13.14 Test of the three-cluster solution with two ANOVAs 486
Fig. 13.15 Initial partition for k-means clustering 487
Fig. 13.16 Hierarchical cluster analysis with SPSS 489

Fig. 13.17 K-means cluster analysis with SPSS 490
Fig. 13.18 Cluster analysis with Stata 491
Fig. 13.19 Hierarchical cluster analysis with R Commander 492
Fig. 13.20 K-means cluster analysis with R Commander 493
Fig. 13.21 Hierarchical cluster analysis. *Source* Bühl (2019, p. 636) 494
Fig. 13.22 Dendrogram 495
Fig. 13.23 Cluster memberships 496
Fig. 13.24 Final cluster centres and cluster memberships 496
Fig. 13.25 Cluster analysis (1) 497
Fig. 13.26 Cluster analysis (2) 498
Fig. 14.1 Toothpaste attributes 502
Fig. 14.2 Correlation matrix of the toothpaste attributes 502
Fig. 14.3 Correlation matrix check 503
Fig. 14.4 Eigenvalues and stated total variance for toothpaste attributes 505
Fig. 14.5 Reproduced correlations and residuals 505
Fig. 14.6 Scree plot of the desirable toothpaste attributes 506
Fig. 14.7 Unrotated and rotated factor matrix for toothpaste attributes 507
Fig. 14.8 Varimax rotation for toothpaste attributes 508
Fig. 14.9 Factor score coefficient matrix 509
Fig. 14.10 Factor analysis with SPSS 511
Fig. 14.11 Factor analysis with Stata 512
Fig. 14.12 Factor analysis with R Commander 514
Fig. E.1 SPSS data viewer and SPSS menu bar 539
Fig. E.2 SPSS variable view 539
Fig. F.1 The Stata window 543
Fig. F.2 The Stata tool bar 543
Fig. G.1 The R console 546
Fig. G.2 The R commander window. *Source* https://cran.r-project.org/doc/contrib/Karp-Rcommander-intro.pdf, p. 5 547

List of Tables

Table 2.1 External data sources at international institutions 21
Table 3.1 Example of mean calculation from classed data 44
Table 3.2 Harmonic mean 48
Table 3.3 Share of sales by age class for diaper users 50
Table 4.1 Scale combinations and their measures of association 80
Table 6.1 Toss probabilities with two loaded dice 168
Table 6.2 Birth-weight study at Baystate Medical Center 169
Table 11.1 Classification table of a binary regression model 438
Table 12.1 Average prices for diesel and petrol in Germany 455
Table 12.2 Sample salary trends for two companies 463
Table 12.3 Chain indices for forward and backward extrapolations 465
Table 14.1 Measure of sampling adequacy (MSA) score intervals 503



Statistics and Empirical Research

1

1.1 Do Statistics Lie?

I don't trust any statistics I haven't falsified myself.

Statistics can be made to prove anything.

One often hears statements such as these when challenging the figures used by an opponent. Benjamin Disraeli, for example, is famously reputed to have declared, "There are three types of lies: lies, damned lies, and statistics". This oft-quoted assertion implies that statistics and statistical methods represent a particularly underhanded form of deception. Indeed, individuals who mistrust statistics often find confirmation for their scepticism when two different statistical assessments of the same phenomenon arrive at diametrically opposed conclusions. Yet if statistics can invariably be manipulated to support one-sided arguments, what purpose do they serve?

Although the disparaging quotes cited above may often be greeted with a nod, grin, or even wholehearted approval, statistics remain an indispensable tool for substantiating argumentative claims. Open a newspaper any day of the week, and you will come across tables, diagrams, and figures. Not a month passes without great fanfare over the latest economic forecasts, survey results, and consumer confidence data. And, of course, innumerable investors rely on the market forecasts issued by financial analysts when making investment decisions.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*,
Springer Texts in Business and Economics,
https://doi.org/10.1007/978-3-031-78070-7_1

We are thus caught in the middle of a seeming contradiction. Why do statistics have such a bad reputation even as they exude an irresistible magic—that is, the promise of precise figures, of a neatly quantifiable world? How can statistics be a lie and simultaneously the foundation upon which individuals and companies plan their futures? Swoboda (1971, p. 16) has identified two reasons for this ambivalence with regard to statistical procedures:

- First, there is a *lack of knowledge* concerning the role, methods, and limits of statistics.
- Second, many figures which are regarded as statistics are in fact *pseudo-statistics*.

The first point in particular has become increasingly relevant since the 1970s. In the era of the computer, anyone who has a command of basic arithmetic might feel capable of conducting statistical analysis, as off-the-shelf software programmes allow one to easily produce statistical tables, graphics, or regressions. Yet when laymen are entrusted with statistical tasks, basic methodological principles are often violated, and information may be intentionally or unintentionally displayed in an incomplete fashion. Furthermore, it frequently occurs that carefully generated statistics are interpreted or cited incorrectly by readers. Even when statistics are carefully prepared, they are often interpreted incorrectly or reported on erroneously. Yet naïve readers are not the only ones who fall victim to statistical fallacies. In scientific articles, one also regularly encounters what Swoboda has termed *pseudo-statistics*, i.e. statistics based on incorrect methods or even invented from whole cloth. The intentional or unintentional misapplication of statistical methods and the intentional or unintentional misinterpretation of their results are the real reasons why people distrust statistics. Fallacious conclusions and errors are as contagious as chicken pox, and spread accordingly, write Dubbern and Beck-Bornholdt (2007, p. 17). Those who have survived an infection are often inoculated against a new one, and those who later recognize an error do not so easily make one again (Dubbern & Beck-Bornholdt, 2007, p. 17). Herein lies the purpose of this book. In addition to teaching readers about the methods of statistics as lucidly as possible, it seeks to vaccinate them against fallacious conclusions and misuse.

Krämer (2015) distinguishes between *false statistics* as follows: "Some statistics are intentionally manipulated, while others are only selected improperly. In some cases, the numbers themselves are incorrect; in others they are merely presented in a misleading fashion. In any event, we regularly find apples and oranges cast together, questions posed in a suggestive manner, trends carelessly carried forward, rates or averages calculated improperly, probabilities abused, and samples distorted". In this book, we will examine numerous examples of false interpretations or attempts to manipulate. In this way, the goal of this book is clear. In a world in which data, figures, trends, and statistics constantly surround us, it is imperative to understand and be capable of using quantitative methods. Indeed, this was clear even to the German poet *Johann Wolfgang von Goethe*, who famously said in a conversation with Eckermann (31st of January 1830), "that the world is governed

by numbers; this I know, that from numbers we can find out whether it is well or ill governed". Statistical models and methods are one of the most important tools in business and economic analyses, decision-making, and business planning. Against this backdrop, the aim of this book is not just to present the most important statistical methods and their applications, but also to sharpen the reader's ability to recognize sources of error and attempts to manipulate.

You may have thought previously that common sense is sufficient for using statistics and that mathematics or statistical models play a secondary role. Yet no one who has taken a formal course in statistics would endorse this opinion. Naturally, a textbook such as this one cannot avoid some recourse to formulas. And how could it? Qualitative descriptions quickly exhaust their usefulness, even in everyday settings. When a professor is asked about the failure rate on a statistics test, no student would be satisfied with the answer *not too bad*. A quantitative answer—such as 10%—is expected, and such an answer requires a calculation—in other words, a formula.

Consequently, the formal presentation of mathematical methods and means cannot be entirely neglected in this book. Nevertheless, any diligent reader with a mastery of basis analytical principles will be able to understand the material presented herein.

1.2 Different Types of Statistics

What are the characteristics of statistical methods that avoid sources of error or attempts to manipulate? To answer this question, we first need to understand the purpose of statistics.

Historically, statistical methods were used long before the birth of Christ. In the sixth century BC, the constitution enacted by Servius Tullius provided for a periodic census of all citizens. Many readers are likely familiar with the following story: "In those days Caesar Augustus issued a decree that a census should be taken of the entire Roman world. This was the first census that took place while Quirinius was governor of Syria. And everyone went to his own town to register"[1] (Luke 2.1-5).

As this Biblical passage demonstrates, politicians have long had an interest in assessing the wealth of the populace—yet not for altruistic reasons, but rather for taxation purposes. Data were collected about the populace so that the governing elite had access to information about the lands under their control. The effort to gather data about a country represents a form of statistics.

Until the beginning of the twentieth century, all statistical analyses took the form of a *full survey* in the sense that an attempt was made to literally count every

[1] In 6/7 A.D., Judea (along with Edom and Samaria) became Roman protectorates. This passage probably refers to the census that was instituted under Quirinius, when all residents of the country and their property were registered for the purpose of tax collection. It could be, however, that the passage is referring to an initial census undertaken in 8/7 B.C.

person, animal, and object. It was during this era that the field of descriptive statistics emerged. The term *descriptive statistics* refers to all techniques used to obtain information based on the description of data from a population. The calculations of figures and parameters as well as the generation of graphics and tables are just some of the methods and techniques used in descriptive statistics.

It was not until the beginning of the twentieth century that the now common form of *inductive statistics* was developed in which one attempts to draw conclusions about a total population based on a sample. Key figures in this development were Jacob Bernoulli (1654–1705), Abraham de Moivre (1667–1754), Thomas Bayes (1702–1761), Pierre-Simon Laplace (1749–1827), Carl Friedrich Gauss (1777–1855), Pafnuti Lwowitsch Chebyshev (1821–1894), Francis Galton (1822–1911), Ronald A. Fisher (1890–1962), and William Sealy Gosset (1876–1937). A large number of inductive techniques can be attributed to the aforementioned statisticians. Thanks to their work, we no longer have to count and measure each individual within a population, but can instead conduct a smaller, more manageable survey. It would be prohibitively expensive, for example, for a firm to ask all potential customers how a new product should be designed. For this reason, firms instead attempt to query a representative sample of potential customers. Similarly, election researchers can hardly survey the opinions of all voters. In this and many other cases, the best approach is not to attempt a complete survey of an entire population, but instead to investigate a representative sample.

When it comes to the assessment of the gathered data, this means that the knowledge that is derived no longer stems from a full survey, but rather from a sample. The conclusions that are drawn must therefore be assigned a certain level of uncertainty, which can be statistically defined. This uncertainty is the price paid for the simplifying approach of inductive statistics.

Descriptive and inductive statistics are a scientific discipline used in business, economics, the natural sciences, humanities, and the social sciences. It is a discipline that encompasses methods for the description and analysis of mass phenomena with the aid of numbers and data. The analytical goal is to draw conclusions concerning the properties of the investigated objects on the basis of a full survey or partial sample. The discipline of statistics is an assembly of methods that allows us make *reasonable* decisions in the face of uncertainty. For this reason, statistics are a key foundation of decision theory.

The two main purposes of statistics are thus clearly evident: descriptive statistics aim to portray data in a purposeful, summarized fashion, and, in this way, to transform data into information. When this information is analysed using the assessment techniques of inductive statistics, *generalizable knowledge* is generated that can be used to inform political or strategic decisions. Figure 1.1 illustrates the relationship between data, information, and knowledge.

In addition, the statistical methods can also be distinguished regarding the number of analysed variables. If only one characteristic, e.g. age, is statistically analysed, this is commonly referred to as *univariate analysis*. The corresponding statistical methods are presented in Chap. 3. By contrast, when researchers analyse the relationship between two variables—for example, the relationship between

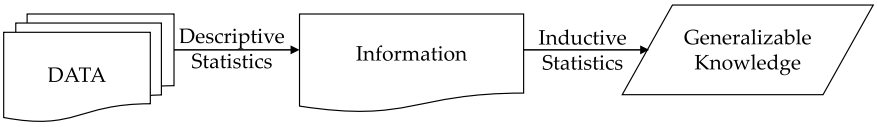


Fig. 1.1 Data beget information, which in turn begets knowledge

gender and income—this is called *bivariate analysis* (see Chap. 4). With relationships between more than two variables, one speaks of *multivariate analysis*. Let us imagine a market research study in which researchers have determined the following information for a five-year period:

- product unit sales;
- the price of the product under review and the prices of competing products, all of which remained constant;
- the shelf position of the product under review and the shelf positions of competing products, all of which remained constant;
- neither the product manufacturer nor its competitors ran any advertising.

If the product manufacturer had signed off on an ad campaign at some point during the five-year period, then bivariate analysis—in this case, the *t*-test for independent samples—would be pretty good at assessing the advertising effect. All we need to do is compare average sales before and after the ad was released. This is only possible because product prices and shelf locations remained constant. But when do market conditions ever remain constant? Don't retailers usually adjust prices when launching ad campaigns? And don't they almost always respond to the price cuts and ads of their competitors?

This example shows that under real-world conditions, changes can rarely be reduced to a single cause. In most cases, there is an assortment of influences whose combined effects and interactions need to be investigated. In this book, we will learn about several techniques for analysing more than two variables at once. Such multivariate techniques can be divided into two groups (see Fig. 1.2). Among others, *Exploratory Multivariate Analysis* includes cluster analysis (see Chap. 13) and factor analysis (see Chap. 14). These methods allow us to detect patterns in datasets that contain many observations and variables. Cluster analysis, for instance, can be used to create different segments of customers for a product, grouping them, say, by purchase frequency and purchase amount.

Unfortunately, many empirical studies end after they undertake exploratory analysis. However, exploratory analysis does not check the statistical significance of the detected pattern or structure. A technique like cluster analysis can identify different customer groups but it cannot guarantee that they differ from each other significantly. To do that, we need *testing techniques* such as regression analysis (see Chap. 10) and analysis of variance (see Sect. 9.5.1).

Exploratory techniques	Objective
Factor analysis	Reduce (i.e. pool) variables to most important factors
Cluster analysis	Pool objects/subjects in homogeneous groups
Testing techniques	**Objective**
(Logistic) Regression analysis	Test the influence of independent variables on one or more dependent variables
Analysis of Variance (ANOVA)	

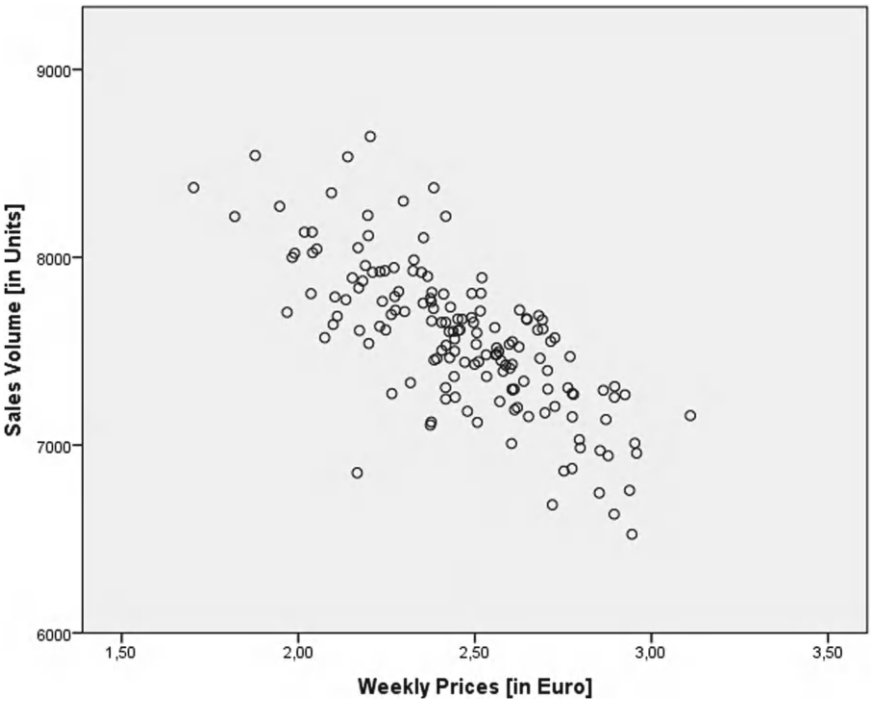
Fig. 1.2 Techniques for multivariate analysis

1.3 The Generation of Knowledge Through Statistics

The fundamental importance of statistics in the human effort to generate new knowledge should not be underestimated. Indeed, the process of knowledge generation in science and professional practice typically involves both of the aforementioned descriptive and inductive steps. This fact can be easily demonstrated with an example.

Imagine that a market researcher in the field of dentistry is interesting in figuring out the relationship between the price and volume of sales for a specific brand of toothpaste. The researcher would first attempt to gain an understanding of the market by gathering individual pieces of information. He could, for example, analyse weekly toothpaste prices and sales over the last three years (see Fig. 1.3). As is often the case when gathering data, it is likely that sales figures are not available for some stores, such that no full survey is possible, but rather only a partial sample. Imagine that our researcher determines that in the case of high prices, sales figures fall, as demand moves to other brands of toothpaste, and that, in the case of lower prices, sales figures rise once again. However, this relationship, which has been determined on the basis of descriptive statistics, is not a finding solely applicable to the present case. Rather, it corresponds precisely to the microeconomic price and demand function. Invariably in such cases, it is the methods of descriptive statistics that allow us to draw insights concerning specific phenomena, insights which, on the basis of individual pieces of data, demonstrate the validity (or, in some cases, non-validity) of existing expectations or theories.

At this stage, our researcher will ask himself whether the insights obtained on the basis of this partial sample—insights which he, incidentally, expected beforehand—can be viewed as representative of the entire population. Generalizable information in descriptive statistics is always initially speculative. With the aid of inductive statistical techniques, however, one can estimate the error probability associated with applying insights obtained through descriptive statistics to an overall population. The researcher must decide for himself which level of error probability renders the insights insufficiently qualified and inapplicable to the overall population.



The figure shows the average weekly prices and associated sales volumes over a three year period. Each point represents the amount of units sold at a certain price within a given week.

Fig. 1.3 Price and demand function for sensitive toothpaste

Yet even if all stores reported their sales figures, thus providing a full survey of the population, it would be necessary to ask whether, ceteris paribus, the determined relationship between price and sales will also hold true in the future. Data from the future are of course not available. Consequently, we are forced to forecast the future based on the past. This process of forecasting is what allows us to verify theories, assumptions, and expectations. Only in this way can information be transformed into generalizable knowledge (in this case, for the firm).

Descriptive and inductive statistics thus fulfil various purposes in the research process. For this reason, it is worthwhile to address each of these domains separately, and to compare and contrast them. In university courses on statistics, these two domains are typically addressed in separate lectures.

1.4 The Phases of Empirical Research

The example provided above additionally demonstrates that the process of knowledge generation typically goes through specific phases. These phases are illustrated in Fig. 1.4. In the *Problem Definition Phase*, the goal is to establish a common understanding of the problem and a picture of potential interrelationships. This may require discussions with decision-makers, interviews with experts, or an initial screening of data and information sources. In the subsequent *Theory Phase*, these potential interrelationships are then arranged within the framework of a cohesive model.

1.4.1 From Exploration to Theory

Although the practitioner uses the term *theory* with reluctance, for he fears being labelled *overly academic* or *impractical*, the development of a theory is a necessary first step in all efforts to advance knowledge. The word theory is derived from the Greek term *theorema* which can be translated as *to view*, *to behold*, or *to investigate*. A theory is thus knowledge about a system that takes the form of a speculative description of a series of relationships (Crow, 2005, p. 14). On this basis, we see that the postulation of a theory hinges on the observation and linkage

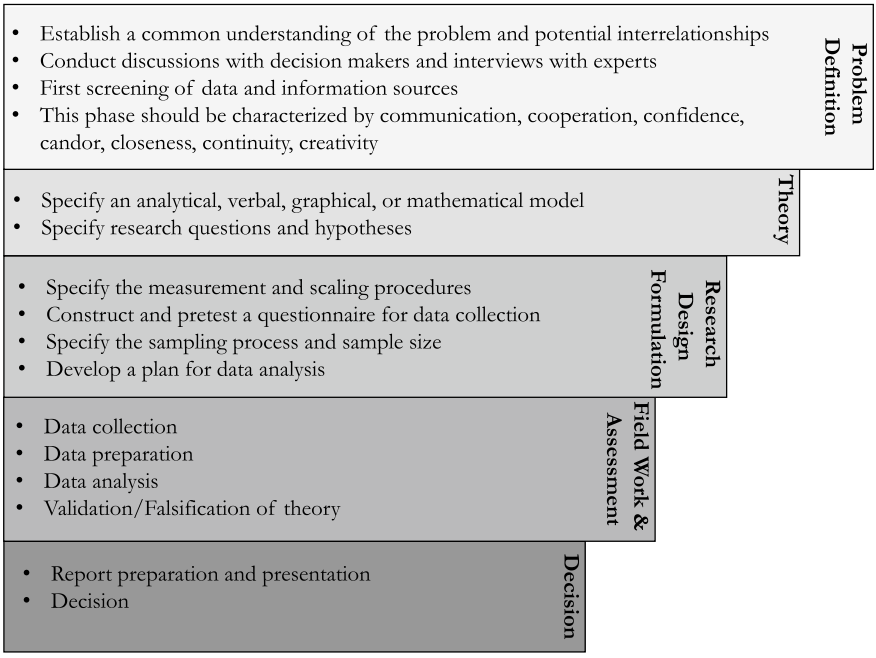


Fig. 1.4 Phases of empirical research

of individual events, and that a theory cannot be considered generally applicable without being verified. An empirical theory draws connections between individual events so that the origins of specific observed conditions can be deduced. The core of every theory thus consists in the establishment of a unified terminological system according to which cause-and-effect relationships can be deduced. In the case of our toothpaste example, this means that the researcher first has to consider which causes (i.e. factors) have an impact on sales of the product. The most important causes are certainly apparent to researcher based on a *gut feeling*: the price of one's own product, the price of competing products, advertising undertaken by one's own firm and competitors, as well as the target customers addressed by the product, to name but a few.

Alongside these factors, other causes which are hidden to those unfamiliar with the sector also normally play a role. Feedback loops for the self or third-person verification of the determinations made thus far represent a component of both the Problem Definition and Theory Phases. In this way, a quantitative study always requires strong communicative skills. All properly conducted quantitative studies rely on the exchange of information with outside experts—e.g. in our case, product managers—who can draw attention to hidden events and influences. Naturally, this also applies to studies undertaken in other departments of the company. If the study concerns a procurement process, purchasing agents need to be queried. Alternatively, if we are dealing with an R&D project, engineers are the ones to contact, and so on. Yet this gathering of perspectives doesn't just improve a researcher's understanding of causes and effects. It also prevents the embarrassment of completing a study only to have someone point out that key influencing factors have been overlooked.

1.4.2 From Theories to Models

Work on constructing a model can begin once the theoretical interrelationships that govern a set of circumstances have been established. The terms *theory* and *model* are often used as synonyms, although, strictly speaking, *theory* refers to a language-based description of reality. If one views mathematical expressions as a language with its own grammar and semiotics, then a theory could also be formed on the basis of mathematics. In professional practice, however, one tends to use the term *model* in this context—a model is merely a theory applied to a specific set of circumstances.

Models are a technique by which various theoretical considerations are combined in order to render an approximate description of reality (see Fig. 1.5). An attempt is made to take a specific real-world problem and, through *abstraction* and *simplification*, to represent it formally in the form of a structurally cohesive model. The model is structured to reflect the totality of the traits and relationships that characterize a specific subset of reality. Thanks to models, the problem of mastering the complexity that surrounds economic activity initially seems to be solved: it would appear that in order to reach rational decisions that ensure

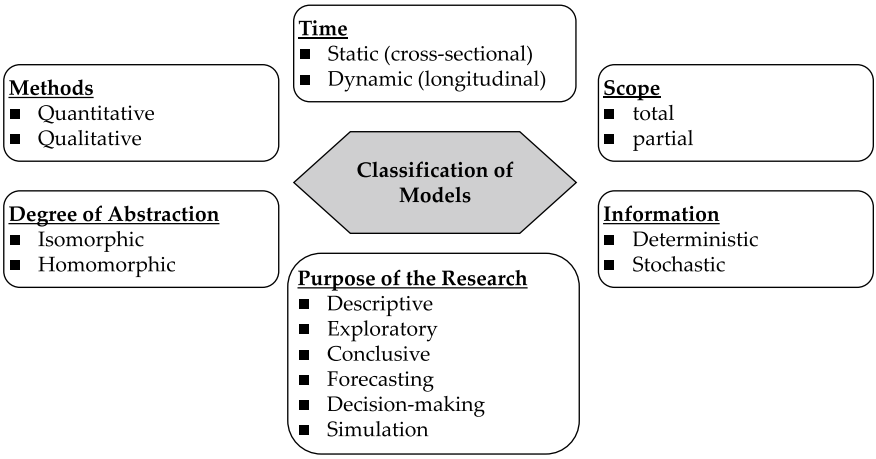


Fig. 1.5 A systematic overview of model variants

the prosperity of a firm or the economy as a whole, one merely has to assemble data related to a specific subject of study, evaluate these data statistically, and then disseminate one's findings. In actual practice, however, one quickly comes to the realization that the task of providing a comprehensive description of economic reality is hardly possible, and that the decision-making process is an inherently messy one. The myriad aspects and interrelationships of economic reality are far too complex to be comprehensively mapped. The mapping of reality can never be undertaken in a manner that is structurally homogenous—or, as one also says, *isomorphic*. No model can fulfil this task. Consequently, models are almost invariably reductionist, or *homomorphic*.

The accuracy with which a model can mirror reality—and, by extension, the process of model enhancement—has limits. These limits are often dictated by the imperatives of practicality. A model should not be excessively complex such that it becomes unmanageable. It must reflect the key properties and relations that characterize the problem for which it was created to analyse, and it must not be alienated from this purpose. Models can thus be described as mental constructions built out of abstractions that help us portray complex circumstances and processes that cannot be directly observed (Bonhoeffer, 1948, p. 3). A model is solely an approximation of reality in which complexity is sharply reduced. Various methods and means of portrayal are available for representing individual relationships. The most vivid one is the *physical* or *iconic* model. Examples include dioramas (e.g. wooden, plastic, or plaster models of a building or urban district), maps, and blueprints. As economic relationships are often quite abstract, they are extremely difficult to represent with a physical model.

Symbolic models are particularly important in the field of economics. With the aid of language, which provides us with a system of symbolic signs and an accompanying set of syntactic and semantic rules, we use symbolic models to investigate

and represent the structure of the set of circumstances in an approximate fashion. If everyday language or a specific form of jargon serves as the descriptive language, then we are speaking of a *verbal model* or of a *verbal theory*. At its root, a verbal model is an assemblage of symbolic signs and words. These signs don't necessarily produce a given meaning. Take, for example, the following constellation of words: "Spotted lives in Chicago my grandma rabbit". Yet even the arrangement of the elements in a syntactically valid manner—"My grandma is spotted and her rabbit lives in Chicago"—does not necessarily produce a reasonable sentence. The verbal model only makes sense when semantics are taken into account and the contents are linked together in a meaningful way: "My grandma lives in Chicago and her rabbit is spotted".

The same applies to artificial languages such as logical and mathematical systems, which are also known as *symbolic models.* These models also require character strings (variables), and these character strings must be ordered syntactically and semantically in a system of equations. To refer once again to our toothpaste example, one possible verbal model or theory could be the following:

- There is an inverse relationship between toothpaste sales and the price of the product, and a direct relationship between toothpaste sales and marketing expenditures during each period (i.e. calendar week).
- The equivalent formal symbolic model is thus as follows: $y_i = f(p_i, w_i) = \alpha_1 \cdot p_i + \alpha_2 \cdot w_i + \beta$. p_i: price at point in time i; w_i: marketing expenditures at point in time i; α refers to the effectiveness of each variable; β is a possible constant.

Both of these models are homomorphic *partial models*, as only one aspect of the firm's business activities—in this case, the sale of a single product—is being examined. For example, we have not taken into account changes in the firm's employee headcount or other factors. This is exactly what one would demand from a *total model*, however. Consequently, the development of total models is in most cases prohibitively laborious and expensive. Total models thus tend to be the purview of economic research institutes.

Stochastic, homomorphic, and partial models are the models that are used in statistics (much to the chagrin of many students in business and economics). Yet what does the term *stochastic* mean? Stochastic analysis is a type of inductive statistics that deals with the assessment of non-deterministic systems. *Chance* or *randomness* are terms we invariably confront when we are unaware of the causes that lead to certain events, i.e. when events are *non-deterministic*. When it comes to future events or a population that we have surveyed with a sample, it is simply impossible to make forecasts without some degree of uncertainty. Only the past is certain. The poor chap in Fig. 1.6 demonstrates how certainty can be understood differently in everyday contexts.

Yet economists have a hard time dealing with the notion that everything in life is uncertain and that one simply has to accept this. To address uncertainty, economists attempt to estimate the probability that a given event will occur using inductive statistics and stochastic analysis. Naturally, the young man depicted in



Fig. 1.6 What is certain? © Marco Padberg

the image of Fig. 1.6 would have found little comfort had his female companion indicated that there was a 95% probability (i.e. very high likelihood) that she would return the following day. Yet this assignment of probability clearly shows that the statements used in everyday language—i.e. *yes* or *no*, and *certainly* or *certainly not*—are always to some extent a matter of conjecture when it comes to future events. However, statistics cannot be faulted for its conjectural or uncertain declarations, for statistics represents the very attempt to quantify certainty and uncertainty and to take into account the random chance and incalculables that pervade everyday life (Swoboda, 1971, p. 30).

Another important aspect of a model is its purpose. In this regard, we can differentiate between the following model types:

- Descriptive models
- Explanatory models or forecasting models
- Decision models or optimization models
- Simulation models.

The question asked and its complexity ultimately determines the purpose a model must fulfil.

Descriptive models merely intend to describe reality in the form of a model. Such models do not contain general hypotheses concerning causal relationships in real systems. A profit and loss statement, for example, is nothing more than an attempt to depict the financial situation of a firm within the framework of a model. Assumptions concerning causal relationships between individual items in the statement are not depicted or investigated.

Explanatory models, by contrast, attempt to codify theoretical assumptions about causal connections and then test these assumptions on the basis of empirical data. Using an explanatory model, for example, one can seek to uncover interrelationships between various firm-related factors and attempt to project these factors

into the future. In the latter case—i.e. the generation of forecasts about the future—one speaks of *forecasting models*, which are viewed as a type of explanatory model. To return to our toothpaste example, the determination that a price reduction of €0.10 leads to a sales increase of 10,000 tubes of toothpaste would represent an explanatory model. By contrast, if we forecasted that a price increase of €0.10 *this* week (i.e. at time t) would lead to a fall in sales *next* week (i.e. at time $t + 1$), then we would be dealing with a forecasting, or prognosis, model.

Decision models, which are also known as optimization models, are understood by Grochla (1969, p. 382) to be "systems of equations aimed at deducing recommendations for action". The effort to arrive at an optimal decision is characteristic of decision models. As a rule, a mathematical target function that the user hopes to optimize while adhering to specific conditions serves as the basis for this type of model. Decision models are used most frequently in Operations Research and are less common in statistical data analysis (Runzheimer et al., 2005).

Simulation models are used to "recreate" procedures and processes—for example, the phases of a production process. The random-number generator function in statistical software allows us to uncover interdependencies between the examined processes and stochastic factors (e.g. variance in production rates). Yet roleplaying exercises in leadership seminars or Family Constellation sessions can also be viewed as simulations.

1.4.3 From Models to Business Intelligence

Statistical methods can be used to gain a better understanding of even the most complicated circumstances and situations. While not all of the analytical methods that are employed in practice can be portrayed within the scope of this textbook, it takes a talented individual to master all of the techniques that will be described in the coming pages. Indeed, everyone is probably familiar with a situation similar to the following: an exuberant but somewhat overintellectualized professor seeks to explain the advantages of the Heckman Selection Model to a group of business professionals (see Heckman, 1976). Most listeners will be able to follow the explanation for the first few minutes—or at least for the first few seconds. Then uncertainty sets in, as each listener asks: am I the only one who understands nothing right now? But a quick look around the room confirms that others are equally confused. The audience slowly loses interest, and minds wander. After the talk is over, the professor is thanked for his illuminating presentation. And those in attendance never end up using the method that was presented.

Thankfully, some presenters are aware of the need to avoid excessive technical detail, and they do their best to explain the results that have been obtained in a matter that is intelligible to mere mortals. Indeed, the purpose of data analysis is not the analysis itself, but rather the communication of findings in an audience-appropriate manner. Only findings that are understood and accepted by decision-makers can affect decisions and future reality. Analytical procedures must

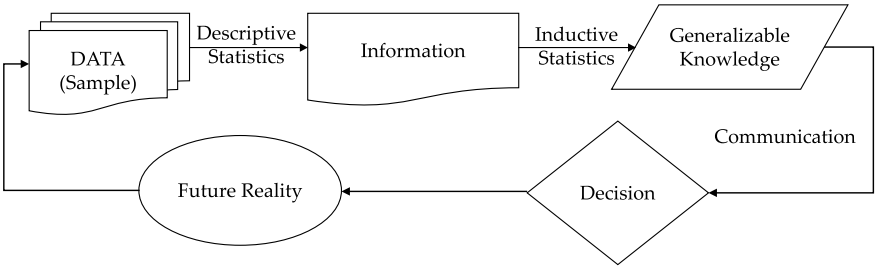


Fig. 1.7 Intelligence cycle. *Source* Own graphic, adapted from Harkleroad (1996, p. 45)

therefore be undertaken in a goal-oriented manner, with an awareness for the informational needs of a firm's management (even if these needs are not clearly defined in advance).

Consequently, the communication of findings, which is the final phase of an analytical project, should be viewed as an integral component of any rigorously executed study. In Fig. 1.7, the processes that surround the construction and implementation of a decision model are portrayed schematically as an *intelligence cycle* (Kunze, 2000, p. 70). The intelligence cycle is understood as "the process by which raw information is acquired, gathered, transmitted, evaluated, analysed, and made available as finished intelligence for policymakers to use in decision-making and action" (Kunze, 2000, p. 70). In this way, the intelligence cycle is "[…] an analytical process that transforms disaggregated […] data into actionable strategic knowledge […]" (Bernhardt, 1994, p. 12).

In the following chapter of this book, we will look specifically at the activities that accompany the assessment phase (cf. Fig. 1.4). In these phases, raw data are gathered and transformed into information with strategic relevance by means of descriptive assessment methods, as portrayed in the intelligence cycle above.

References

Bernhardt, D. C. (1994). I want it fast, factual, actionable—Tailoring competitive intelligence to executives' needs. *Long Range Planning, 27*(1), 12–24.

Bonhoeffer, K. F. (1948). *Über physikalisch-chemische Modelle von Lebensvorgängen.* Akademie Verlag.

Crow, D. (2005). *Zeichen. Eine Einführung in die Semiotik für Grafikdesigner*. Stiebner.

Dubbern, H.-H., & Beck-Bornholdt, H. P. (2007). *Der Hund der Eier legt. Erkennen von Fehlinformation durch Querdenken* (2nd ed.). Rowohlt Taschenbuch Verlag.

Grochla, E. (1969). Modelle als Instrumente der Unternehmensführung. *Zeitschrift für betriebswirtschaftliche Forschung (ZfbF), 21*, 382–397.

Harkleroad, D. (1996). Actionable competitive intelligence, society of competitive intelligence professionals. In *Annual International Conference & Exhibit Conference Proceedings* (pp. 43–52). Alexandria, VA.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models. *The Annals of Economic and Social Measurement, 5*(4), 475–492.

Krämer, W. (2015). *So lügt man mit Statistik* (17th ed.). Campus.

Kunze, C. W. (2000). *Competitive Intelligence. Ein ressourcenorientierter Ansatz strategischer Frühaufklärung*. Shaker.

Runzheimer, B., Cleff, T., & Schäfer, W. (2005). *Operations Research 1: Lineare Planungsrechnung und Netzplantechnik* (8th ed.). Gabler.

Swoboda, H. (1971). *Exakte Geheimnisse: Knaurs Buch der modernen Statistik*. Knaur.



From Disarray to Dataset

2

2.1 Get Your Software Ready

Statistical analysis is the backbone of modern data-driven decision-making. In today's data-centric world, effective statistical and multivariate data analysis can unlock hidden patterns, trends, and correlations within vast datasets. Whether you are working in academia, industry, or research, ensuring your computer is primed for statistical analysis will enable you to extract meaningful information, make informed decisions, and uncover actionable solutions to real-world problems. It empowers researchers, analysts, and professionals across various disciplines to derive valuable insights from complex datasets.

It is crucial to have a well-prepared computer and software environment that can efficiently handle data processing, statistical modelling, and visualization tasks in order to embark on the journey of statistical (multivariate) data analysis. Therefore, statistical software should provide efficient and user-friendly interfaces for a wide range of statistical analyses and should include data visualization tools and data management capabilities, ensuring that datasets can be well prepared for analyses. And indeed, today, users can choose from a wide range of paid or free statistical software packages so that sometimes the choice in favour or against a particular statistical software is not that easy.

Although this book shows how to use R, SPSS, Stata, and Excel, each user should choose a software based on his or her specific needs, the complexity of the targeted data analysis tasks, the user's familiarity with statistical software, and last but not least on the available budget to obtain the software. Many analysts and researchers use a combination of these tools to leverage their respective strengths

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78070-7_2.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*,
Springer Texts in Business and Economics,
https://doi.org/10.1007/978-3-031-78070-7_2

in different aspects of data analysis. Key differences are primarily related to (a) the complexity, (b) the specialization, and (c) the cost of the software.

- Complexity: Excel is the simplest and most general-purpose tool, while R is probably the most powerful and flexible but has a steeper learning curve.
- Specialization: R, Stata, and SPSS are designed specifically for statistical analysis, while Excel is more of a general-purpose spreadsheet tool.
- Cost: Excel may come bundled with Microsoft Office, while Stata and SPSS typically require separate licenses, with R being open-source and free.

Excel is a versatile spreadsheet software primarily used for data entry, basic data analysis, and general-purpose calculations. It is not specifically designed for statistical analysis but can handle simple statistical functions and basic charting. It is widely used by professionals in various fields for tasks such as budgeting, financial analysis, and data tracking.

That is why Excel is a widely accessible and user-friendly option for basic and intermediate statistical tasks, covering some of the statistical techniques presented in this book well. It contains a number of pre-programmed statistical functions.[1] Using these, more complex methods such as Regression Analysis or Analysis of Variance can be performed. However, as the complexity of the used statistical methods increases, both the user-friendliness and the range of possible options related to the respective method decrease. In addition, the requirements for the user's programming abilities increase. For statistical methods such as logistic regression, cluster analysis, and factor analysis, the basic version of Excel does not provide any pre-programmed function. While Excel is a powerful tool for basic and intermediate statistical analysis, it may not be as feature-rich or specialized as dedicated statistical software like R, Stata, or SPSS. Complex or specialized analyses may require these more advanced tools.

SPSS (Statistical Package for the Social Sciences) is a comprehensive statistical software package designed for data analysis, data mining, and reporting. It is widely used in social sciences, psychology, marketing, and market research, offers a user-friendly interface, and is suitable for basic and advanced statistical analyses. It is favoured by researchers and analysts who appreciate a point-and-click interface for performing advanced statistical tasks, but also by experienced analysts working with syntax and customized scripts.[2]

[1] These functions can be found under *Formulas* → *Insert Function* (select the category *Statistical* to set the constraints). It is also possible to use the Add-ins Manager to permanently activate the *Analysis ToolPak* and the *Analysis ToolPak VBA* for Excel. The Add-Ins Manager can be accessed via *File* → *Options* → *Add-ins* → Manage: Excel Add-ins → Go. Next, go to *Data* → *Data Analysis* and choose a more complex statistical technique.

[2] SPSS is organized with a core system that provides basic data management and analysis tools, and users can enhance its functionality by adding specific modules or extensions (so-called add-ins) that cater to their analytical needs. These add-ins can be purchased separately and its combination

Undoubtedly, for beginners who are new to data analysis or lack programming skills, the point-and-click interface provides an easy entry point without the need to learn a new syntax language first. Beginners can initially concentrate only on the interpretation of statistical analyses. However, it is important to note that over time many users, especially as they progress to more advanced data analysis tasks, often transition to learning and using syntax programming because it provides more control, automation capabilities, and flexibility. Please refer to Appendix for more information on how to get SPSS ready for use.

Stata is a statistical software package designed for advanced data management and statistical analysis. It is commonly used in academic research, economics, social sciences, and healthcare research and is particularly useful for managing and analysing large datasets and conducting a wide range of statistical methods. In addition to the Stata basic functions, users can install so-called ado files on their systems to benefit from added functionalities and methods created by the Stata community. While SPSS is often preferred for its user-friendly interface and is easy to use in basic statistical analyses, Stata is known for its wide scope of robust statistical capabilities and for its efficiency in handling large datasets and complex survey data. Similar to SPSS, Stata can be navigated either by a user interface or by its own syntax, which makes it easy to work with the software also for beginners. Please refer to Appendix for more information on how to get Stata ready for use.

Among many others, SPSS and Stata are popular statistical software packages, each with its own strengths and weaknesses. In recent years, however, R has emerged as the cutting-edge statistical software. R was first developed in the year 1993 by the statisticians Ross Ihaka and Robert Gentleman at the University of Auckland in New Zealand. From the very beginning on, R was initiated as an open-source software project to provide a free alternative to commercial statistical software programmes. Since then, R has evolved into one of the most widely used statistical programming languages and environments, and it is used by statisticians, data analysts, and scientists across various disciplines worldwide. R has a vast and active community of users and developers who have created thousands of packages for a wide range of statistical and data analysis tasks. Therefore, R tends to be at the forefront of statistical research. New statistical methods and techniques are often implemented in R packages before they become available in Stata or SPSS. This is especially important for researchers and analysts who need access to the latest advancements in statistics. In addition, R can be easily integrated with other programming languages and tools, such as Python, SQL, and LaTeX, allowing for seamless workflows and automation.

For a long time, however, the main disadvantage of R was the lack of a graphical user interface, which made the use of Stata and SPSS much easier for novice users of statistical software. The R Commander, often referred to as “Rcmdr”, has filled this gap by providing a sound graphical user interface for the R statistical

with the core system makes SPSS a versatile tool for a wide range of basic and advanced statistical and data analysis tasks. For further information, see https://www.ibm.com/docs/en/spss-statistics.

software. R Commander was initially developed by John Fox in 2003, a professor of sociology at McMaster University in Canada. He created it as a response to the need for a user-friendly interface for R, primarily for the benefit of students and researchers who were not familiar with programming in R.

All calculations based on R are shown in conjunction with the R Commander in order to facilitate the learning of R in the context of this book. Please refer to Appendix for more information on how to get R and the R Commander ready for use.

In conclusion, it is important to note that the choice between Excel, R, Stata, or SPSS depends on individual preferences, specific project requirements, and the user's level of expertise. Some users may find Stata or SPSS more user-friendly for certain tasks, while others may prefer R's flexibility and extensive package ecosystem for more complex and specialized analyses.

2.2 Data Collection

Let us begin with the first step of the intelligence cycle: data collection. Many businesses gather crucial information—on expenditures and sales, say—but few enter it into a central database for systematic evaluation. The first task of the statistician is to mine this valuable information. Often, this requires skills of persuasion: employees may be hesitant to give up data for the purpose of systematic analysis, for this may reveal past failures.

But even when a firm has decided to systematically collect data, preparation may be required prior to analysis. Who should be authorized to evaluate the data? Who possesses the skills to do so? And who has the time? Businesses face questions like these on a daily basis, and they are no laughing matter. Consider the following example: when tracking customer purchases with loyalty cards, companies obtain extraordinarily large datasets. Administrative tasks alone can occupy an entire department, and this is before systematic evaluation can even begin.

In addition to the data they collect themselves, firms can also find information in public databases. Sometimes these databases are assembled by private marketing research firms such as ACNielsen or the GfK Group, which usually charge a data access fee. The databases of research institutes, federal and local statistics offices, and many international organizations (Eurostat, the OECD, the World Bank, etc.) may be used for free. Either way, public databases often contain valuable information for business decisions. Table 2.1 provides a list of links to some interesting sources of data.

Let's take a closer look at how public data can aid business decisions. Imagine a procurement department of a company that manufacturers intermediate goods for machine construction. In order to lower costs, optimize stock levels, and fine-tune order times, the department is tasked with forecasting stochastic demand for materials and operational supplies. They could of course ask the sales department about future orders, and plan production and material needs accordingly. But experience shows that sales departments vastly overestimate projections to ensure

Table 2.1 External data sources at international institutions

German Federal Statistical Office	destatis.de	Offers links to diverse international databases
Eurostat	https://ec.europa.eu/eurostat/	Various databases
OECD	oecd.org	Various databases
World Bank	worldbank.org	World and country-specific development indicators
UN	un.org	Diverse databases
ILO	ilo.org	Labour statistics and databases
IMF	imf.org	Global economic indicators, financial statistics, information on direct investment, etc.

delivery capacity. So the procurement (or inventory) department decides to consult the most recent Ifo Business Climate Index.[3] Using this information, the department staff can create a valid forecast of the end-user industry for the next six months. If the end-user industry sees business as trending downward, the sales of our manufacturing company are also likely to decline, and vice versa. In this way, the procurement department can make informed order decisions using public data instead of conducting its own surveys.

Public data may come in various states of aggregation. Such data may be based on a category of company or group of people, but only rarely on a single firm or individual. For example, the Centre for European Economic Research (ZEW) conducts recurring surveys on industry innovation. These surveys never contain data on a single firm, but rather data on a group of firms—say, the R&D expenditures of chemical companies with between 20 and 49 employees. This information can then be used by individual companies to benchmark their own indices. Another example is the GfK household panel, which contains data on the purchase activity of households, but not of individuals. Loyalty card data also provide, in effect, aggregate information, since purchases cannot be traced back reliably to particular cardholders (as a husband, for example, may have used his wife's card to make a purchase). Objectively speaking, loyalty card data reflect only a household, but not its members.

[3] The Ifo Business Climate Index is released each month by Germany's Ifo Institute. It is based on a monthly survey that queries some 7000 companies in the manufacturing, construction, wholesaling, and retailing industries about a variety of subjects: the current business climate, domestic production, product inventory, demand, domestic prices, order change over the previous month, foreign orders, exports, employment trends, three-month price outlook, and six-month business outlook.

To collect information about individual persons or firms, one must conduct a *survey*. Typically, this is most expense form of data collection. But it allows companies to specify their own questions. Depending on the subject, the survey can be oral or written. The traditional form of survey is the questionnaire, though telephone and Internet surveys are also becoming increasingly popular.

2.3 Level of Measurement

It would go beyond the scope of this textbook to present all of the rules for the proper construction of questionnaires. For more on questionnaire design, the reader is encouraged to consult other sources (see, for instance, Malhotra et al. (2017)). Consequently, we focus below on the criteria for choosing a specific quantitative assessment method.

Let us begin with an example. Imagine you own a little grocery store in a small town. Several customers have requested that you expand your selection of butter and margarine. Because you have limited space for display and storage, you want to know whether this request is *representative* of the preferences of all your customers. You thus hire a group of students to conduct a survey using the short questionnaire in Fig. 2.1.

Within a week, the students have collected questionnaires from 850 customers. Each individual survey is a *statistical unit* with certain relevant *traits*. In this questionnaire, the relevant traits are *gender*, *age*, *body weight*, *preferred bread spread*, and *selection rating*. One customer—we'll call him Mr. Smith—has the *trait values* of *male*, *67 years old*, *74 kg*, *margarine*, and *fair*. Every survey requires that the designer first define the statistical unit (who to question?), the relevant traits or variables (what to question?), and the trait values (what answers can be given?). Variables can be classified as either discrete or continuous variables. *Discrete*

Sex: ☐ male ☐ female

Age: _____

Body weight: _____ kg

Which spread do you prefer? *(Choose one answer)*
☐ butter ☐ margarine ☐ other

On a scale of 1 (poor) to 5 (excellent) how do rate the selection of your preferred spread at our store?

☐(1)	☐(2)	☐(3)	☐(4)	☐(5)
poor	fair	average	good	excellent

Fig. 2.1 Retail questionnaire

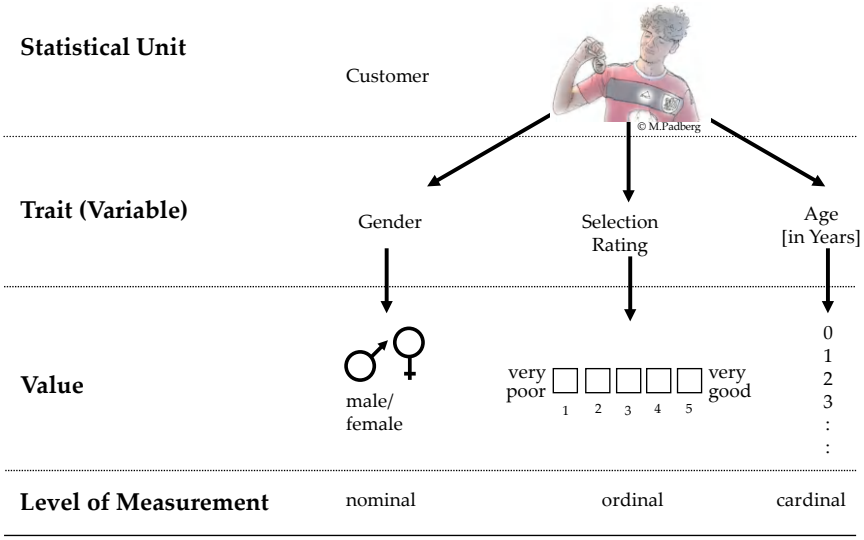


Fig. 2.2 Statistical units/traits/trait values/level of measurement

variables can only take on certain given numbers—normally whole numbers—as possible values. There are usually gaps between two consecutive outcomes. The *size of a family* (1, 2, 3, …) is an example of a discrete variable. *Continuous variables* can take on any value within an interval of numbers. All numbers within this interval are possible. Examples are variables such as *weight* or *height*.

Generally speaking, the statistical units are the subjects (or objects) of the survey. They differ in terms of their values for specific traits. The traits *gender*, *selection rating*, and *age* shown in Fig. 2.2 represent the three levels of measurement in quantitative analysis: the nominal scale, the ordinal scale, and the cardinal scale, respectively.

The lowest level of measurement is the *nominal scale*. With this level of measurement, a number is assigned to each possible trait (e.g. $x_i = 1$ for *male* or $x_i = 2$ for *female*). A *nominal variable* is sometimes also referred to as *qualitative variable*, or *attribute*. The values serve to assign each statistical unit to a specific group (e.g. the group of *male* respondents) in order to differentiate it from another group (e.g. the *female* respondents). Every statistical unit can only be assigned to one group and all statistical units with the same trait status receive the same number. Since the numbers merely indicate a group, they do not express qualities such as larger/smaller, less/more, or better/worse. They only designate membership or non-membership in a group ($x_i = x_j$ versus $x_i \neq x_j$). In the case of the trait *gender*, a *one* for *male* is no better or worse than a *two* for *female*; the data are merely segmented in terms of male and female respondents. Neither does rank play a role in other nominal traits, including profession (e.g. 1 = *butcher*; 2 = *baker*; 3 = *chimney sweep*), nationality, class year, etc.

This leads us to the next highest level of measurement, the *ordinal scale*. With this level of measurement, numbers are also assigned to individual value traits, but here they express a rank. The typical examples are answers based on scales from one to x, as with the trait *selection rating* in the sample survey. This level of measurement allows researchers to determine the intensity of a trait value for a statistical unit compared to that of other statistical units. If Ms. Peters and Ms. Miller both check the third box under *selection rating*, we can assume that both have the same perception of the store's selection. As with the nominal scale, statistical units with the same values receive the same number. If Mr. Martin checks the fourth box, this means both that his perception is different from that of Ms. Peters and Ms. Miller, and that he thinks the selection is *better* than they do. With an ordinal scale, traits can be ordered, leading to qualities such as larger/smaller, less/more, and better/worse ($x_i = x_j$; $x_i > x_j$; $x_i < x_j$).

What we cannot say is how large the distance is between the third and fourth boxes. We cannot even assume that the distance between the first and second boxes is as large as that between other neighbouring boxes. Consider an everyday example of an ordinal scale: standings at athletic competitions. The difference between each place does not necessarily indicate a proportional difference in performance. In a swimming competition, the time separating first and second place may be one one-thousandth of a second, with third place coming in two seconds later, yet only one place separates each.

The highest level of measurement is the *metric* or *cardinal scale*. It contains not only the information of the ordinal scales—larger/smaller, less/more, better/worse ($x_i = x_j$; $x_i > x_j$; $x_i < x_j$)—but also the distance between value traits held by two statistical units. Age is one example. A 20-year-old is not only older than an 18-year-old; a 20-year-old is exactly two years older than an 18-year-old. Moreover, the distance between a 20-year-old and a 30-year-old is just as large as the distance between an 80-year-old and a 90-year-old. The graduations on a cardinal scale are always equidistant. In addition to age, typical examples for cardinal scales are currency, weight, length, and speed.

Cardinal scales are frequently differentiated into absolute scales,[4] ratio scales,[5] and interval scales.[6] These distinctions tend to be academic and seldom play much role in deciding which statistical method to apply. This cannot be said of the distinction between cardinal- and ordinal-scaled variables, however. On account of the much greater variety of analysis methods for cardinal scales in relation to ordinal methods, researchers often tend to see ordinal variables as cardinal in nature. For example, researchers might assume that the gradations on the five-point scale used for rating selection in our survey example are identical. We frequently find such assumptions in empirical studies. More serious researchers note in passing that *equidistance* has been assumed or offer justification for such *equidistance*. Schmidt

[4] A metric scale with a natural zero point and a natural unit (e.g. age).

[5] A metric scale with a natural zero point but without a natural unit (e.g. surface).

[6] A metric scale without a natural zero point and without a natural unit (e.g. geographical longitude).

and Opp (1976, p. 35) have proposed a rule of thumb according to which ordinal-scaled variables can be treated as cardinal-scaled variables: the ordinal scale must have more than four possible outcomes and the survey must have more than 100 observations (see also Pell (2005), Carifio and Perla (2008)). Still, interpreting a difference of 0.5 between two ordinal scale averages is difficult and is a source of many headaches among empirical researchers.

As this section makes clear, a variable's scale is crucial because it determines which statistical method to apply. For a nominal variable like *profession* it is impossible to determine the mean value of three backers, five butchers, and two chimney sweeps. Later in the book, I will discuss which statistical method goes with which level of measurement or combination of measurements.

Before data analysis can begin, the collected data must be transferred from paper or from an online survey platform to a form that can be read and processed by a computer. We will continue to use the 850 questionnaires collected by the students as an example.

2.4 Scaling and Coding

To emphasize again, the first step in conducting a survey is to define the level of measurement for each trait. In most cases, it is impossible to raise the level of measurement after a survey has been implemented (i.e. from nominal to ordinal, or from ordinal to cardinal). If a survey asks respondents to indicate their age not by years but by age group, this variable must remain on the ordinal scale. This can be a great source of frustration: among other things, it makes it impossible to determine the exact average age of respondents in retrospect. It is therefore always advisable to set a variable's level of measurement as high as possible beforehand (e.g. age in years or expenditures for a consumer good).

The group or person who commissions a survey may stipulate that questions remain on a lower level of measurement in order to ensure anonymity. When a company's works council is involved in implementing a survey, for example, one may encounter such a request. Researchers are normally obligated to accommodate such wishes.

In our above sample survey, the following levels of measurement were used:

- Nominal: gender; preferred spread
- Ordinal: selection rating
- Cardinal: age; body weight.

Now, how can we communicate this information to the computer? Every statistics application contains an Excel-like spreadsheet in which data can be entered directly (see, for instance, Fig. 3.1). While columns in Excel spreadsheets are typically named A, B, C, etc., the columns in more professional spreadsheets are labelled with the *variable name*. Typically, variable names may be no longer than eight characters. So, for instance, the variable *selection rating* is given as

"*selection*". For clarity's sake, a variable name can be linked to a longer *variable label* or to an entire survey question. The software commands use the variable names while the printout of the results displays the complete label.

The next step is to enter the survey results into the spreadsheet. The answers from questionnaire #1 go in the first row, those from questionnaire #2 go in the second row, and so on. A computer can only "understand" numbers. For cardinal-scaled variables, this is no problem, since all of the values are numbers anyway. Suppose person #1 is 31 years old and weighs 63 kg. Simply enter the numbers 31 and 63 in the appropriate row for respondent #1. Nominal and ordinal variables are more difficult and require that all contents be coded with a number. In the sample dataset, for instance, the nominal scale traits *male* and *female* are assigned the numbers "0" and "1", respectively. The number assignments are recorded in a label book, as shown in Fig. 2.3. Using this system, you can now enter the remaining results.

```
-------------------------------------------------------------------------------
value label: selection
-------------------------------------------------------------------------------
  definition
           1   poor
           2   fair
           3   average
           4   good
           5   excellent
   variables:  selection
-------------------------------------------------------------------------------
value label: brd_sprd
-------------------------------------------------------------------------------
  definition
           0   butter
           1   margarine
           2   other
   variables:  bread_spread
-------------------------------------------------------------------------------
value label: gender
-------------------------------------------------------------------------------
  definition
           0   male
           1   female
   variables:  gender
```

Fig. 2.3 Label book

2.5 Missing Values

A problem that becomes immediately apparent when evaluating survey data is the omission of answers and frequent lack of opinion (i.e. responses like *I don't know*). The reasons can be various: deliberate refusal, missing information, respondent inability, indecision, etc.

Faulkenberry and Mason (1978, p. 533) distinguish between two main types of answer omissions:

(a) *No opinion*: respondents are indecisive about an answer (due to an ambiguous question, say).
(b) *Non-opinion*: respondents have no opinion about a topic.

The authors find that respondents who tend to give the first type of omission (no opinion) are more reflective and better educated than respondents who tend to give the second type of omission (non-opinion). They also note that the gender, age, and ethnic background of the respondents (among other variables) can influence the likelihood of an answer omission.

This observation brings us to the problem of *systematic bias* caused by answer omission. Some studies show that lack of opinion can be up to 30% higher when respondents are given the option of *I don't know* (Schumann & Presser, 1981, p. 117). But simply eliminating this option as a strategy for its avoidance can lead to biased results. This is because the respondents who tend to choose *I don't know* often do not feel obliged to give truthful answers when the *I don't know* option is not available. Such respondents typically react by giving a random answer or no answer at all. This creates the danger that an identifiable, systematic error attributable to frequent *I don't know* responses will be transformed into an undiscovered, systematic error at the level of actual findings. From this perspective, it is hard to understand those who recommend the elimination of the *I don't know* option. More important is the question of how to approach answer omissions during data analysis.

In principle, the omissions of answers should not lead to values that are interpreted during analysis, which is why some analysis methods do not permit the use of missing values. The presence of missing values can even necessitate that other data be excluded. In regression or factor analysis, for example, when a respondent has missing values, the remaining values for that respondent must be omitted as well. Since answer omissions often occur and no one wants large losses of information, the best alternative is to use some form of substitution. There are five general approaches:

(a) The best and most time-consuming way to eliminate missing values is to fill them in yourself, provided it is possible to obtain accurate information through further research. In many cases, missing information in questionnaires on revenue, R&D expenditures, etc., can be discovered through a careful study of financial reports and other published materials.

(b) If the variables in question are qualitative (nominally scaled), missing values can be avoided by creating a new class. Consider a survey in which some respondents check the box *previous customer*, some the box *not a previous customer*, and others check *neither*. In this case, the respondents who provided no answer can be assigned to a new class; let's call it *customer status unknown.* In the frequency tables, this class then appears in a separate line titled *missing values.*

(c) If it is not possible to address missing values conducting additional research or creating a new category, missing variables can be substituted with the total arithmetic mean of existing values, provided they are on a cardinal scale.

(d) Missing cardinal values can also be substituted with the arithmetic mean of a group. For instance, in a survey gathering statistics on students at a given university, missing information is better replaced by the arithmetic mean of students in the respective course of study rather than by the arithmetic mean of the entire student body.

(e) We must remember to verify that the omitted answers are indeed non-systematic; otherwise, attempts to compensate for missing values will produce grave distortions. When answers are omitted in non-systematic fashion, missing values can be estimated with relative accuracy. Nevertheless, care must be taken not to understate value distribution and, by extension, misrepresent the results. "In particular", note Roderick et al., "variances from filled-in data are clearly understated by imputing means, and associations between variables are distorted. Thus, the method yields an inconsistent estimate of the covariance matrix" (1995, p. 45). The use of complicated estimation techniques becomes necessary when the number of missing values is large enough that the insertion of mean values significantly changes the statistical indices. These techniques mostly rely on regression analysis, which estimates missing values using existing dependent variables in the dataset. Say a company provides incomplete information about their R&D expenditures. If you know that R&D expenditures depend on company sector, company size, and company location (West Germany or East Germany, for instance), you can use available data to roughly extrapolate the missing data. Regression analysis is discussed in more detail in Chap. 10.

Generally, you should take care when subsequently filling in missing values. Whenever possible, the reasons for the missing values should remain clear. In a telephone interview, for instance, you can distinguish between:

- respondents who do not provide a response because they do not know the answer;
- respondents who have an answer but do not want to communicate it; and
- respondents who do not provide a response because the question is directed to a different age group than theirs.

In the last case, an answer is frequently just omitted (missing value due to study design). In the first two cases, however, values may be assigned but are later defined as *missing values* by the analysis software.

2.6 Outliers and Obviously Incorrect Values

A problem similar to missing values is that of obviously incorrect values. Standardized customer surveys often contain both. Sometimes a respondent checks the box marked *unemployed* when asked about job status but enters some outlandish figure like €1,000,000,000 when asked about income. If this response were included in a survey of 500 people, the average income would increase by €2,000,000. This is why obviously incorrect answers must be eliminated from the dataset. Here, the intentionally wrong income figure could be marked as a missing value or given an estimated value using one of the techniques described in Sect. 2.5.

Obviously incorrect values are not always deliberate. They can also be the result of error. Business surveys, for instance, often ask for revenue figures in thousands of euros, but some respondents invariably provide absolute values, thus indicating revenues one-thousand times higher than they actually are. If discovered, mistakes like these must be corrected before data analysis.

A more difficult case is when the data are unintentionally false but cannot be easily corrected. For example, when you ask businesses to provide a breakdown of their expenditures by category and per cent, you frequently receive total values amounting to more than 100%. Similar errors also occur with private individuals.

Another tricky case is when the value is correct but an outlier. Suppose a company wants to calculate future employee pensions. To find the average retirement age, they average the ages at which workers retired in recent years. Now suppose that of one of the recent retirees, the company's founder left the business just shy of 80. Though this information is correct—and though the founder is part of the target group of retired employees—the inclusion of this value would distort the average retirement age, since it is very unlikely that other employees will also retire so late in the game. Under certain circumstances it thus makes sense to exclude outliers from the analysis—provided, of course, that the context warrants it. One general solution is to *trim* the dataset values, eliminating the highest and lowest five per cent. I will return to this topic once more in Sect. 3.2.2.

2.7 Chapter Exercises

Exercise 1
For each of the following statistical units, provide traits and trait values:

(a) Patient cause of death
(b) Length of university study
(c) Alcohol content of a drink

Exercise 2
For each of the following traits, indicate the appropriate level of measurement:

(a) Student part-time jobs
(b) Market share of a product between 0 and 100%
(c) Students' chosen programme of study
(d) Time of day
(e) Blood alcohol level
(f) Vehicle fuel economy
(g) IQ
(h) Star rating for a restaurant

Exercise 3
Use R, Stata, SPSS, or Excel for the questionnaire in Fig. 2.1 and enter the data from Fig. 3.1. Allow for missing values in the dataset.

2.8 Exercise Solutions

Solution 1

(a) Deceased patients; cause of death; heart attack, stroke, etc.
(b) Student; semester; 1st, 2nd etc.
(c) Type of beverage; alcohol content; 3%, 4%, etc.

Solution 2
(a) Nominal; (b) metric; (c) nominal; (d) interval scaled (metric); (e) ratio scaled (metric); (f) ratio scaled (metric); (g) ordinal; (h) ordinal

Solution 3
See the respective file "*Chapter2_spread_small*" at the book's supplementary material (see DOI: https://doi.org/10.1007/978-3-031-78070-7_2).

References

Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical Education, 42*, 1150–1152.
Faulkenberry, G. D., & Mason, R. (1978). Characteristics of nonopinion and no opinion response groups. *Public Opinion Quarterly, 42*, 533–543.
Malhotra, N. K., Nunan, D., & Birks, D. F. (2017). *Marketing research. An applied approach* (5th ed.). Pearson.
Pell, G. (2005). Use and misuse of Likert scales. *Medical Education, 39*, 970.

Roderick, J. A., Little, & Schenker, N. (1995). Missing data. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modelling for the social and behavioral sciences* (pp. 39–75). Plenum Press.

Schmidt, P., & Opp, K.-D. (1976). *Einführung in die Mehrvariablenanalyse.* Rowohlt.

Schumann, H., & Presser, S. (1981). *Questions and answers in attitude surveys.* Academic Press.



Univariate Data Analysis

3

3.1 First Step in Data Analysis

Let us return to our students from the previous chapter. After completing their survey of bread spreads, they have now coded the data from the 850 respondents and entered them into a computer (see dataset *Chapter3_spread* in the supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_3). In the first step of data assessment, they investigate each variable—for example, average respondent age—separately. This is called *univariate analysis* (see Fig. 3.1). By contrast, when researchers analyse the relationship between two variables—for example, between gender and choice of spread—this is called *bivariate analysis* (see Chap. 4). With relationships between more than two variables, one speaks of *multivariate analysis* (see Sect. 9.5 and Chaps. 10, 11, 13, 14).

How can the results of 850 responses be "distilled" to create a realistic and accurate impression of the surveyed attributes and their relationships? Here the importance of statistics becomes apparent. Recall the professor who was asked about the results of the last final exam. The students expect distilled information, e.g. "*the average score was 75%*" or "*the failure rate was 29.4%*". Based on this information, students believe they can accurately assess general performance: "*an average score of 75% is worse than the 82% average on the last final exam*". A single distilled piece of data—in this case, the average score—appears sufficient to sum up the performance of the entire class.[1]

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78070-7_3.

[1] It should be noted here that the student assessment assumes a certain kind of distribution. An average score of 75% is obtained whether all students receive a score of 75%, or whether half score

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*,
Springer Texts in Business and Economics,
https://doi.org/10.1007/978-3-031-78070-7_3

	index	gender	age	Bodyweight	spread	offer
1	1	male	31	63.1	butter	very poor
2	2	male	73	77.5	butter	very poor
3	5	male	45	82.1	butter	very poor
4	6	male	57	61.7	butter	very poor
5	9	male	38	36.5	butter	very poor
6	11	male	27	64.0	butter	very poor
7	12	male	36	70.9	butter	very poor
8	13	male	60	70.4	butter	very poor
9	15	male	21		butter	very poor
10	16	male	26			
11	18	male	55			
12	22	male	27		butter	very poor
13	25	male	30	72.7	butter	very poor
14	26	male	33	77.8	butter	very poor
15	27	male	33	90.8	butter	very poor
16	28	male	58	62.4	butter	very poor
17	29	male	23	91.2	butter	very poor

Analysis of only one variable: Univariate Analysis

**Using SPSS or Stata: The data editor can usually be set to display the codes or labels for the variables, though the numerical values are stored.

Fig. 3.1 Survey data entered in the data editor

This chapter and the next will describe methods of distilling data and their attendant problems. The above survey will be used throughout as an example.

Graphical representations or frequency tables can be used to create an overview of the univariate distribution of nominal- and ordinal-scaled variables. In the *frequency table* in Fig. 3.2, each variable trait receives its own line, and each line intersects the columns *absolute frequency*, *relative frequency [in %]*,[2] *valid percentage values*, and *cumulative percentage*. The relative frequency of trait x_i is abbreviated algebraically by $f(x_i)$. Any missing values are indicated in a separate line with a percentage value. Missing values are not included in the calculations of *valid percentage values*[3] and *cumulative percentage*. The cumulative percentage reflects the sum of all rows up to and including the row in question. The figure of 88.1% given for the rating *average* in Fig. 3.2 indicates that 88.1% of the

50% and the other half score 100%. Although the average is the same, the qualitative difference in these two results is obvious. Average alone, therefore, does not suffice to describe the results.

[2] Relative frequency ($f(x_i)$) equals the absolute frequency ($h(x_i)$) relative to all valid and invalid observations ($N = N_{\text{valid}} + N_{\text{invalid}}$): $f(x_i) = h(x_i)/N$.

[3] Valid percentage ($gf(x_i)$) equals the absolute frequency ($h(x_i)$) relative to all valid observations (N_{valid}): $g(x_i) = h(x_i)/N_{\text{valid}}$.

	Absolute frequency	Relative frequency [in %]	Valid percentage values	Cumulative percentage
Poor	391	46.0	46.0	46.0
Fair	266	31.3	31.3	77.3
Average	92	10.8	10.8	88.1
Good	62	7.3	7.3	95.4
Excellent	39	4.6	4.6	100.0
Total	850	100.0	100.0	

Fig. 3.2 Frequency table for selection ratings

respondents described the selection as average or worse. Algebraically, the cumulative frequencies are expressed as a *distribution function*, abbreviated $F(x)$, and calculated as follows:

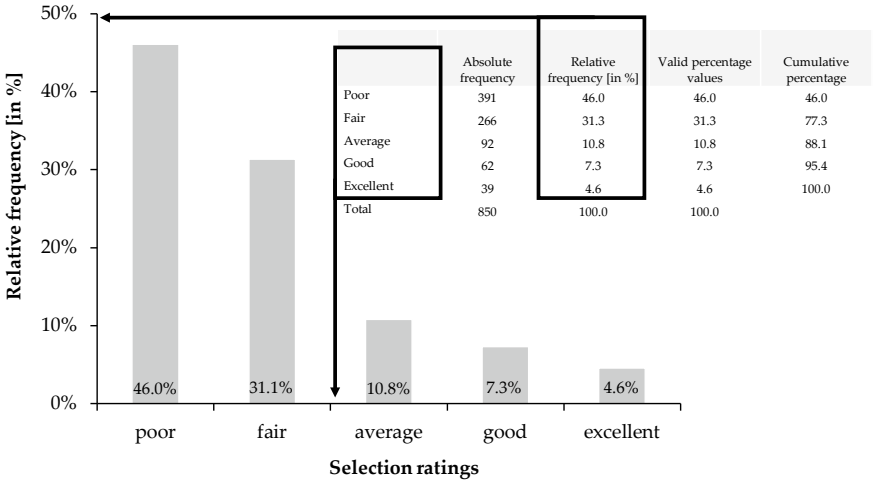
$$F\left(x_p\right) = f\left(x_1\right) + f\left(x_2\right) + \cdots + f\left(x_p\right) = \sum_{i=1}^{p \leq n} f\left(x_i\right) \tag{3.1}$$

These results can also be represented graphically as a *pie chart*, a *horizontal bar chart*, or a *vertical bar chart*. All three diagram forms can be used with nominal and ordinal variables, though pie charts are used mostly for nominal variables.

The traits of the frequency table in the bar chart (poor, fair, average, good, excellent) are assigned to the x-axis and the relative or absolute frequency to the y-axis. The height of a bar equals the frequency of each x-value. If the relative frequencies are assigned to the y-axis, a graph of the frequency function is obtained (see Fig. 3.3).

In addition to the frequency table, we can also represent the distribution of an ordinally scaled variable (or higher) using the $F(x)$ distribution function. This function leaves the traits of the x-variables in question on the x-axis, and assigns the cumulative percentages to the y-axis, generating a *step function*. The data representation is analogous to the column with cumulative percentages in the frequency table (see Fig. 3.4).

In many publications, the scaling on the y-axis of a vertical bar chart begins not with zero but with some arbitrary value. As Fig. 3.5 shows, this can lead to a misunderstanding at first glance. Both graphs represent the same content—the relative frequency of male and female respondents (49% and 51%, respectively). But because the y-axis is cut off in the first graph, the relative frequency of the genders appears to change. The first graph appears to show a relationship of five females to one male, suggesting that there are five times as many female observations as male observations in the sample. The interval in the first graph is misleading—a problem we'll return to below—so that the difference of two percentage points



	Absolute frequency	Relative frequency [in %]	Valid percentage values	Cumulative percentage
Poor	391	46.0	46.0	46.0
Fair	266	31.3	31.3	77.3
Average	92	10.8	10.8	88.1
Good	62	7.3	7.3	95.4
Excellent	39	4.6	4.6	100.0
Total	850	100.0	100.0	

Fig. 3.3 Bar chart/frequency distribution for the selection variable

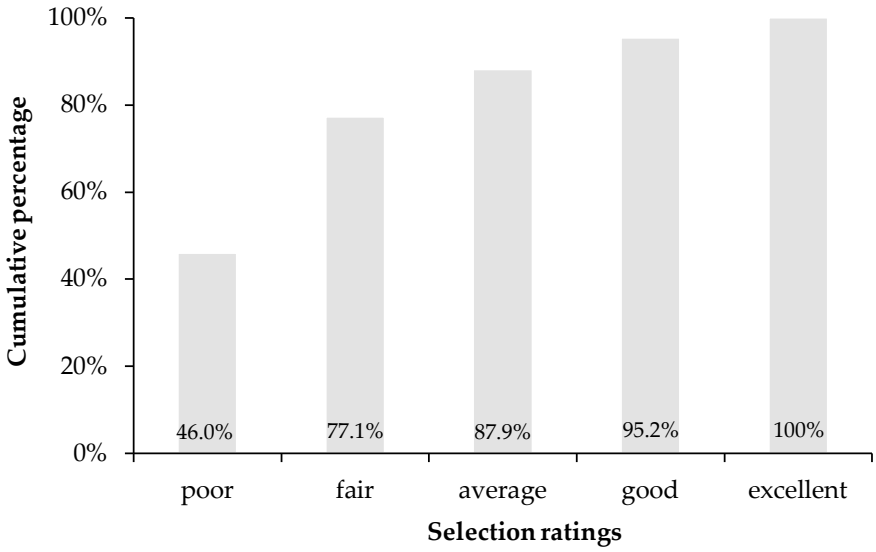
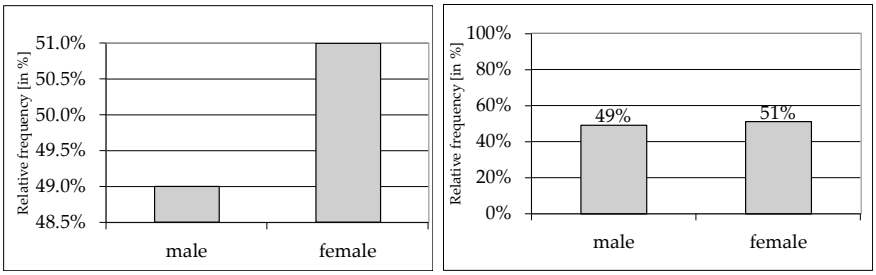


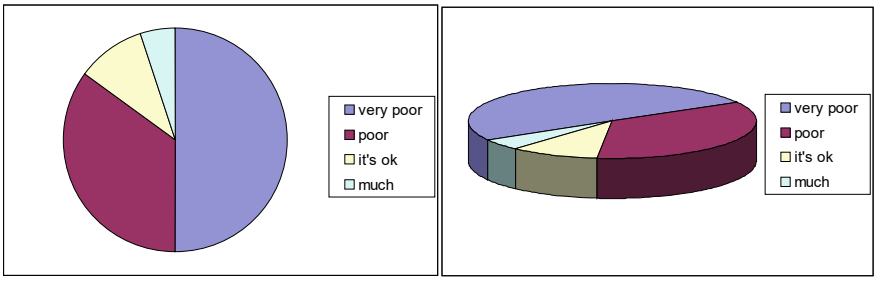
Fig. 3.4 Distribution function for the selection variable

seems larger than it actually is. For this reason, the second graph in Fig. 3.5 is the preferable form of representation.



Part 1 Part 2

Fig. 3.5 Different representations of the same data (1)



Part 1 Part 2

Fig. 3.6 Different representations of the same data (2)

Similar distortions can arise when two alternate forms of a pie chart are used. In the first chart in Fig. 3.6, the size of each wedge represents relative frequency. The chart is drawn by weighting the circle segment angles such that each angle $\alpha_i = f(x_i) \cdot 360°$.

Since most viewers read pie charts clockwise from the top, the traits to be emphasized should be placed in the twelve o'clock position whenever possible. Moreover, the chart shouldn't contain too many segments—otherwise the graph will be hard to read. They should also be ordered by some system—for example, by size or content (Krämer, 2008).

The second graph in Fig. 3.6, which is known as a "perspective" or "3D" pie chart, looks more modern, but the downside is that the area of each wedge no longer reflects relative frequency. The representation is thus somewhat misleading. The pie chart segments in the foreground seem larger. The edge of the pie segments in the front can be seen, but not those in the back. The "lifting up" of a particular wedge can amplify this effect even more.

And what of cardinal variables? How should they be represented? The novice might attempt to represent bodyweight using a vertical bar diagram—as shown in graph 1 of Fig. 3.7. But the variety of possible traits generates too many bars, and

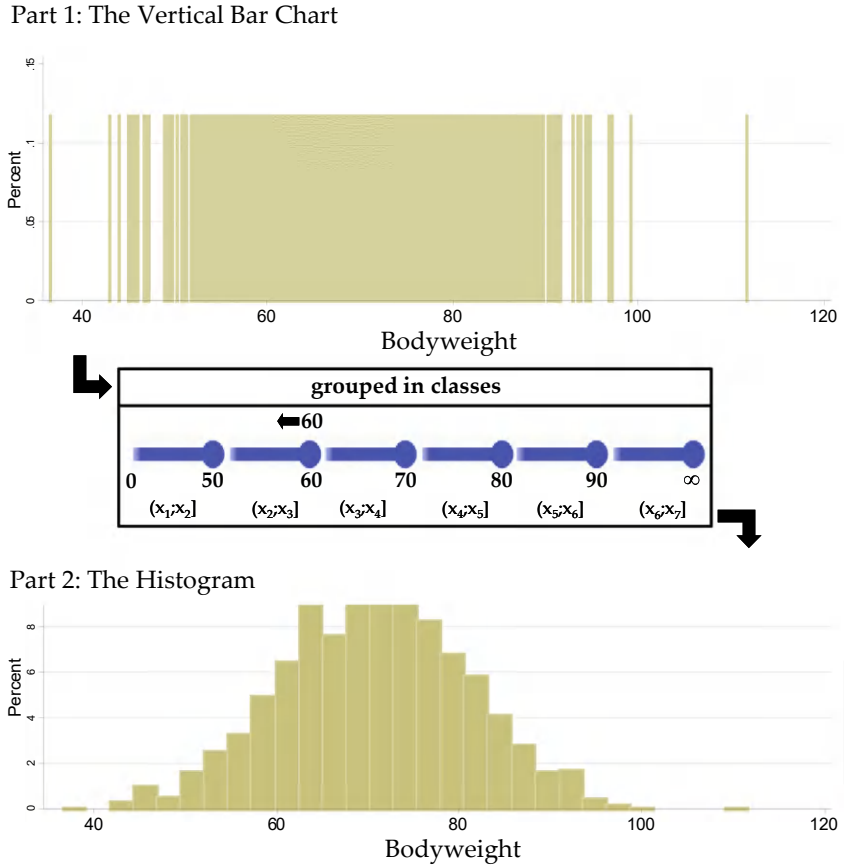


Fig. 3.7 Using a histogram to classify data

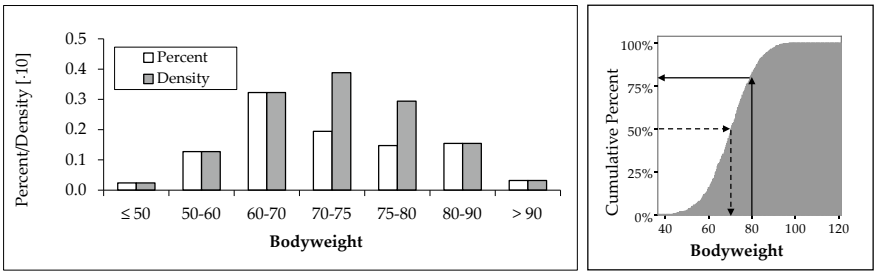
their heights rarely vary. Frequently, a trait appears only once in a collection of cardinal variables. In such cases, the goal of presenting all the basic relationships at a glance is destined to fail. For this reason, the individual values of cardinal variables should be grouped in classes, or classed. Bodyweight, for instance, could be assigned to the classes shown in Fig. 3.7.[4]

By standard convention, the upper limit value in a class belongs to that class; the lower limit value does not. Accordingly, persons who are 60 kg belong to the 50–60 kg group, while those who are 50 kg belong to the class below. Of course, it is up to the persons assessing the data to determine class size and class membership at the boundaries. When working with data, however, one should clearly indicate the decisions made in this regard.

[4] For each *i*th class, the following applies: $x_i < X \leq x_{i+1}$ with $i \in \{1, 2, \ldots, k\}$.

A *histogram* is a classed representation of cardinal variables. What distinguishes the histogram from other graphic representations is that it expresses relative class frequency not by height but by area (height · width). The height of the bars represents frequency density. The denser the bars are in the bar chart in part 1 of Fig. 3.7, the more observations there are for that given class and the greater its frequency density. As the frequency density for a class increases, so too does its area (height · width). The histogram obeys the principle that the intervals in a diagram should be selected so that the data are not distorted. In the histogram, the share of area for a specific class relative to the entire area of all classes equals the relative frequency of the specific class. To understand why the selection of suitable intervals is so important consider part 1 of Fig. 3.8, which represents the same information as Fig. 3.7 but uses unequal class widths. In a vertical bar chart, height represents relative frequency. The white bars in the figure represent relative frequency. The graph appears to indicate that a bodyweight between 60 and 70 kg is the most frequent class. Above this range, frequency drops off before rising again slightly for the 80–90 kg class. This impression is created by the distribution of the 70–80 kg group into two classes, each with a width of five kg, or half that of the others. If the data are displayed without misleading intervals, the frequency densities can be derived from the grey bars. With the same number of observations in a class, the bars would only be the same height if the classes were equally wide. By contrast, with a class half as large and the same number of observations, the observations will be twice as dense. Here we see that, in terms of class width, the density for the 70–75 kg range is the largest.

It would be useful if the histogram's differences in class width were indicated to scale by different widths on the *x*-axis. Unfortunately, no currently available statistics or graphical software can perform this function. Instead, they avoid the problem by permitting equal class widths only.



Part 1 Part 2

Fig. 3.8 Distorting interval selection with a distribution function

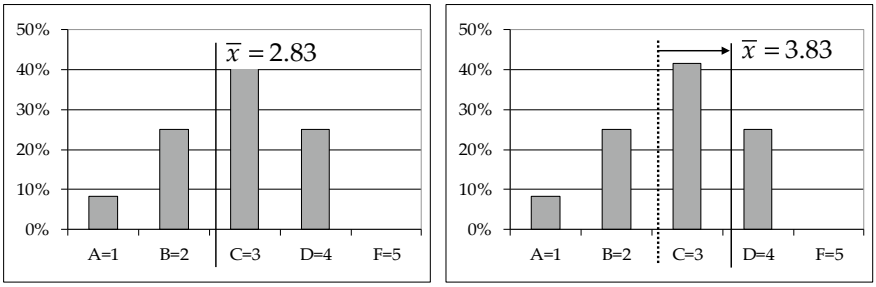


Fig. 3.9 Grade averages for two final exams

The distribution function of a cardinal variable can be represented as unclassed. Here too, the frequencies are cumulative as one moves along the *x*-axis. The values of the distribution function rise evenly and remain between zero and one. The distribution function for the bodyweight variable is represented in part 2 of Fig. 3.8. Here, one can obtain the cumulated percentages for a given bodyweight and vice versa. Some 80% of the respondents are 80 kg or under, and 50% of the respondents are 70 kg or under.

3.2 Measures of Central Tendency

The previous approach allowed us to reduce the diversity of information from the questionnaires—in our sample there were 850 responses—by creating graphs and tables with just a few lines, bars, or pie wedges. But how and under which conditions can this information be reduced to a single number or measurement that summarizes the distinguishing features of the dataset and permits comparisons with others? Consider again the student who, to estimate the average score on the last final exam, looks for a single number—the average grade or failure rate. The average score for two final exams is shown in Fig. 3.9.[5]

Both final exams have an identical distribution; in the second graph (part 2), this distribution is shifted one grade to the right on the *x*-axis. This shift represents a mean value one grade higher than the first exam. Mean values or similar parameters that express a general trend of a distribution are called *measures of central tendency*. Choosing the most appropriate measure usually depends on context and the level of measurement.

[5] The grade scale is taken here to be cardinal scaled. This assumes that the difference in scores between A and B is identical to the difference between B and C, etc. But because this is unlikely in practice, school grades, strictly speaking, must be seen as ordinal scaled.

3.2.1 Mode or Modal Value

The most basic measure of central tendency is known as the *mode* or *modal value*. The mode identifies the value that appears most frequently in a distribution. In part 1 of Fig. 3.9, the mode is the grade C. The mode is the "champion" of the distribution. Another example is the item selected most frequently from five competing products. This measure is particularly important with voting, though its value need not be clear. When votes are tied, there can be more than one modal value. Most software programmes designate only the smallest trait. When values are far apart, this can lead to misinterpretation. For instance, when a cardinal variable for age and the traits 18 and 80 appear in equal quantities and more than all the others, many software packages still indicate the mode as 18.

3.2.2 Mean

The *arithmetic mean*—colloquially referred to as the *average*—is calculated differently depending on the nature of the data. In empirical research, data most frequently appear in a raw data table that includes all the individual trait values. For raw data tables, the mean is derived from the formula:

$$\frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.2}$$

All values of a variable are added and divided by n. For instance, given the values 12, 13, 14, 16, 17, and 18 the mean is

$$\bar{x} = \frac{1}{6}(12 + 13 + 14 + 16 + 17 + 18) = 15. \tag{3.3}$$

The mean can be represented as a balance scale (see Fig. 3.10), and the deviations from the mean can be regarded as weights. If, for example, there is a deviation of (− 3) units from the mean, then a weight of three grams is placed on the left side of the balance scale. The further a value is away from the mean, the heavier the weight. All negative deviations from the mean are placed on the left side of the mean, and all positive deviations on the right. The scale is exactly balanced. With an arithmetic mean, the sum of negative deviations equals the sum of positive deviations:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0 \tag{3.4}$$

In real life, if a heavy weight is on the one side of the scale and many smaller weights are on the other, the scale can still be balanced (cf. Fig. 3.10). But the

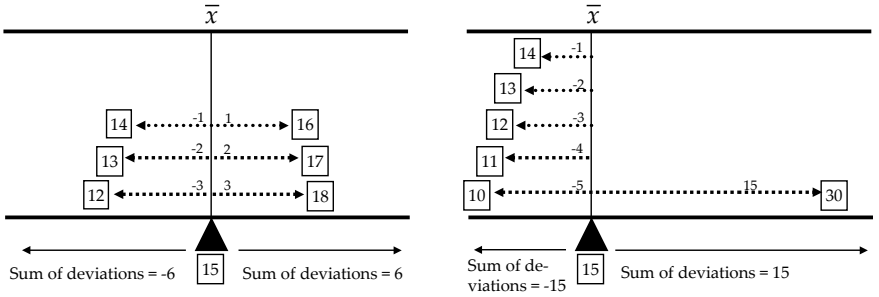


Fig. 3.10 Mean expressed as a balanced scale

mean is not a good estimate for this kind of distribution: it could over- or underestimate the many smaller weights. We encountered this problem in Sect. 2.6; in such cases, an outlier value is usually responsible for distorting the results. Assume you want to calculate the central tendency of the age of animals in a zoo terrarium containing five snakes, nine spiders, five crocodiles, and one turtle (see also the dataset *Chapter3_zoo* in the supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_3). The last animal—the turtle—is 120 years old, while all the others are no older than four (see Fig. 3.11).

Based on these ages, the mean would be 7.85 years. To "balance" the scale, the ripe old turtle would have to be alone on the right side, while all the other animals are on the left side. We find that the mean value is a poor measure to describe the average age in this case because only one other animal is older than three. To reduce or eliminate the outlier effect, practitioners frequently resort to a *trimmed mean*. This technique "trims" the smallest and largest 5% of values before calculating the mean, thus partly eliminating outliers. In our example, the 5% trim covers both the youngest and oldest observation (the terrarium has 20 animals), thereby eliminating the turtle's age from the calculation. This results in an average age of two years, a more realistic description of the age distribution. We should

		Age					
		1	2	3	4	120	Total
Animal	Snake	2	1	1	1	0	5
	Turtle	0	0	0	0	1	1
	Crocodile	1	2	2	0	0	5
	Spider	4	4	1	0	0	9
Total		7	7	4	1	1	20

Mean=7.85 years; 5% trimmed mean=2 years.

Fig. 3.11 Mean or trimmed mean using the zoo example

remember, however, that this technique eliminates 10% of the observations, and this can cause problems, especially with small samples.

Let us return to the "normal" mean, which can be calculated from a frequency table (such as an overview of grades) using the following formula:

$$\bar{x} = \frac{1}{n} \sum_{v=1}^{k} x_v \cdot n_v = \sum_{v=1}^{k} x_v \cdot f_v \tag{3.5}$$

We will use the frequency table in Fig. 3.2 as an example. Here the index v runs through the different traits of the observed ordinal variables for selection (*poor*, *fair*, *average*, *good*, *excellent*). The value n_v equals the absolute number of observations for a trait. The trait *good* yields a value of $n_v = n_4 = 62$. The variable x_v assumes the trait value of the index v. The trait *poor* assumes the value $x_1 = 1$, the trait *fair* the value $x_2 = 2$, etc. The mean can be calculated as follows:

$$\bar{x} = \frac{1}{850} \cdot (391 \cdot 1 + 266 \cdot 2 + 92 \cdot 3 + 62 \cdot 4 + 39 \cdot 5) = 1.93 \tag{3.6}$$

The respondents gave an average rating of 1.93, which approximately corresponds to *fair*. The mean could also have been calculated using the relative frequencies of the traits f_v:

$$\bar{x} = (0.46 \cdot 1 + 0.313 \cdot 2 + 0.108 \cdot 3 + 0.073 \cdot 4 + 0.046 \cdot 5) = 1.93 \tag{3.7}$$

Finally, the mean can also be calculated from traditional classed data according to this formula:

$$\bar{x} = \frac{1}{n} \sum_{v=1}^{k} n_v m_v = \sum_{v=1}^{k} f_v m_v, \text{ where } m_v \text{ is the mean of class number } v. \tag{3.8}$$

Students often confuse this with the calculation from frequency tables, as even the latter contain classes of traits. With classed data, the mean is calculated from cardinal variables that are summarized into classes by making certain assumptions. In principle, the mean can be calculated this way from a histogram. Consider again Fig. 3.7. The calculation of the mean bodyweight in part 1 agrees with the calculation from the raw data table. But what about when there is no raw data table, only the information in the histogram, as in part 2 of Fig. 3.7? Figure 3.12 shows a somewhat more simplified representation of a histogram with only six classes.

We start from the implicit assumption that all observations are distributed evenly within a class. Accordingly, cumulated frequency increases linearly from the lower limit to the upper limit of the class. Here class frequency average necessarily equals the mean. To identify the total mean, add all products from the class midpoint and the attendant relative frequencies.

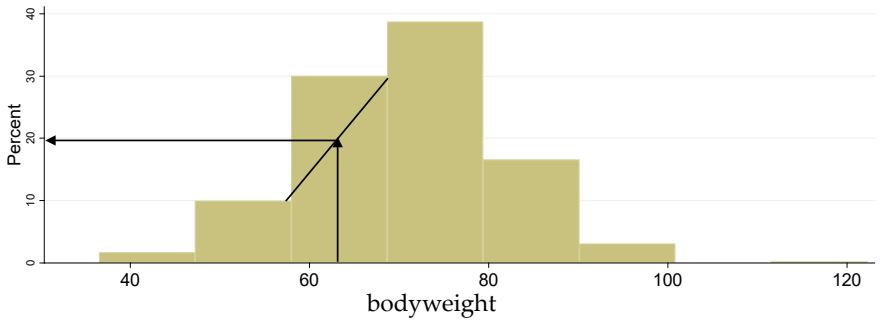


Fig. 3.12 Calculating the mean from classed data

Table 3.1 Example of mean calculation from classed data

Water use [in l]	0–200	200–400	400–600	600–1000
Rel. frequency	0.2	0.5	0.2	0.1

Source Schwarze (2013, p. 16), translated from the German

Here is another example to illustrate the calculation. Consider the following information on water use by private households in Table 3.1.

The water-use average can be calculated as follows:

$$\bar{x} = \sum_{v=1}^{k} f_v m_v = \sum_{v=1}^{4} f_v m_v = 0.2 \cdot 100 + 0.5 \cdot 300 + 0.2 \cdot 500 + 0.1 \cdot 800 = 350 \quad (3.9)$$

With all formulas calculating the mean, we assume equidistant intervals between the traits. This is why the mean cannot be determined for nominal variables. This is also why, strictly speaking, no mean can be calculated for ordinal variables. But this is only true if one takes a dogmatic position. Practically minded researchers who possess sufficiently large samples (approx. $n > 99$) often calculate the mean by assuming equidistance.

The informational value of the mean was previously demystified in Sect. 3.2.1 using the example of average test grades. An average grade of C occurs when all students receive C. The same average results when half of the students receive an A and the other half an F. The same kind of problem could result by selecting travel destinations based on temperature averages. Beijing, Quito, and Milan all have an average temperature of 12°C, but the experience of temperature in the three cities varies greatly. The winter in Beijing is colder than in Stockholm and the summer is hotter than in Rio de Janeiro. In Milan, the temperatures are Mediterranean, fluctuating seasonally, while the altitude in Quito ensures that the temperature stays pretty much the same the whole year over (Swoboda, 1971, p. 36).

The average is not always an information-rich number that uncovers all that remains hidden in tables and figures. When no information can be provided on

distribution (e.g. average deviation from average) or when weightings and reference values are withheld, the average can also be misleading. The list of amusing examples is long, as described by Krämer (2015). Here are a few:

- Means rarely result in whole numbers. For instance, what do we mean by the decimal place when we talk of 1.7 children per family or 3.5 sexual partners per person?
- When calculating the arithmetic mean, all values are treated equally. Imagine a proprietor of an eatery in the Wild West who, when asked about the ingredients of his stew, says: *Half and half. One horse and one jackrabbit*. It is not always accurate to consider the values as equal in weight. The cook might advertise his concoction as a *wild game stew*, but if the true weights of the inputs were taken into account, it would be more accurately described as horse goulash. Consider an example from the economy: if the average female salary is 20 MUs (monetary units) and the average male salary is 30 MUs, the average employee salary is not necessarily 25 MUs. If males constitute 70% of the workforce, the average salary will be: $0.7 \cdot 30\ \text{MU} + 0.3 \cdot 20\ \text{MU} = 27\ \text{MU}$. One speaks here of a *weighted arithmetic mean* or a *scaled arithmetic mean.* The Federal Statistical Office of Germany calculates the rate of price increase for products in a basket of commodities in a similar fashion. The price of a banana does not receive the same weight as the price of a vehicle; its weight is calculated based on its average share in a household's consumption.
- The choice of *reference base*—i.e. the dominator for calculating the average—can also affect the interpretation of data. Take the example of traffic deaths. Measured by deaths per passenger-kilometres travelled, trains have a rate of nine traffic deaths per ten billion kilometres travelled and planes three deaths per ten billion kilometres travelled. Airlines like to cite these averages in their ads. But if we consider traffic deaths not in relation to distance but in relation to time of travel, we find completely different risks. For trains, there are seven fatalities per 100 million passenger-hours, and for planes there are 24 traffic deaths per 100 million passenger-hours. Both reference bases can be asserted as valid. The job of empirical researchers is to explain their choice. Although I have a fear of flying, I agree with Krämer (2015) when he argues that passenger-hours is a better reference base. Consider the following: few of us are scared of going to bed at night, yet the likelihood of dying in bed is nearly 99%. Of course, this likelihood seems less threatening when measured against the time we spend in bed.

3.2.3 Geometric Mean

The above problems frequently result from a failure to apply weightings or by selecting a wrong or poor reference base. But sometimes the arithmetic mean as a measure of general tendency can lead to faulty results even when the weighting and reference base are appropriate. This is especially true in economics when measuring rates of change or growth. These rates are based on data observed over time, which is why such data are referred to as time series. Figure 3.13 shows an example of sales and their rates of change over five years.

Using the arithmetic mean to calculate the average rate of change yields a value of 1.25%. This would mean that yearly sales have increased by 1.25%. Based on this growth rate, the €20,000 in sales in 2014 should have increased to €21,018.91 by 2018, but actual sales in 2018 were €20,691.00. Here we see how calculating average rates of change using arithmetic mean can lead to errors. This is why the *geometric mean for* rates of change is used. In this case, the parameter links initial sales in 2014 with the subsequent rates of growth each year until 2018. The result is:

$$\begin{aligned} U_6 &= U_5 \cdot (1 + 0.1) = (U_4 \cdot (1 - 0.1)) \cdot (1 + 0.1) = \ldots \\ &= (U_2 \cdot (1 + 0.1)) \cdot (1 - 0.05) \cdot (1 - 0.1) \cdot (1 + 0.1). \end{aligned} \tag{3.10}$$

To calculate the average change in sales from this chain, the four rates of change (1 + 0.1) · (1–0.05) · (1–0.1) · (1 + 0.1) must yield the same value as the fourfold application of the average rate of change:

$$\left(1 + \overline{p}_{\text{geom}}\right) \cdot \left(1 + \overline{p}_{\text{geom}}\right) \cdot \left(1 + \overline{p}_{\text{geom}}\right) \cdot \left(1 + \overline{p}_{\text{geom}}\right) = \left(1 + \overline{p}_{\text{geom}}\right)^4 \tag{3.11}$$

For the geometric mean, the yearly rate of change is thus:

$$\overline{p}_{\text{geom}} = \sqrt[4]{(1 + 0.1)(1 - 0.05)(1 - 0.1)(1 + 0.1)} - 1 = 0.853 = 0.853 \tag{3.12}$$

			Changes in sales when using	
Year	Sales [mio.]	Rate of change [in %]	arithm. mean	geom. mean
2014	€20,000.00		€20,000.00	€20,000.00
2015	€22,000.00	10.000%	€20,250.00	€20,170.56
2016	€20,900.00	-5.000%	€20,503.13	€20,342.57
2017	€18,810.00	-10.000%	€20,759.41	€20,516.04
2018	€20,691.00	10.000%	€21,018.91	€20,691.00
Arithmetic mean		1.250%		
Geometric mean		0.853%		

Fig. 3.13 An example of geometric mean

The last column in Fig. 3.13 shows that this value correctly describes the sales growth between 2014 and 2018. Generally, the following formula applies for identifying *average rates of change*:

$$\overline{p}_{\text{geom}} = \sqrt[n]{(1+p_1)\cdot(1+p_2)\cdot\ldots\cdot(1+p_n)} - 1 = \sqrt[n]{\prod_{i=1}^{n}(1+p_i)} - 1 \quad (3.13)$$

The geometric mean for rates of change is a special instance of the *geometric mean* and is defined as follows:

$$\overline{x}_{\text{geom}} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i} \quad (3.14)$$

The geometric mean equals the arithmetic mean of the logarithms[6] and is only defined for positive values. For observations of different sizes, the geometric mean is always smaller than the arithmetic mean.

3.2.4 Harmonic Mean

A measure seldom required in economics is the so-called harmonic mean. Because of the rarity of this measure, researchers tend to forget it and instead use the arithmetic mean. However, sometimes the arithmetic mean produces false results. The harmonic mean is the appropriate method for averaging ratios consisting of numerators and denominators (unemployment rates, sales productivity, kilometres per hour, price per litre, people per square metre, etc.) when the values in the numerator are not identical. Consider, for instance, the sales productivity (as measured in revenue per employee) of three companies with differing headcounts but identical revenues. The data are given in Table 3.2.

To compare the companies, we should first examine the sales productivity of each firm regardless of its size. Every company can be taken into account with a simple weighted calculation. We find average sales per employee as follows:

$$\overline{x} = \frac{1}{3}\left(\frac{S_1}{E_1} + \frac{S_2}{E_2} + \frac{S_3}{E_3}\right) = €433.33 \quad (3.15)$$

If this value were equally applicable to all employees, the firms—which have 16 employees together—would have sales totalling $16 \cdot €433.33 \approx €6933$, but the table shows that actual total sales are only €3000. When calculating company sales, it must be taken into account that the firms employ varying numbers

[6] If all values are available in logarithmic form, the following applies to the arithmetic mean: $\frac{1}{n}(\ln(x_1) + \cdots + \ln(x_n)) = \frac{1}{n}\ln(x_1 \cdot \ldots \cdot x_n) = \ln(x_1 \cdot \ldots \cdot x_n)^{\frac{1}{n}} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \overline{x}_{\text{geom}}$.

Table 3.2 Harmonic mean

	Sales	Employees	Sales per employee (SP)	Formula in Excel
Company 1	€1000	10	€100.00	
Company 2	€1000	5	€200.00	
Company 3	€1000	1	€1000.00	
Sum	€3000	16	€1300.00	SUM(D3:D5)
Arithmetic mean			€433.33	AVERAGE(D3:D5)
Harmonic mean			€187.50	HARMEAN(D3:D5)

of employees and that the employees contribute in different ways to total productivity. This becomes clear from the fact that companies with equal sales (identical numerators) have different headcounts and hence different values in the denominator. To identify the contribution made by each employee to sales, one must weight the individual observations ($i = 1, \ldots, 3$) of sales productivity (SP_i) with the number of employees (n_i), add them and then divide by the total number of employees. The result is an arithmetic mean weighted by the number of employees:

$$\frac{n_1 \cdot \mathrm{SP}_1 + n_2 \cdot \mathrm{SP}_2 + n_3 \cdot \mathrm{SP}_3}{n} = \frac{10 \cdot €100}{16} + \frac{5 \cdot €200}{16} + \frac{1 \cdot €1000}{16} \approx €187.50 \tag{3.16}$$

Using this formula, the 16 employees generate the real total sales figure of €3000. If the weighting for the denominator (i.e. the number of employees) is unknown, the value for $k = 3$ sales productivity must be calculated using *an unweighted harmonic mean*:

$$\bar{x}_{\mathrm{harm}} = \frac{k}{\sum_{i=1}^{k} \frac{1}{x_i}} = \frac{k}{\sum_{i=1}^{k} \frac{1}{\mathrm{SP}_i}} = \frac{3}{\frac{1}{€100} + \frac{1}{€200} + \frac{1}{€1000}} = \frac{€187.50}{\mathrm{Employee}} \tag{3.17}$$

Let's look at another example that illustrates the harmonic mean. A student must walk three kilometres to his university campus by foot. Due to the nature of the route, he can walk the first kilometre at 2 km/h, the second kilometre at 3 km/h, and the last kilometre at 4 km/h. As in the last example, the arithmetic mean yields the wrong result:

$$\bar{x} = \frac{1}{3}\left(2\,\frac{\mathrm{km}}{\mathrm{h}} + 3\,\frac{\mathrm{km}}{\mathrm{h}} + 4\,\frac{\mathrm{km}}{\mathrm{h}}\right) = 3\,\frac{\mathrm{km}}{\mathrm{h}}, \tag{3.18}$$

or one hour walk to his campus. But if we break down the route by kilometre, we get 30 min for the first kilometre, 20 min for the second kilometre, and 15 min for the last kilometre. The durations indicated in the denominator vary by route segment, resulting in a total of 65 min walk to his campus. The weighted average

speed is thus 2.77 km/h.[7] This result can also be obtained using the harmonic mean formula and $k = 3$ for the route segments:

$$\bar{x}_{\text{harm}} = \frac{k}{\sum_{i=1}^{k} \frac{1}{x_i}} = \frac{3}{\frac{1}{2\,\frac{\text{km}}{\text{h}}} + \frac{1}{3\,\frac{\text{km}}{\text{h}}} + \frac{1}{4\,\frac{\text{km}}{\text{h}}}} = 2.77\,\frac{\text{km}}{\text{h}} \tag{3.19}$$

In our previous examples, the values in the numerator were identical for every observation. In the first example, all three companies had sales of €1000, and in the second example all route segments were one kilometre. If the values are not identical, the *unweighted harmonic mean* must be calculated. For instance, if the $k = 3$ companies mentioned previously had sales of $n_1 = €1000$, $n_2 = €2000$, and $n_3 = €5000$, we would use the following calculation:

$$\bar{x}_{\text{harm}} = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}} = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{\text{SP}_i}} = \frac{€1000 + €2000 + €5000}{\frac{€1000}{€100} + \frac{€2000}{€200} + \frac{€5000}{€1000}} = \frac{€500}{\text{Employee}} \tag{3.20}$$

As we can see here, the unweighted harmonic mean is a special case of the weighted harmonic mean.

Fractions do not always necessitate the use of the harmonic mean. For example, if the calculation involving the route to the university campus included different times instead of different segments, the arithmetic mean should be used to calculate the average speed. If one student walked an hour long at two km/h, a second hour at three km/h, and the last hour at four km/h, the arithmetic mean yields the correct average speed. Here the size of the denominator (time) is identical and yields the value of the numerator (i.e. the length of the partial route):

$$\bar{x} = \frac{1}{3}\left(2\,\frac{\text{km}}{\text{h}} + 3\,\frac{\text{km}}{\text{h}} + 4\,\frac{\text{km}}{\text{h}}\right) = 3\,\frac{\text{km}}{\text{h}} \tag{3.21}$$

The harmonic mean must be used when: (1) ratios are involved and (2) relative weights are indicated by numerator values (e.g. km). If the relative weights are given in the units of the denominator (e.g. hours), the arithmetic mean should be used. It should also be noted that the harmonic mean—like the geometric mean—is only defined for positive values greater than zero. For unequally sized observations, the following applies:

$$\bar{x}_{\text{harm}} < \bar{x}_{\text{geom}} < \bar{x} \tag{3.22}$$

[7] (30 min · 2 km/h + 20 min · 3 km/h + 15 min · 4 km/h)/65 min = 2.77 km/h.

Table 3.3 Share of sales by age class for diaper users

Age class	Under 1	1	2–4	5–10	11–60	61–100
Relative frequency (%)	30	15	25	4	3	23
Cumulated: $F(x)$ (%)	30	45	70	74	77	100

3.2.5 The Median

As the mean is sometimes not "representative" to measure the central tendency of a distribution, an alternative is required to identify the central tendency. Consider the following example: you work at an advertising agency and must determine the average age of diaper users for a diaper ad. You collect the data from Table 3.3.

Based on what we learned above about calculating the mean using the class midpoint of classed data, we get[8]:

$$\begin{aligned}\bar{x} &= 0.3 \cdot 0.5 + 0.15 \cdot 1.5 + 0.25 \cdot 3.5 + 0.04 \cdot 8 + 0.03 \cdot 36 + 0.23 \cdot 81 \\ &\approx 21 \text{ years.}\end{aligned} \tag{3.23}$$

This would mean that the average diaper user is college age! This is doubtful, of course, and not just because of the absence of baby-care rooms at universities. The high values on the outer margins—classes 0–1 and 61–100—create a bimodal distribution and paradoxically produce a mean in the age class in which diaper use is lowest.

So what other methods are available for calculating the central tendency of the age of diaper users? Surely one way would be to find the modal value of the most important age group: 0–1. Another option is the so-called *median*. This value not only offers better results in such cases. The median is also the value that divides the size-ordered dataset into two equally large halves. Exactly 50% of the values are smaller and 50% of the values are larger than the median.[9]

Figure 3.14 shows five weights ordered by *heaviness*. The median is $\tilde{x} = x_{0.5} = x_{(3)} = 9$, as 50% of the weights are to the left and right of weight number three.

There are several formulas for calculating the median. When working with a raw data tables—i.e. with unclassed data—most statistics textbooks suggest these formulas:

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)} \tilde{x} = x_{\left(\frac{n+1}{2}\right)} \text{ for an odd number of observations } (n) \tag{3.24}$$

[8] To find the value for the last class midpoint, take half the class width—$(101 - 61)/2 = 20$—and from that we get $61 + 20 = 81$ years for the midpoint.

[9] Strictly speaking, this only applies when the median lies between two observations, which is to say, only when there are an even number of observations. With an odd number of observations, the median corresponds to a single observation. In this case, 50% of $(n - 1)$ observations are smaller and 50% of $(n - 1)$ observations are larger than the median.

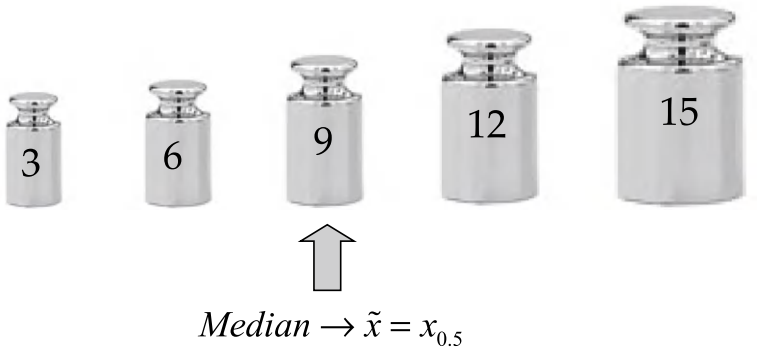


Fig. 3.14 The median: the central value of unclassed data

and

$$\tilde{x} = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) \text{ for an even number of observations.} \tag{3.25}$$

If one plugs in the weights from the example into the first formula, we get:

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{5+1}{2}\right)} = x_{(3)} = 9 \tag{3.26}$$

The trait of the weight in the third position of the ordered dataset equals the median. If the median is determined from a classed dataset, as in our diaper example, the following formula applies:

$$\tilde{x} = x_{0.5} = x_{i-1}^{\text{UP}} + \frac{0.5 - F\left(x_{i-1}^{\text{UP}}\right)}{f(x_i)}\left(x_i^{\text{UP}} - x_i^{\text{LOW}}\right) \tag{3.27}$$

First we identify the class in which 50% of observations are just short of being exceeded. In our diaper example, this corresponds to the one-year-olds. The median is above the upper limit x_{i-1}^{UP} of the class, or one year. But how many years above the limit? There is a difference of 5 percentage points between the postulated value of 0.5 and the upper limit value of $F\left(x_{i-1}^{\text{UP}}\right) = 0.45$:

$$0.5 - F\left(x_{i-1}^{\text{UP}}\right) = 0.5/0.45 = 0.05 \tag{3.28}$$

This five percentage points must be accounted for from the next largest (*i*th) class, as it must contain the median. The five percentage points are then set in relation to the relative frequency of the entire class:

$$\frac{0.5 - F\left(x_{i-1}^{\text{UP}}\right)}{f(x_i)} = \frac{0.5 - 0.45}{0.25} = 0.2 \tag{3.29}$$

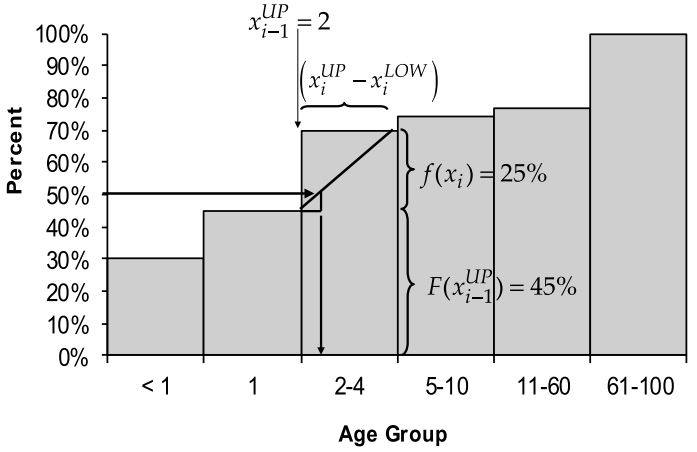


Fig. 3.15 The median: the middle value of classed data

Twenty per cent of the width of the age class that contains the median must be added on by age. This results in a Δi of three years, as the class contains all persons who are two, three, and four years old. This produces a median of $\tilde{x} = 2 + 20\% \cdot 3 = 2.6$ years. This value represents the "average user of diapers" better than the value of the arithmetic mean. Here I should note that the calculation of the median in a bimodal distribution can, in principle, be just as problematic as calculating the mean. The more realistic result here has almost everything to do with the particular characteristics of the example. The median is particularly suited when many outliers exist (see Sect. 2.6). Figure 3.15 traces the steps for us once more.

3.2.6 Quartile and Percentile

In addition to the median, there are several other important measures of central tendency that are based on the quantization of an ordered dataset. These parameters are called *quantiles*. When quantiles are distributed over 100 equally sized intervals, they are referred to as *percentiles*. Their calculation requires an ordinal or cardinal scale and can be defined in a manner analogous to the median. In an ordered dataset, the p percentile is the value at which no less than p per cent of the observations are smaller or equal in value and no less than $(1 - p)$ per cent of the observations are larger or equal in value. For instance, the 17th percentile of age in our grocery store survey is 23 years old. This means that 17% of the respondents are 23 years or younger, and 83% are 23 years old or older. This interpretation is similar to that of the median. Indeed, the median is ultimately a special case ($p = 50\%$) of a whole class of measures that partitions the ordered dataset into parts, i.e. quantiles.

In practical applications, one particular important group of quantiles is known as the quartiles. It is based on an ordered dataset divided into four equally sized parts. These are called the first quartile (the lower quartile or 25th percentile), the second quartile (the median or 50th percentile), and the third quartile (the upper quartile or 75th percentile).

Although there are several methods for calculating quantiles from raw data tables, the weighted average method is considered particularly useful and can be found in many statistics programmes. For instance, if the ordered sample has a size of $n = 850$, and we want to calculate the lower quartile ($p = 25\%$), we first have to determine the product $(n + 1) \cdot p$. In our example, $(850 + 1) \cdot 0.25$ produces the value 212.75. The result consists of an integer before the decimal mark ($i = 212$) and a decimal fraction after the decimal mark ($f = 0.75$). The integer (i) helps indicate the values between which the desired quantile lies—namely, between the observations (i) and ($i + 1$), assuming that (i) represents the ordinal numbers of the ordered dataset. In our case, this is between rank positions 212 and 213. Where exactly does the quantile in question lie between these ranks? Above we saw that the total value was 212.75, which is to say, closer to 213 than to 212. The figures after the decimal mark can be used to locate the position between the values with the following formula:

$$(1 - f) \cdot x_{(i)} + f \cdot x_{(i+1)} \tag{3.30}$$

In our butter example, the variable bodyweight produces these results:

$$\begin{aligned}(1 - 0.75) \cdot x_{(212)} + 0.75 \cdot x_{(213)} &= 0.25 \cdot 63.38 + 0.75 \cdot 63.44 \\ &= 63.43\,\text{kg}\end{aligned} \tag{3.31}$$

Another example for the calculation of the quartile is shown in Fig. 3.16.

It should be noted here that the weighted average method cannot be used with extreme quantiles. For example, to determine the 99% quantile for the five weights in Fig. 3.16 a sixth weight is needed, since $(n + 1) \cdot p = (5 + 1) \cdot 0.99 = 5.94$. This weight does not actually exist. It is fictitious, just like a weight of zero for

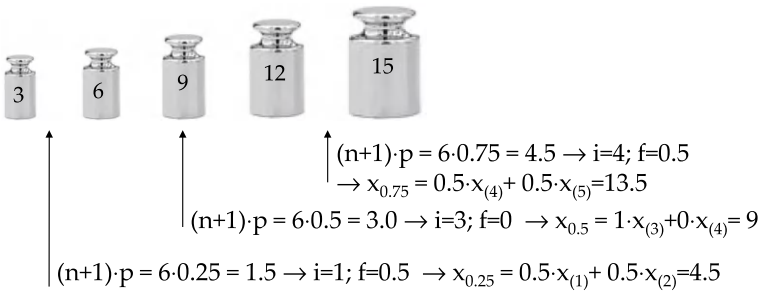


Fig. 3.16 Calculating quantiles with five weights

determining the one per cent quantile ($(n + 1) \cdot p = (5 + 1) \cdot 0.01 = 0.06$). In such cases, software programmes indicate the largest and smallest variable traits as quantiles. In the example case, we thus have:

$$x_{0.99} = 15 \text{ and } x_{0.01} = 3 \tag{3.32}$$

3.3 The Boxplot: A First Look at Distributions

We have now seen some basic measures of central tendency. All of these measures attempt to reduce dataset information to a single number expressing a general tendency. We learned that this reduction does not suffice to describe a distribution that contains outliers or special forms of dispersion. In practice, so-called boxplots are used to get a general sense of dataset distributions.

The boxplot combines various measures. Let's look at an example: imagine that over a three-year period researchers recorded the weekly sales of a certain brand of Italian salad dressing, collecting a total of 156 observations.[10] Part 1 of Fig. 3.17 shows the boxplot of weekly sales. The plot consists of a central box whose lower edge indicates the lower quartile and whose upper edge indicates the upper quartile. The values are chartered along the *y*-axis and come to 51,093 bottles sold for the lower quartile and 54,612 bottles sold for the upper quartile. The edges frame the middle 50% of all observations, which is to say: 50% of all observed weeks saw no less than 51,093 and no more than 54,612 bottles sold. The difference between the first and third quartile is called the *interquartile range*. The line in the middle of the box indicates the median position (53,102 bottles sold). The lines extending from the box describe the smallest and largest 25% of sales. Known as whiskers, these lines terminate at the lowest and highest observed values, provided they are no less than 1.5 times the box length (interquartile range) below the lower quartile or no more than 1.5 times the box length (interquartile range) above the upper quartile. Values beyond these ranges are indicated separately as potential *outliers*. Some statistical packages like SPSS differentiate between *outliers* and *extreme values*—i.e. values that are less than three times the box length (interquartile range) below the lower quartile or more than three times the box length (interquartile range) above the upper quartile. These extreme values are also indicated separately. It is doubtful whether this distinction is helpful, however, since both outliers and extreme values require separate analysis (see Sect. 2.6).

From the boxplot in Part 1 of Fig. 3.17, we can conclude the following:

- Observations 37 and 71 are outliers above the maximum (60,508 bottles sold) and below the minimum (45,682 bottles sold), respectively. These values are fairly close to the edges of the whiskers, indicating weak outliers.

[10] The data can be found in the file *Chapter3_salad_dressing* on the springer website of the book (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_3).

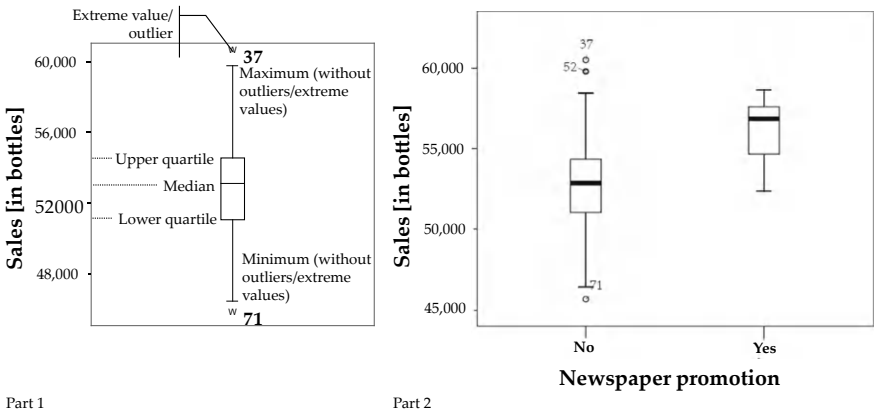


Fig. 3.17 Boxplot of weekly sales

- Some 15,000 bottles separate the best and worst sales weeks. The smallest observation (45,682 bottles) represents a deviation from the best sales week of more than 30%.
- In this example, the median lies very close to the centre of the box. This means that the central 50% of the dataset is symmetrical: the interval between the lower quartile and the median is just as large as the interval between the median and the upper quartile. Another aspect of the boxplot's symmetry is the similar length of the whiskers: the range of the lowest 25% of sales is close to that of the highest 25%.

Figure 3.18 summarizes different boxplot types and their interpretations. The boxplots are presented horizontally, not vertically, though both forms are common in practice. In the vertical form, the values are read from the *y*-axis; in the horizontal form, they are read from the *x*-axis.

If the boxplot is symmetrical—i.e. with the median in the centre of the box and whiskers of similar length—the distribution is symmetrical. When the value spread is large, the distribution is flat and lacks a clear-cut modal value. Such a distribution results, for instance, when plotting ages at a party with guests from various generations. If the value spread is small—i.e. with a compact box and whiskers—the distribution is narrow. This type of distribution results when plotting ages at a party with guests from a single generation. Boxplots can also express asymmetrical datasets. If the median is shifted to the left and the left whisker is short, then the middle 50% falls within a narrow range of relatively low values. The remaining 50% of observations are mostly higher and distributed over a large range. The resulting histogram is right-skewed and has a peak on the left side. Such a distribution results when plotting the ages of guests at a student party. Conversely, if the median is shifted to the right and the right whisker is relatively

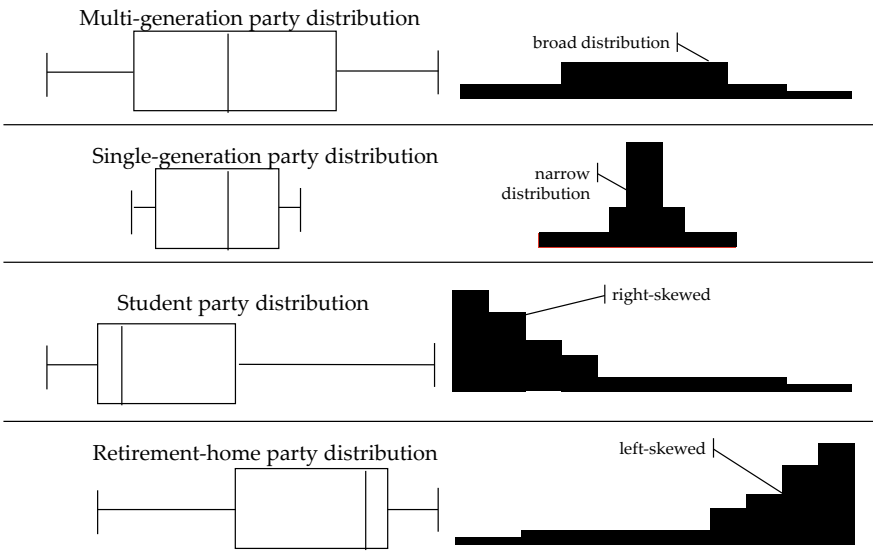


Fig. 3.18 Interpretation of different boxplot types

short, then the distribution is skewed left and has a peak on the right side. Such a distribution results when plotting the ages of guests at a retirement-home birthday party.

In addition to providing a quick overview of distribution, boxplots allow comparison of two or more distributions or groups. Let us return again to the salad dressing example. Part 2 of Fig. 3.17 displays sales for weeks in which ads appeared in daily newspapers compared with sales for weeks in which no ads appeared. The boxplots show which group (i.e. weeks with or without newspaper ads) has a larger median, a larger interquartile range, and a greater dispersion of values. Since the median and the boxplot box are larger in weeks with newspaper ads, one can assume that these weeks had higher average sales. In terms of theory, this should come as no surprise, but the boxplot also shows a left-skewed distribution with a shorter spread and no outliers. This suggests that the weeks with newspaper ads had relatively stable sales levels and a concentration of values above the median.

3.4 Dispersion Parameters

The boxplot provides an indication of the value spread around the median. The field of statistics has developed parameters to describe this spread, or dispersion, using a single measure. In the last section, we encountered our first dispersion

parameter: the *interquartile range*, i.e. the difference between the upper and lower quartile, which is formulated as

$$\text{IQR} = (x_{0.75} - x_{0.25}) \tag{3.33}$$

The larger the range, the further apart the upper and lower values of the mid-spread. Some statistics books derive from the IQR the *mid-quartile range*, or the IQR divided by two, which is formulated as

$$\text{MQR} = 0.5 \cdot (x_{0.75} - x_{0.25}) \tag{3.34}$$

The easiest dispersion parameter to calculate is one we've already encountered implicitly: *range*. This parameter results from the difference between the largest and smallest values:

$$\text{Range} = \text{Max}(x_i) - \text{Min}(x_i) \tag{3.35}$$

If the data are classed, the range results from the difference between the upper limit of the largest class of values and the lower limit of the smallest class of values. Yet we can immediately see why range is problematic for measuring dispersion. No other parameter relies so much on extreme distribution values for calculation, making range highly susceptible to outliers. If, for instance, 99 values are gathered close together and a single value appears as an outlier, the resulting range predicts a high dispersion level. But this belies the fact that 99% of the values lie very close together. To calculate dispersion, it makes sense to use as many values as possible, and not just two.

One alternative parameter is the *median absolute deviation*. Using the median as a measure of central tendency, this parameter is calculated by adding the absolute deviations of each observation and dividing the sum by the number of observations:

$$\text{MAD} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \tilde{x}| \tag{3.36}$$

In empirical practice, this parameter is less important than that of variance, which we present in the next section.

3.4.1 Standard Deviation and Variance

An accurate measure of dispersion must indicate average deviation from the mean. The first step is to calculate the deviation of every observation. Our intuition tells us to proceed as with the arithmetic mean—that is, by adding the values of the deviations and dividing them by the total number of deviations:

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) \tag{3.37}$$

Here, however, we must recall a basic notion about the mean. In an earlier section, we likened the mean to a balance scale: the sum of deviations on the left side equals the sum of deviations on the right. Adding together the negative and positive deviations from the mean always yields a value of zero. To prevent the substitution of positive with negative values, we can add the absolute deviation amounts and divide these by the total number of observations:

$$\left(\frac{1}{n}\sum_{i=1}^{n}|x_i-\overline{x}|\right) \tag{3.38}$$

Yet statistics always make use of another approach: squaring both positive and negative deviations, thus making all values positive. The squared values are then added and divided by the total number of observations. The resulting dispersion parameter is called *empirical variance*, or *population variance*, and represents one of the most important dispersion parameters in empirical research:

$$\mathrm{Var}(x)_{\mathrm{emp}} = S^2_{\mathrm{emp}} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 \tag{3.39}$$

The root of the variance yields the *population standard deviation*, or the *empirical standard deviation*:

$$S_{\mathrm{emp}} = \sqrt{\mathrm{Var}(x)_{\mathrm{emp}}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2} \tag{3.40}$$

Its value equals the average deviation from the mean. The squaring of the values gives a few large deviations more weight than they would have otherwise. To illustrate, consider the observations 2, 2, 4, and 4. Their mean is three, or $\overline{x} = 3$. Their distribution has four deviations of one unit each. The squared sum of the deviations is:

$$\sum_{i=1}^{n}(x_i-\overline{x})^2 = 1^2+1^2+1^2+1^2 = 4\,\text{units} \tag{3.41}$$

Another distribution contains the observations 2, 4, 4, and 6. Their mean is four, or $\overline{x} = 4$, and the total sum of deviations again is $2 + 2 = 4$ units. Here, two observations have a deviation of two and two observations have a deviation of zero. But the sum of the squared deviation is larger:

$$\sum_{i=1}^{n}(x_i-\overline{x})^2 = 2^2+0^2+0^2+2^2 = 8\,\text{units} \tag{3.42}$$

Although the sum of the unsquared deviations is identical in each case, a few large deviations lead to a larger empirical variance than many small deviations with the same quantity ($S^2_{\text{emp}} = 1$ vs. $S^2_{\text{emp}} = 2$). This is yet another reason to think carefully about the effect of outliers in a dataset.

Let us consider an example of variance. In our grocery store survey, the customers have an average age of 38.62 years and an empirical standard deviation of 17.50 years. This means that the average deviation from the mean age is 17.50 years.

Almost all statistics textbooks contain a second and slightly modified formula for variance or standard deviation. Instead of dividing by the total number of observations (n), one divides by the total number of observations minus one ($n - 1$). Here one speaks of *unbiased sample variance*, or of *Bessel's corrected variance*:

$$\text{Var}(x) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3.43}$$

Unbiased sample variance can then be used to find the *unbiased sample standard deviation*:

$$S = \sqrt{\text{Var}(x)} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{3.44}$$

This is a common cause of confusion among students, who frequently ask "What's the difference?" Unbiased sample variance is used when we want to infer a population deviation from a sample deviation. This method of measuring variance is necessary to make an unbiased estimation of a population deviation from a sample distribution when the mean of the population is unknown. If we use the empirical standard deviation (S_{emp}) of a sample instead, we invariably underestimate the true standard deviation of the population. Since, in practice, researchers work almost exclusively from samples, many statistics textbooks even forgo discussions of empirical variance. When large samples are being analysed, it makes little difference whether the divisor is n or $(n - 1)$. Ultimately, this is why many statistics packages indicate only the values of unbiased sample variance (standard deviation), and why publications and statistics textbooks mean unbiased sample variance whenever they speak of variance, or S^2. Readers should nevertheless be aware of this fine distinction.

3.4.2 The Coefficient of Variation

Our previous example of customer age shows that, like the mean, the standard deviation has a unit—in our survey sample, years of age. But how do we compare dispersions measured in different units? Figure 3.19 shows the height of five children in centimetres and inches. Body height is dispersed $S_{\text{emp}} = 5.1$ cm—or S_{emp}

		Child no.					Mean	S_{emp}	Coefficient of variation
		1	2	3	4	5			
cm	x	120	130	125	130	135	128.0	5.1	0.04
in	y	48	52	50	52	54	51.2	2.0	0.04

Fig. 3.19 Coefficient of variation

= 2.0 in—around the mean. Just because the standard deviation for the inches unit is smaller than the standard deviation for the centimetres unit does not mean the dispersion is any less. If two rows are measured with different units, then the values of the standard deviation cannot be used as the measure of comparison for the dispersion. In such cases, the *coefficient of variation* is used. It is equal to the quotient of the (empirical or unbiased) standard deviation and the absolute value of the mean:

$$V = \frac{S}{|\bar{x}|}, \text{ provided the mean does not have the value } \bar{x} = 0 \tag{3.45}$$

The coefficient of variation has no unit and expresses the dispersion as a percentage of the mean. Figure 3.19 shows that the coefficient of variation—0.04—has the same value regardless of whether body height is measured in inches or centimetres.

Now, you might ask, why not just convert the samples into a single unit (for example, centimetres) so that the standard deviation can be used as a parameter for comparison? The problem is that there are always real-life situations in which conversion either is impossible or demands considerable effort. Consider the differences in dispersion when measuring

- the consumption of different screws, if one measure counts the number of screws used, and the other total weight in grams;
- the value of sales for a product in countries with different currencies. Even if the average exchange rate is available, conversion is always approximate.

In such—admittedly rare—cases, the coefficient of variation should be used.

3.5 Skewness and Kurtosis

The boxplots in Fig. 3.18 not only provide information about central tendencies and dispersions, but also describe the symmetries of the distributions. Recall for a moment that the student party produced a distribution that was right-skewed (peak on the left), and the retirement-home birthday party produced a distribution that was left-skewed (peak on the right). Skewness is a measure of distribution asymmetry. A simple parameter from *Yule and Pearson* uses the difference between

median and mean in asymmetric distributions. Look at the examples in Fig. 3.20: in the right-skewed distribution, there are many observations on the left side and few observations on the right. The student party has many young students (ages 20, 21, 22, 23, 24) but also some older students and young professors (ages 41 and 45). The distinguishing feature of the right-skewed distribution is that the mean is always to the right of the median, which is why $\overline{x} > \tilde{x}$. The few older guests pull the mean upward, but leave the median unaffected. In the left-skewed distribution, the case is reversed. There are many older people at the retirement-home birthday party, but also a few young caregivers and volunteers. The latter pull the mean downwards, moving it to the left of the median ($\overline{x} < \tilde{x}$). *Yule and Pearson* express the difference between median and mean as a degree of deviation from symmetry:

$$\text{Skew} = \frac{3 \cdot (\overline{x} - \tilde{x})}{S} \tag{3.46}$$

Values larger than zero indicate a right-skewed distribution, values less than zero indicate a left-skewed distribution, and values that are zero indicate a symmetric distribution.

The most common parameter to calculate the skewness of a distribution is the so-called *third central moment*:

$$\text{Skew} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{S^3} \tag{3.47}$$

To understand this concept, think again about the left-skewed distribution of the retirement-home birthday party in Fig. 3.21. The mean is lowered by the young caregivers, moving it from around 91 to 72 years. Nevertheless, the sum of deviations on the left and right must be identical. The residents of the retirement home

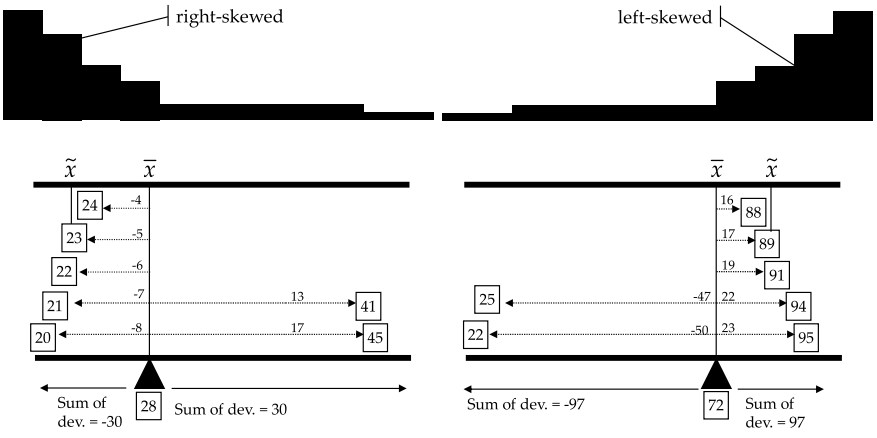
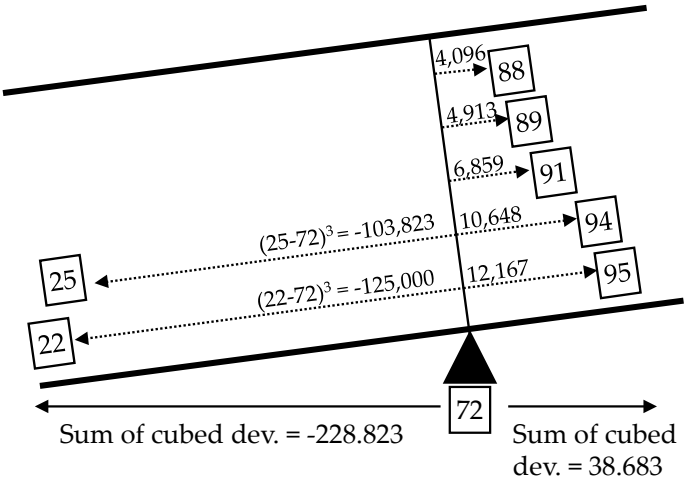


Fig. 3.20 Skewness. The numbers in the boxes represent ages. The mean is indicated by the arrow. Like a balance scale, the deviations to the left and right of the mean are in equilibrium



**The numbers in the boxes represent ages. The mean is indicated by the triangle. Like a balance scale, the cubed deviations to the left and right of the mean are in disequilibrium.

Fig. 3.21 Third central moment. The numbers in the boxes represent ages. The mean is indicated by the triangle. Like a balance scale, the cubed deviations to the left and right of the mean are in disequilibrium

create many small upward deviations on the right side of the mean (16, 17, 19, 22, 23). The sum of these deviations—97 years—corresponds exactly to the few large deviations on the left side of the mean caused by the young caregivers (47 and 50 years).

But what happens if the deviations from the mean for each observation are cubed $(x_i - \bar{x})^3$ before being summed? Cubing produces a value for caregiver ages of − 228,823 and a value for resident ages of 38,683. While the sums of the basic deviations are identical, the sums of the cubed deviations are different. The sum on the side with many small deviations is smaller than the sum on the side with a few large deviations. This disparity results from the mathematical property of exponentiation: relatively speaking, larger numbers raised to a higher power increase more than smaller numbers raised to a higher power. One example of this is the path of a parabolic curve.

The total sum of the values from the left and right hand sides results in a negative value of − 190,140 (= − 228,823 + 38,683) for the left-skewed distribution. For a right-skewed distribution, the result is positive, and for symmetric distributions the result is close to zero. A value is considered different than zero when the absolute value of the skewness is more than twice as large as the *standard error* of the skewness. This means that a skewness of 0.01 is not necessarily different than zero. The standard error is always indicated in statistics programmes and does not need to be discussed further here.

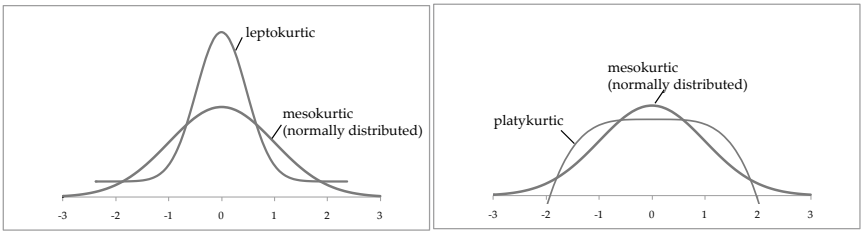


Fig. 3.22 Kurtosis distributions

Above we described the symmetry of a distribution with a single parameter. Yet what is missing is an index describing the bulge (pointy or flat) of a distribution. Using the examples in Fig. 3.18, the contrast is evident between the wide distribution of a multi-generation party and the narrow distribution of a single-generation party. *Kurtosis* is used to help determine which form is present. Defined as the *fourth central moment*, kurtosis is described by the following formula:

$$\text{Kurt} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{S^4} \tag{3.48}$$

A unimodal normal distribution as shown in Fig. 3.22 has a kurtosis value of three. This is referred to as a *mesokurtic* distribution. With values larger than three, the peak of the distribution becomes steeper, provided the edge values remain the same. This is called a *leptokurtic* distribution. When values are smaller than three, a flat peak results, also known as a *platykurtic* distribution. Figure 3.22 displays the curves of leptokurtic, mesokurtic, and platykurtic distributions.

When using software such as Excel or SPSS, similar parameters are sometimes calculated and displayed as an *excess*. But they normalize to a value of zero, not three. The user must be aware of which formula is being used when calculating kurtosis.

3.6 Robustness of Parameters

We previously discussed the effects of outliers. Some parameters, such as mean or variance, react sensitively to outliers; others, like the median in a bigger sample, don't react at all. The latter group are referred to as *robust* parameters. If the data include only robust parameters, there is no need to search for outliers. Figure 3.23 provides a summary of the permitted scales for each parameter and its robustness.

Parameter	Level of Measurement			robust?
	nominal	ordinal	cardinal	
Mean	not permitted	not permitted	permitted	not robust
Median	not permitted	permitted	permitted	robust
Quantile	not permitted	permitted	permitted	robust
Mode	permitted	permitted	permitted	robust
Sum	not permitted	not permitted	permitted	not robust
Variance	not permitted	not permitted	permitted	not robust
Interquartile range	not permitted	not permitted	permitted	robust
Range	not permitted	not permitted	permitted	not robust
Skewness	not permitted	not permitted	permitted	not robust
Kurtosis	not permitted	not permitted	permitted	not robust

**Many studies use mean, variance, skewness, and kurtosis with ordinal scales as well. Section 2.3 described the conditions necessary for this to be possible.

Fig. 3.23 Robustness of parameters. Many studies use mean, variance, skewness, and kurtosis with ordinal scales as well. Section 2.3 described the conditions necessary for this to be possible

3.7 Measures of Concentration

The above measures of dispersion dominate empirical research. They answer (more or less accurately) the following question: to what extent do observations deviate from a location parameter? Occasionally, however, another question arises: how concentrated is a trait (e.g. sales) within a group of particular statistical units (e.g. a series of firms). For instance, the EU's Directorate General for *Competition* may investigate whether a planned takeover will create excessively high concentration in a given market. To this end, indicators are needed to measure the concentration of sales, revenues, etc.

The simplest way of measuring concentration is by calculating the *concentration ratio*. Abbreviated as CR_g, the concentration ratio indicates the percentage of a quantity (e.g. revenues) achieved by g statistical units with the highest trait values. Let's assume that five companies each have a market share of 20%. The market concentration ratio CR_2 for the two largest companies is $0.2 + 0.2$, or 0.4. The other concentration ratios can be calculated in a similar fashion: $CR_3 = 0.2 + 0.2 + 0.2 = 0.6$, etc. The larger the concentration ratio is for a given g, the greater the market share controlled by the g largest companies, and the larger the concentration. In Germany, g has a minimum value of three in official statistics. In the United States, the minimum value is four. Smaller values are not published because they would allow competitors to determine each other's market shares with relative precision, thus violating confidentiality regulations (Kulick & Card, 2022).

Another very common measure of concentration is the *Herfindahl index*. First proposed by O. C. Herfindahl in a 1950 study of concentration in the US steel industry, the index is calculated by summing the squared shares of each trait (Herfindahl, 1950):

$$H = \sum_{i=1}^{n} f(x_i)^2 \tag{3.49}$$

Let us take again the example of five equally sized companies (an example of low concentration in a given industry). Using the above formula, this produces the following results:

$$H = \sum_{i=1}^{n} f(x_i)^2 = 0.2^2 + 0.2^2 + 0.2^2 + 0.2^2 + 0.2^2 = 0.2 \tag{3.50}$$

Theoretically, a company with 100% market share would have a Herfindahl index value of

$$H = \sum_{i=1}^{n} f(x_i)^2 = 1^2 + 0^2 + 0^2 + 0^2 + 0^2 = 1 \tag{3.51}$$

The value of the Herfindahl index thus varies between 1/n (provided all statistical units display the same shares and there is no concentration) and one (only one statistical unit captures the full value of a trait for itself; i.e. full concentration).

A final and important measure of concentration can be derived from the graphical representation of the *Lorenz curve*. Consider the curve in Fig. 3.25 with the example of a medium level of market concentration in Fig. 3.24. Each company represents 20% of the companies acting on the market, or 1/5 of all companies. The companies are then ordered by the size of the respective trait variable (e.g. sales), from smallest to largest, on the *x*-axis. In Fig. 3.25, the *x*-axis is spaced at 20 percentage point intervals, with the corresponding cumulative market shares on the *y*-axis. The smallest company (i.e. the lowest 20% of companies) generates 10% of sales. The two smallest companies (i.e. the lowest 40% of the companies) generate 20% of sales, while the three smallest companies generate 30% of sales, and so on.

The result is a "sagging" curve. The extent to which the curve sags depends on market concentration. If the market share is distributed equally (i.e. five companies, each representing 20% of all companies), then every company possesses 20% of the market. In this case, the Lorenz curve precisely bisects the coordinate plane. This 45-degree line is referred to the *line of equality*. As concentration increases or deviates from the uniform distribution, the Lorenz curve sags more, and the area between it and the bisector increases. If one sets the area in relationship to the entire area below the bisector, an index results between zero (uniform distribution,

	Concentration		
	Minimum	Medium	Maximum
Share of Company 1	20%	50%	100%
Share of Company 2	20%	20%	0%
Share of Company 3	20%	10%	0%
Share of Company 4	20%	10%	0%
Share of Company 5	20%	10%	0%
CR_2	40%	70%	100%
CR_3	60%	80%	100%
Herfindahl	0.20	0.32	1.00
GINI	0	0.36	0.80
$GINI_{norm.}$	0	0.45	1

Fig. 3.24 Measure of concentration

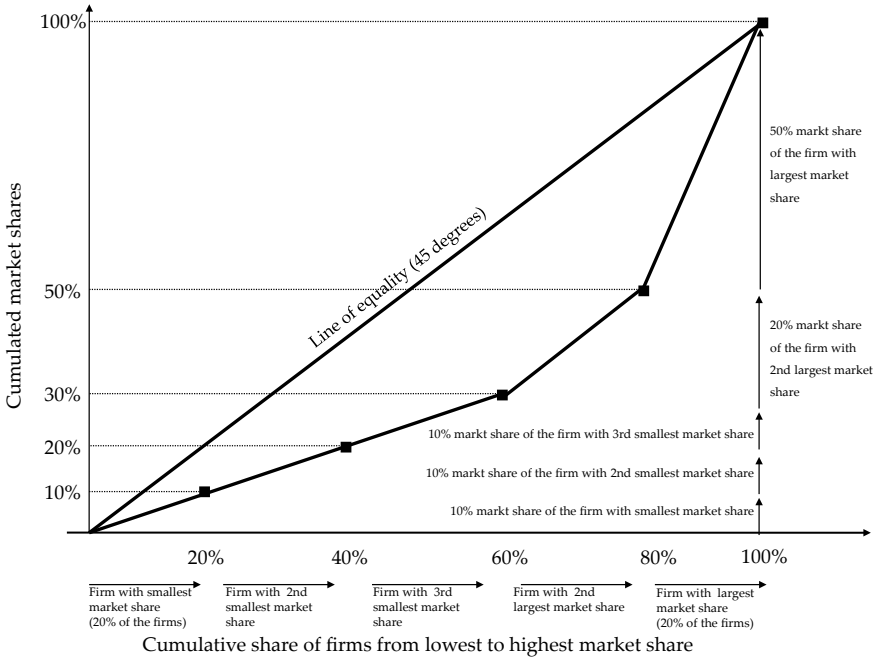


Fig. 3.25 Lorenz curve

since otherwise the area between the bisector and the Lorenz curve would be zero) and $(n-1)/n$ (full possession of all shares by one statistical unit):

$$\text{GINI} = \frac{\text{Area between bisector and the Lorenz curve}}{\text{Entire area below the bisector}} \tag{3.52}$$

This index is called the *Gini coefficient*. The following formulas are used to calculate the Gini coefficient:

(a) For unclassed ordered raw data:

$$\text{GINI} = \frac{2\sum_{i=1}^{n} i \cdot x_i - (n+1)\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i} \tag{3.53}$$

(b) For unclassed ordered relative frequencies:

$$\text{GINI} = \frac{2\sum_{i=1}^{n} i \cdot f_i - (n+1)}{n} \tag{3.54}$$

For the medium level of concentration shown in Fig. 3.24, the Gini coefficient can be calculated as follows:

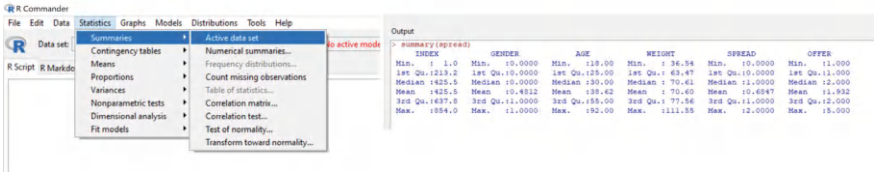
$$\begin{aligned}\text{GINI} &= \frac{2\sum_{i=1}^{n} i \cdot f_i - (n+1)}{n} \\ &= \frac{2 \cdot (1 \cdot 0.1 + 2 \cdot 0.1 + 3 \cdot 0.1 + 4 \cdot 0.2 + 5 \cdot 0.5) - (5+1)}{5} \\ &= 0.36\end{aligned} \tag{3.55}$$

In the case of full concentration, the Gini coefficient depends on the number of observations (n). The value GINI = 1 can be approximated only when a very large number of observations (n) are present. When there are few observation numbers ($n < 100$), the Gini coefficient must be normalized by multiplying each of the above formulas by $n/(n-1)$. This makes it possible to compare concentrations among different observation quantities. A full concentration always yields the value $\text{GINI}_{\text{norm.}} = 1$.

3.8 Using the Computer to Calculate Univariate Parameters

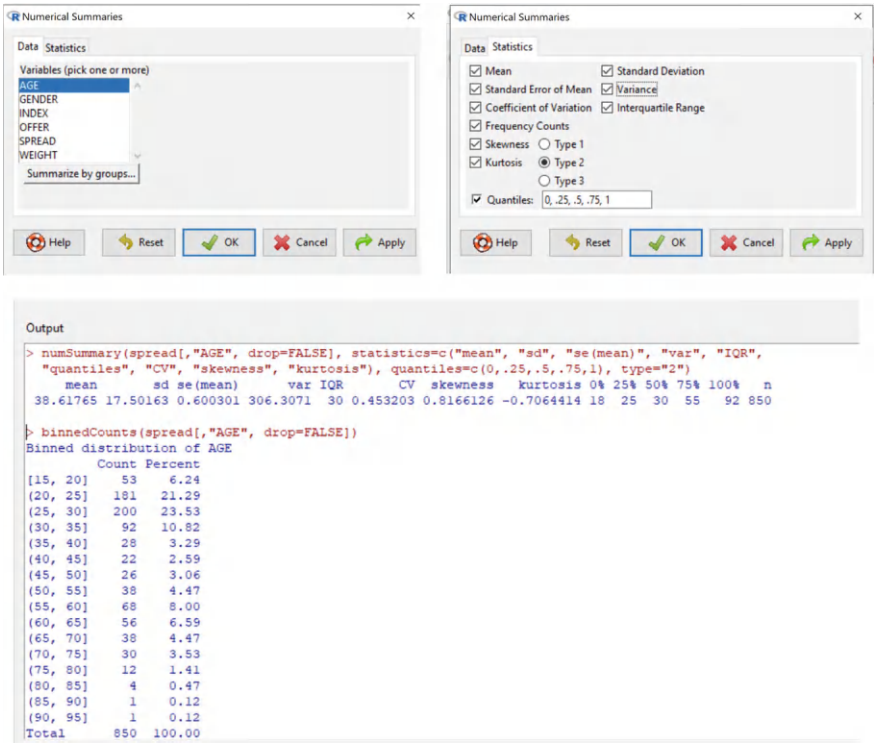
3.8.1 Calculating Univariate Parameters with R Commander

This section uses the sample dataset *Chapter3_spread.RData* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_3). The dataset is loaded by first specifying the working directory and then loading the R data file. This can be done by typing in the following R-Syntax:



Applicable syntax commands: summary(spread)

Fig. 3.26 Univariate parameters with R and R Commander (1)



Applicable syntax commands: > numSummary(spread[,"AGE", drop=FALSE], statistics=c("mean", "sd", "se(mean)", "var", "IQR", "quantiles", "CV", "skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1), type="2")

Fig. 3.27 Univariate parameters with R and R Commander (2)

setwd("C:/Data sets/R")
load("C:/Data sets/R/spread_small.RData")
or by clicking *File* → *Change working directory*…, followed by clicking *Data* → *Load data set*… in the R Commander.

Many descriptive parameters can be calculated by clicking the menu items *Statistic* → *Summaries* → *Active data set* (see Fig. 3.26). A more detailed list of univariate parameters is available through the menu sequence *Statistic* → *Summaries* → *Numerical Summaries*… (see Fig. 3.27). Furthermore, there are several other R packages, such as "pastecs" and "psych", supporting the calculation of univariate parameters. The graph discussed in this chapter can also be created using the R Commander menu. Choose *Graphs* to generate a boxplot or a histogram among many other types of graphs.

3.8.2 Calculating Univariate Parameters with SPSS

This section uses the sample dataset *Chapter3_spread.sav* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_3). There are two ways to calculate univariate parameters with SPSS. Most descriptive parameters can be calculated by clicking the menu items *Analyze* → *Descriptive Statistics* → *Frequencies*. In the menu that opens, first select the variables that are to be calculated for the univariate statistics. If there's a cardinal variable among them, deactivate the option *Display frequency tables*. Otherwise, the application will calculate contingency tables that don't typically produce meaningful results for cardinal variables. Select *Statistics*… from the submenu to display the univariate parameters for calculation.

SPSS uses a standard kurtosis of zero, not three. Figure 3.28 shows the menu and the output for the age variable from the sample dataset.

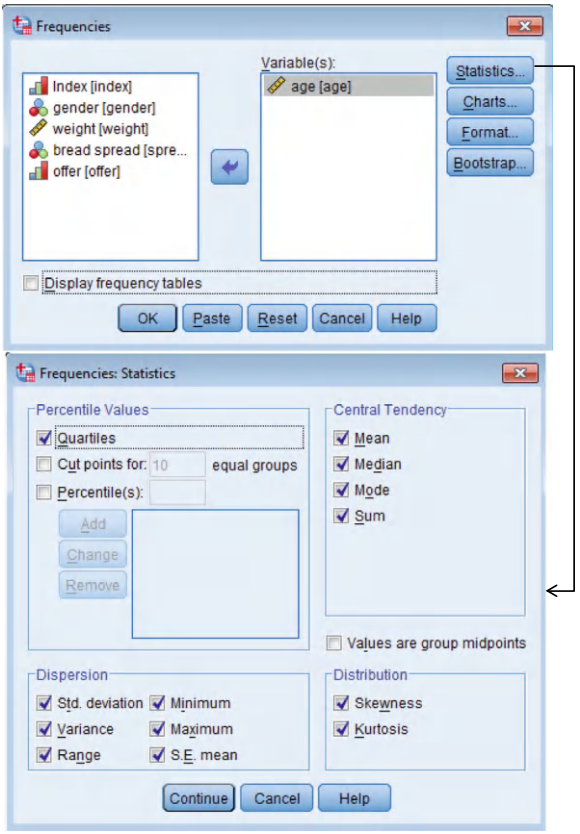
Another way to calculate univariate statistics can be obtained by selecting *Analyze* → *Descriptive Statistics* → *Descriptives*…. Once again, select the desired variables and indicate the univariate parameters in the submenu *Options.*

Choose *Graphs* → *Chart Builder*… to generate a boxplot or other graphs.

3.8.3 Calculating Univariate Parameters with Stata

Let's return again to the file *Chapter3_spread.dta* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_3). The calculation of univariate parameters with Stata can be found under *Statistics* → *Summaries, tables, and tests* → *Summary and descriptive statistics* → *Summary statistics*. From the menu select the variables to be calculated for univariate statistics. To calculate the entire range of descriptive statistics, make sure to select *Display additional statistics*, as otherwise only the mean, variance, and smallest and greatest values will be displayed. Figure 3.29 shows the menu and the output for the variable age in the sample dataset.

To see the graphs (boxplot, pie charts, etc.) select *Graphics* from the menu.



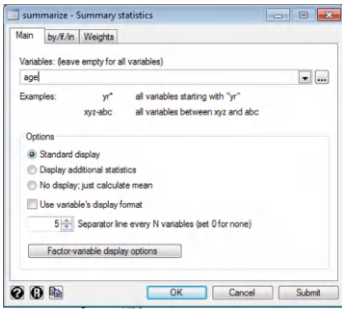
Statistics

age

N	Valid	854
	Missing	0
Mean		38.58
Std. Error of Mean		0.598
Median		30.00
Mode		25
Std. Deviation		17.472
Variance		305.262
Skewness		.823
Std. Error of Skewness		.084
Kurtosis		-.694
Std. Error of Kurtosis		.167
Range		74
Minimum		18
Maximum		92
Sum		32946
Percentiles	25	25.00
	50	30.00
	75	55.00

Applicable syntax commands: Frequencies; Descriptives

Fig. 3.28 Univariate parameters with SPSS




```
. summarize age

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |       850    38.61765    17.50163        18         92

. summarize age, detail

                            alter
-------------------------------------------------------------
      Percentiles      Smallest
 1%           18            18
 5%           20            18
10%           22            18       Obs                 850
25%           25            18       Sum of Wgt.         850

50%           30                     Mean           38.61765
                       Largest       Std. Dev.      17.50163
75%           55            83
90%           66            85       Variance       306.3071
95%           71            89       Skewness       .8151708
99%           80            92       Kurtosis       2.290657
```

Applicable syntax commands for univariate parameters:
ameans; centile; inspect; mean; pctile; summarize; mean; tabstat; tabulate summarize

Fig. 3.29 Univariate parameters with Stata

Example: Calculation of univariate parameters of the dataset *Chapter3_spread.xls* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_3)

Variable Age

Parameter	Symbol	Result	Excel Command/Function
Count	N	850	=COUNT(Data!$C$2:$C$851)
Mean	$\overline{x}$	38.62	=AVERAGE(Data!$C$2:$C$851)
Median	$\tilde{x}$	30.00	=MEDIAN(Data!$C$2:$C$851)
Mode	x_{mod}	25.00	=MODALWERT(Data!$C$2:$C$851)
Trimmed Mean	x_{trim}	37.62	=TRIMMEAN(Data!$C$2:$C$851;0,1)
Harmonic Mean	x_{harm}	32.33	=HARMEAN(Data!$C$2:$C$851)
25th percentile	$x_{0.25}$	25.00	=PERCENTILE(Data!$C$2:$C$851;0,25)
50th percentile	$x_{0,5}$	30.00	=PERCENTILE(Data!$C$2:$C$851;0,5)
75th percentile	$x_{0,75}$	55.00	=PERCENTILE(Data!$C$2:$C$851;0,75)
Minimum	MIN	18.00	=MIN(Data!$C$2:$C$851)
Maximum	MAX	92.00	=MAX(Data!$C$2:$C$851)
Sum	Σ	32,825.00	=SUM(Data!$C$2:$C$851)
Emp. Stand. Deviation	S_{emp}	17.50	=STDEVP(Data!$C$2:$C$851)
Unb. Stand. Deviation	S	17.49	=STDEV(Data!$C$2:$C$851)
Empirical Variance	VAR_{emp}	306.31	=VARP(Data!$C$2:$C$851)
Unbiased Variance	VAR	305.95	=VAR(Data!$C$2:$C$851)
Skewness		0.82	=SKEW(Data!$C$2:$C$851)
Kurtosis		-0.71	=KURT(Data!$C$2:$C$851)

Fig. 3.30 Univariate parameters with Excel

3.8.4 Calculating Univariate Parameters with Excel

Excel contains a number of pre-programmed statistical functions. These functions can be found under *Formulas → Insert Function*. Select the category *Statistical* to set the constraints. Figure 3.30 shows the Excel functions applied to the dataset *Chapter3_spread.xls* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_3). It is also possible to use the Add-ins Manager[11] to permanently activate the *Analysis ToolPak* and the *Analysis ToolPak VBA* for Excel. Next, go to *Data → Data Analysis → Descriptive Statistics*. This function can calculate the most important parameters. Excel's graphing functions can also generate the most important graphics. The option to generate a boxplot is the only thing missing from the standard range of functionality.

[11] The Add-Ins Manager can be accessed via *File → Options → Add-ins →* Manage: Excel Add-ins → Go…

Go to the website of the *Statistical Services Centre* (https://www.ssc-training.co.uk/ssc-stat.html) for a free non-commercial, Excel statistics add-in (SSC-Stat) download. In addition to many other tools, the add-in allows you to create boxplots.

Excel uses a special calculation method for determining quantiles. Especially with small samples, it can lead to implausible results. In addition, Excel scales the kurtosis to the value zero and not three, which equals a subtraction of three.

3.9 Chapter Exercises

Exercise 1

A spa resort conducts a survey of their hot spring users, asking how often they visit the spa facility. This survey results in the following absolute frequency data:

First time	Rarely	Regularly	Frequently	Every day
15	75	45	35	20

1. Identify the trait (level of measurement).
2. Sketch the relative frequency distribution of the data.
3. Identify the two location parameters that can be calculated and determine their size.
4. Identify one location parameter that can't be calculated. Why?

Exercise 2

Suppose Fig. 3.31 appears in a market research study. What can be said about it?

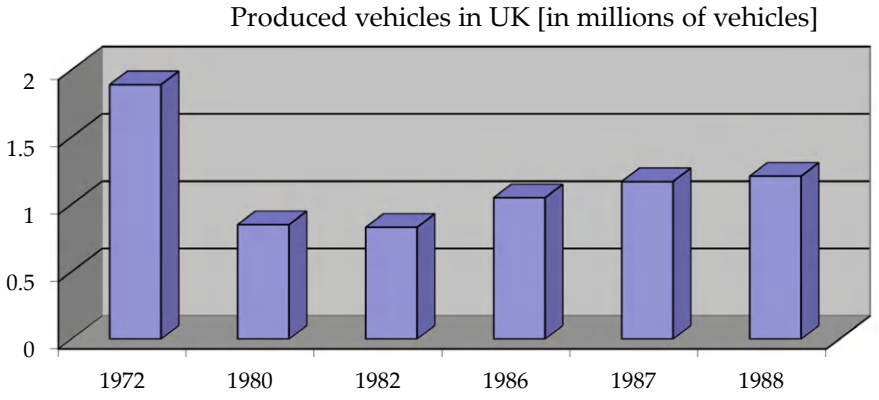


Fig. 3.31 Market research study

Exercise 3
Using the values 4, 2, 5, 6, 1, 6, 8, 3, 4, and 9 calculate

(a) The median,
(b) The arithmetic mean,
(c) The mean absolute deviation from the median,
(d) The empirical variance,
(e) The empirical standard deviation,
(f) The interquartile range.

Exercise 4
The arithmetic mean $\bar{x} = 10$ and the empirical standard deviation $S_{\text{emp}} = 2$ were calculated for a sample ($n = 50$). Later the values $x_{51} = 18$ und $x_{52} = 28$ were added to the sample. What is the new arithmetic mean and empirical standard deviation for the entire sample ($n = 52$)?

Exercise 5
You're employed in the marketing department of an international car dealer. Your boss asks you to determine the most important factors influencing car sales. You receive the following data:

Country	Sales [in 1000s of units]	Number of dealerships	Unit price [in 1000s of €]	Advertising budget [in 100,000s of €]
1	6	7	32	45
2	4	5	33	35
3	3	4	34	25
4	5	6	32	40
5	2	6	36	32
6	2	3	36	43
7	5	6	31	56
8	1	9	39	37
9	1	9	40	23
10	1	9	39	34

(a) What are the average sales (in 1000s of units)?
(b) What are the empirical standard deviation and the coefficient of variation?
(c) What would be the coefficient of variation if sales were given in a different unit of quantity?
(d) Determine the lower, middle, and upper quartile of sales with the help of the "weighted average method".
(e) Draw a boxplot for the variable sales.
(f) Are sales symmetrically distributed across the countries? Interpret the boxplot.

(g) How are company sales concentrated in specific countries? Determine and interpret the Herfindahl index.
(h) Assume that total sales developed as follows over the years: 2016: + 2%; 2017: + 4%; 2018: + 1%. What is the average growth in sales for this period?

Exercise 6
A used car market contains 200 vehicles across the following price categories:

Car price (in €)	Number
Up to 2500	2
Between 2500 and 5000	8
Between 5000 and 10,000	80
Between 10,000 and 12,500	70
Between 12,500 and 15,000	40

(a) Draw a histogram for the relative frequencies. How would you have done the data acquisition differently?
(b) Calculate and interpret the arithmetic mean, the median, and the modal class.
(c) What price is reached by 45% of the used cars?
(d) 80% of used cars in a different market are sold for more than €11,250. Compare this value with the market figures in the above table.

Exercise 7
Unions and employers sign a four-year tariff agreement. In the first year, employees' salaries increase by 4%, in the second year by 3%, in the third year by 2%, and in the fourth year by 1%. Determine the average salary increase to four decimal places.

Exercise 8
A company has sold €30 m worth of goods over the last three years. In the first year, they sold €8 m, in the second year €7 m, and in the third year €15 m. What is the concentration of sales over the last three years? Use any indicator to solve the problem.

3.10 Exercise Solutions

Solution 1

1. Ordinal.
2. Figure based on the following percentages:

First time	Rarely	Frequently	Regularly	Daily
15	75	45	35	20
15/190 = 7.89%	75/190 = 39.47%	45/190 = 23.68%	18/190 = 18.42%	20/190 = 10.53%

3. Mode = 2 (rare); median = 3 (frequently).
4. Mean, as this assumes metric scale.

Solution 2
The distance between years is not uniform. This suggests a rise in motor vehicle production. In reality, production dropped between 1972 and 1979 (not indicated). A histogram would be the right choice for such a case.

Solution 3

(a) First sort the dataset, then:

$$\tilde{x} = \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) = \frac{1}{2}\left(x_{(5)} + x_{(6)}\right) = \frac{1}{2}(4+5) = 4.5 \tag{3.56}$$

(b) $$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{48}{10} = 4.8 \tag{3.57}$$

(c) $$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \tilde{x}| = \frac{20}{10} = 2 \tag{3.58}$$

(d) $$\text{Var}(x)_{\text{emp}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2 = \frac{288}{10} - 4.8^2 = 5.76 \tag{3.59}$$

(e) $$S_{\text{emp}} = \sqrt{\text{Var}(x)_{\text{emp}}} = 2.4 \tag{3.60}$$

(f) Next calculate the lower and upper quartiles.

$$\begin{aligned}
&x_{0.25} : (n+1)\cdot p = (10+1)\cdot 0.25 = 2.75 \\
&\rightarrow x_{0.25} = (1-f)\cdot x_i + f \cdot x_{i+1} = 0.25 \cdot x_2 + 0.75 \cdot x_3 \\
&\qquad = 0.25 \cdot 2 + 0.75 \cdot 3 = 2.75 \\
&\rightarrow x_{0.75} : (n+1)\cdot p = (10+1)\cdot 0.75 = 8.25 \\
&\rightarrow x_{0.75} = 0.75 \cdot x_8 + 0.25 \cdot x_9 = 0.75 \cdot 6 + 0.25 \cdot 8 = 6.5. \\
&\rightarrow \text{ The interquartile range is } x_{0.75} - x_{0.25} = 3.75.
\end{aligned}$$

Solution 4
In the old sample ($n = 50$) the sum of all observations is

$$\sum_{i=1}^{50} x_i = n \cdot \bar{x} = 50 \cdot 10 = 500. \tag{3.61}$$

The new sample has two more observations, for a total sum of

$$\sum_{i=1}^{52} x_i = 500 + 18 + 28 = 546. \tag{3.62}$$

The value for the arithmetic mean is thus

$$\bar{x}_{\text{new}} = \frac{\sum_{i=1}^{52} x_i}{50 + 2} = \frac{546}{52} = 10.5. \tag{3.63}$$

To calculate empirical variance, the following generally applies:

$$S^2_{\text{emp}} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2. \tag{3.64}$$

For the original sample $n = 50$,

$$S^2_{\text{emp}_{\text{old}}} = 4 = \frac{1}{50}\left(\sum_{i=1}^{50} x_i^2\right) - 10^2 \text{ applies,} \tag{3.65}$$

producing the following sum of squares

$$\sum_{i=1}^{50} x_i^2 = 50 \cdot \left(4 + 10^2\right) = 5200. \tag{3.66}$$

From this we can determine the empirical variance of the new sample:

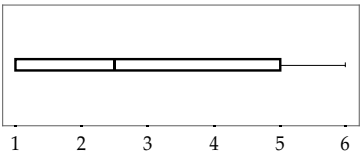
$$\begin{aligned} S^2_{\text{emp}_{\text{new}}} &= \frac{1}{n+2}\left(\sum_{i=1}^{n} x_i^2 + x_{51}^2 + x_{52}^2\right) - \bar{x}^2_{\text{new}} \\ &= \frac{1}{52}\left(5200 + 18^2 + 28^2\right) - 10.5^2 = 11.06. \end{aligned} \tag{3.67}$$

To determine the empirical standard deviation, we must extract the root from the result, for

$$S_{\text{emp}_{\text{new}}} = 3.33. \tag{3.68}$$

Solution 5

(a) $\bar{x} = 3$;
(b) $S_{\text{emp}} = 1.79$; $V = 0.6$;
(c) Identical, as this assumes a coefficient of variation without units;



(d) $x_{0.25} = 1$; $x_{0.5} = 2.5$; $x_{0.75} = 5$;
(e) Min = 1; Max = 6;
(f) Right-skewing tendency;
(g) $H = 0.136$;
(h) $\bar{x}_{\text{geom}} = \sqrt[3]{(1 + 0.02)(1 + 0.04)(1 + 0.01)} - 1 = 2.3\%$

Solution 6

(a) The middle price class is twice as large as the other price classes. A bar graph (see Fig. 3.32 on the left) would be misleading here because the €5000 to €10,000 price class sticks out as a frequently chosen class. But if we consider the width of the class and create a histogram, a different picture emerges: now the €10,000 to €12,500 price class is the highest (most chosen). The height of the bars in the histogram is as follows: 2/2500 = 0.0008; 8/2500 = 0.0032; 80/5000 = 0.016; 70/2500 = 0.028; 40/2500 = 0.016.
(b) The mean can be calculated from the class midpoint: $\bar{x}$ = €9850; the class median must lie above €10,000, as up to €10,000 only 45% = 1% + 4% + 40% of the values come together: $x_{0.5} = 10{,}000 + 2500 \cdot 5/35 =$ €10,357.14; modal class: €10,000 − 12,500.
(c) $x_{0.55} = 10{,}000 + 2500 \cdot (5 + 5)/35 = 10{,}714.28$.
(d) $x_{0.2} = 5000 + 5000 \cdot (15)/40 =$ €6875. The cars on the other used market are more expensive on average.

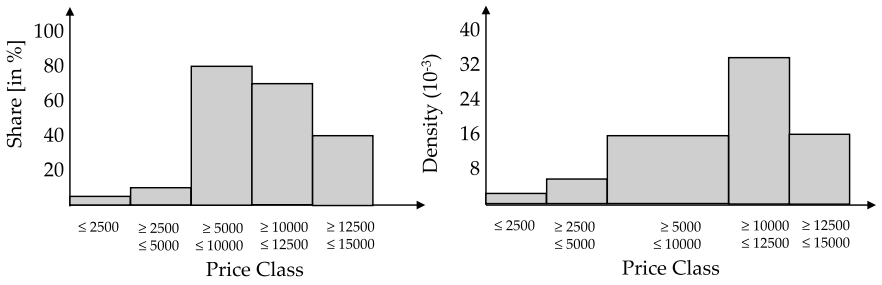


Fig. 3.32 Bar graph and histogram

Solution 7
The question is about growth rates. Here the geometric mean should be applied.

$$\bar{x}_{\text{geom}} = \sqrt[4]{(1+0.04)(1+0.03)(1+0.02)(1+0.01)} - 1 = 0.024939 = 2.49\% \tag{3.69}$$

Solution 8
$\text{CR}_2 = 76.67\%$

$$\text{Herfindahl: } H = \sum_{i=1}^{n} f(x_i)^2 = \left(\frac{7}{30}\right)^2 + \left(\frac{8}{30}\right)^2 + \left(\frac{15}{30}\right)^2 = 0.38 \tag{3.70}$$

$$\text{GINI} = \frac{2\sum_{i=1}^{n} i \cdot f_i - (n+1)}{n} = \frac{2 \cdot \left(1 \cdot \frac{7}{30} + 2 \cdot \frac{8}{30} + 3 \cdot \frac{15}{30}\right) - (3+1)}{3} = 0.18 \tag{3.71}$$

$$\text{GINI}_{\text{norm.}} = \frac{n}{n-1} \cdot \text{GINI} = 0.27 \tag{3.72}$$

References

Herfindahl, O. (1950). *Concentration in the U.S. Steel Industry* (Dissertation). Columbia University.
Krämer, W. (2008). *Statistik verstehen. Eine Gebrauchsanweisung* (8th ed.). Piper.
Krämer, W. (2015). *So lügt man mit Statistik* (17th ed.). Campus.
Kulick, R., & Card, A. (2022). *Industrial concentration in the United States: 2002–2017*. NERA Economic Consulting and US Chamber of Commerce.
Schwarze, J. (2013). *Aufgabensammlung zur Statistik* (7th ed.). nwb.
Swoboda, H. (1971). *Exakte Geheimnisse: Knaurs Buch der modernen Statistik*. Knaur.



Bivariate Association

4

4.1 Bivariate Scale Combinations

In the first stage of data analysis, we learned how to examine variables and survey traits individually, or univariately. In this chapter, we'll learn how to assess the association between two variables using methods known as bivariate analyses. This is where statistics starts getting interesting—practically as well as theoretically. This is because univariate analysis is rarely satisfying in real life. People want to know things like the strength of a relationship

- between advertising costs and product sales,
- between interest rate and share prices,
- between wages and employee satisfaction, or
- between specific tax return questions and tax fraud.

Questions like these are very important, but answering them requires far more complicated methods than the ones we've used so far. As in univariate analysis, the methods of bivariate analysis depend on the scale of the observed traits or variables. Table 4.1 summarizes scale combinations, their permitted bivariate measures of association, and the sections in which they appear.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78070-7_4.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*,
Springer Texts in Business and Economics,
https://doi.org/10.1007/978-3-031-78070-7_4

Table 4.1 Scale combinations and their measures of association

		Nominal	Ordinal	Metric
Nominal	Dichotomous	Phi; Cramer's V [Sect. 4.2]	Biserial rank correlation; Cramer's V [Sects. 4.2 and 4.5.2]	Point-biserial r; classification of metric variables and application of Cramer's V [Sects. 4.2 and 4.5.1]
	Non-dichotomous	Cramer's V; contingency coefficient [Sect. 4.2]	Cramer's V; contingency coefficient [Sect. 4.2]	Classification of metric variables and application of Cramer's V [Sect. 4.2]
Ordinal			Spearman's rho (ρ); Kendall's tau (τ) [Sect. 4.4]	Ranking of metric variables and application of ρ or τ [Sect. 4.4]
Metric				Pearson's correlation (r) [Sect. 4.3.2]

The appropriate measure of association is indicated in the box at the point where the scales intersect. For instance, if one variable is nominal and dichotomous and the other ordinally scaled, then the association can be measured either by the biserial rank correlation or Cramer's V. If both variables are ordinal, then one can use either Spearman's rho or Kendall's tau

4.2 Association Between Two Nominal Variables

4.2.1 Contingency Tables

A common form of representing the association of two nominally scaled variables is the *contingency table* or *crosstab*. The bivariate contingency table takes the univariate frequency table one step further: it records the frequency of value pairs. Figure 4.1 shows a contingency table for the variables *gender* and *selection rating* from our sample survey in Chap. 2.

The right and lower edges of the table indicate the *marginal frequencies*. The values along the right edge of the table show that 441 (51.6%) of the 854 respondents are male and 413 (48.4%) are female. We could have also obtained this information had we calculated a univariate frequency table for the variable *gender*. The same is true for the frequencies of the variable *selection rating* on the lower edge of the contingency table. Of the 854 respondents, 392 (45.9%) find the selection poor, 270 (31.6%) fair, etc. In the interior of the contingency table, we find additional information. For instance, 199 respondents (23.3%) were *male* and found the selection *poor*.

Gender * rating cross tabulation

			offer					Total
			poor	fair	avg	good	excellent	
Gender	male	Count	199	143	52	27	20	441
		Expected count	202.4	139.4	47.0	32.0	20.1	441.0
		% within gender	45.1%	32.4%	11.8%	6.1%	4.5%	100.0%
		% within rating	50.8%	53.0%	57.1%	43.5%	51.3%	51.6%
		% of total	23.3%	16.7%	6.1%	3.2%	2.3%	51.6%
	female	Count	193	127	39	35	19	413
		Expected count	189.6	130.6	44.0	30.0	18.9	413.0
		% within gender	46.7%	30.8%	9.4%	8.5%	4.6%	100.0%
		% within rating	49.2%	47.0%	42.9%	56.5%	48.7%	48.4%
		% of total	22.6%	14.9%	4.6%	4.1%	2.2%	48.4%
Total		Count	392	270	91	62	39	854
		Expected count	392.0	270.0	91.0	62.0	39.0	854.0
		% within gender	45.9%	31.6%	10.7%	7.3%	4.6%	100.0%
		% within rating	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
		% of total	45.9%	31.6%	10.7%	7.3%	4.6%	100.0%

Fig. 4.1 Contingency table (crosstab)

Alongside absolute frequencies and the frequencies expressed relative to the total number of respondents we can also identify *conditional relative frequencies*. For instance, how large is the relative frequency of females within the group of respondents who rated the selection to be *poor*? First look at the subgroup of respondents who checked *poor*. Of these 392 respondents, 193 are female, so the answer must be 49.2%. The formal representation of these conditional relative frequencies is as follows:

$$f(\text{gender} = \textit{female} | \text{selection} = \textit{poor}) = 193/392 = 49.2\% \tag{4.1}$$

The limiting condition appears after the vertical line behind the value in question. The question "What per cent of female respondents rated the selection as good?" would limit the female respondents to 413. This results in the following conditional frequency:

$$f(\text{selection rating} = \textit{good} | \text{gender} = \textit{female}) = 35/413 = 8.5\% \tag{4.2}$$

The formula $f(x = 1 \mid y = 0)$ describes the relative frequency of the value $x = 1$ for the variable x when only observations with the value $y = 0$ are considered.

4.2.2 Chi-Square Calculations

The contingency table gives us some initial indications about the strength of the association between two nominal or ordinal variables. Consider the contingency tables in Fig. 4.2. They show the results of two business surveys. Each survey has $n = 22$ respondents.

The lower crosstab shows that none of the ten male respondents and all twelve female respondents made a purchase. From this we can conclude that all women made a purchase and all men did not, and that all buyers are women and all non-buyers are men. From the value of one variable (*gender*), we can infer the value of the second (*purchase*). The upper contingency table, by contrast, does not permit this conclusion. Of the male respondents, 50% are buyers and 50% non-buyers. The same is true of the female respondents.

These tables express the extremes of association: in the upper table, there is no association between the variables *gender* and *purchase*, while in the lower table there is a perfect association between them. The extremes of association strength can be discerned through close examination of the tables alone. But how can contingency tables be compared whose associations are less extreme? How much weaker, for instance, is the association in the contingency table in Fig. 4.3 compared with the second contingency table in Fig. 4.2?

As tables become more complicated, so do estimations of association. The more columns and rows a contingency table has, the more difficult it is to recognize associations and compare association strengths between tables. The solution is to calculate a parameter that expresses association on a scale from zero (no association) to one (perfect association). To calculate this parameter, we must first

		Gender		
		Female	Male	Total
Purchase	No Purchase	6	5	11
	Purchase	6	5	11
Total		12	10	22

		Gender		
		Female	Male	Total
Purchase	No Purchase	0	10	10
	Purchase	12	0	12
Total		12	10	22

Fig. 4.2 Contingency tables (crosstabs) (1st)

		Gender		Total
		Female	Male	
Purchase	No purchase	1	9	10
	Purchase	11	1	12
Total		12	10	22

Fig. 4.3 Contingency table (crosstab) (2nd)

determine the *expected frequencies*—also known as *expected counts*—for each cell. These are the absolute values that would obtain were there no association between variables is assumed. In other words, one calculates the *expected absolute frequencies* under the assumption of statistical independence.

Let us return again to the first table in Fig. 4.2. A total of 12 of the 22 respondents are female. The relative frequency of females is thus

$$f_{\text{female}} = \frac{12}{22} = 54.5\% \tag{4.3}$$

The relative frequency of a purchase is 11 of 22 persons, or

$$f_{\text{purchase}} = \frac{11}{22} = 50.0\% \tag{4.4}$$

If there is no association between the variables (gender and purchase), then 50% of the women and 50% of the men must make a purchase. Accordingly, the expected relative frequency of female purchases under independence would be:

$$f_{\text{purchase}}^{\text{female}} = f_{\text{purchase}} \cdot f_{\text{female}} = \frac{11}{22} \cdot \frac{12}{22} = 50.0\% \cdot 54.5\% = 27.3\% \tag{4.5}$$

From this we can easily determine the expected counts under independence: six persons, or 27.3% of the 22 respondents, are female and make a purchase:

$$n_{\text{purchase}}^{\text{female}} = f_{\text{purchase}} \cdot f_{\text{female}} \cdot n = \frac{11}{22} \cdot \frac{12}{22} \cdot 22 = \frac{11 \cdot 12}{22} = 6 \tag{4.6}$$

The simplified formula for calculating the expected counts under independence is *row sum (12) multiplied by the column sum (11) divided by the total sum (22)*:

$$n_{ij}^{e} = \frac{\text{row sum} \cdot \text{column sum}}{\text{total sum}} = \frac{n_{i.} \cdot n_{.j}}{n} \tag{4.7}$$

The sum of expected counts in each row or column must equal the absolute frequencies of the row or column. The idea is that a statistical association is not signified by different marginal frequencies but by different distributions of the sums of the marginal frequencies across columns or rows.

By comparing the expected counts n_{ij}^{e} with the actual absolute frequencies n_{ij} and considering their difference $(n_{ij} - n_{ij}^{e})$, we get a first impression of the deviation of actual data from statistical independence. The larger the difference, the more the variables tend to be statistically dependent.

One might be tempted just to add up the deviations of the individual rows. In the tables in Fig. 4.4 the result is always zero, as the positive and negative differences cancel each other out. This happens with every contingency table. This is why we must square the difference in every cell and then divide it by the expected count. For the female buyers in part 1 of the above table, we then have the following value:

$$\frac{\left(n_{12} - n_{12}^{e}\right)^2}{n_{12}^{e}} = \frac{(6-6)^2}{6} = 0. \tag{4.8}$$

These values can then be added up for all cells in the m rows and k columns. This results in the so-called chi-square value (χ^2-square):

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(n_{ij} - n_{ij}^{e}\right)^2}{n_{ij}^{e}} = \frac{(6-6)^2}{6} + \frac{(6-6)^2}{6} + \frac{(5-5)^2}{5} + \frac{(5-5)^2}{5} = 0. \tag{4.9}$$

The chi-square is a value that is independent of the chosen variable code and in which positive and negative deviations do not cancel each other out. If the chi-square has a value of zero, there is no difference to the expected counts with independence. The observed variables are thus independent of each other. In our example this means that *gender* has no influence on purchase behaviour.

As the dependence of the variables increases, the value of the chi-square tends to rise, which Fig. 4.4 clearly shows.

Part 1: No association

			Sex		
			Female	Male	Total
Purchase	No purchase	Count	6	5	11
		Expected Count	6.0	5.0	11.0
	Purchase	Count	6	5	11
		Expected Count	6.0	5.0	11.0
Total		Count	12	10	22
		Expected Count	12.0	10.0	22.0

Part 2: Perfection association

			Sex		
			Female	Male	Total
Purchase	No purchase	Count	0	10	10
		Expected count	5.5	4.5	10.0
	Purchase	Count	12	0	12
		Expected count	6.5	5.5	12.0
Total		Count	12	10	22
		Expected count	12.0	10.0	22.0

Part 3: Strong association

			Sex		
			Female	Male	Total
Purchase	No purchase	Count	1	9	10
		Expected count	5.5	4.5	10.0
	Purchase	Count	11	1	12
		Expected count	6.5	5.5	12.0
Total		Count	12	10	22
		Expected count	12.0	10.0	22.0

Fig. 4.4 Calculation of expected counts in contingency tables

In part 2 one can infer perfectly from one variable (*gender*) to another (*purchase*), and the other way around. All women buy something and all men do not. All non-buyers are male and all buyers are female. For the chi-square this gives us:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{\left(n_{ij} - n_{ij}^{e}\right)^2}{n_{ij}^{e}} = \frac{(0-5.5)^2}{5.5} + \frac{(12-6.5)^2}{6.5} + \frac{(10-4.5)^2}{4.5} + \frac{(0-5.5)^2}{5.5} = 22 \tag{4.10}$$

Its value also equals the number of observations ($n = 22$).

Let us take a less extreme situation and consider the case in part 3 of Fig. 4.4. Here one female respondent does not make a purchase and one male respondent does make a purchase, reducing the value for of the chi-square:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{\left(n_{ij} - n_{ij}^{e}\right)^2}{n_{ij}^{e}} = \frac{(1-5.5)^2}{5.5} + \frac{(11-6.5)^2}{6.5} + \frac{(9-4.5)^2}{4.5} + \frac{(1-5.5)^2}{5.5} = 14.7 \tag{4.11}$$

Unfortunately, the strength of association is not the only factor that influences the size of the chi-square value. As the following sections show, the chi-square value tends to rise with the size of the sample and the number of rows and columns in the contingency tables, too. Adopted measures of association based on the chi-square thus attempt to limit these undesirable influences.

4.2.3 The Phi Coefficient

In the last section, we saw that the value of the chi-square rises with the dependence of the variables and the size of the sample. Figure 4.5 shows two contingency tables with perfect association: the chi-square value is $n = 22$ in the table with $n = 22$ observations and $n = 44$ in the table with $n = 44$ observations.

As these values indicate, the chi-square does not achieve our goal of measuring association independent of sample size. For a measure of association to be independent, the associations of two tables whose sample sizes are different must be comparable. For tables with two rows ($2 \times k$) or two columns ($m \times 2$), it is best to use the phi coefficient. The phi coefficient results from dividing the chi-square value by the number of observations and taking its square root:

$$\text{PHI} = \varphi = \sqrt{\frac{\chi^2}{n}} \tag{4.12}$$

Part 1: Perfect association with n=22 observations

			Gender		
			Female	Male	Total
Purchase	No Purchase	Count	0	10	10
		Expected Count	5.5	4.5	10.0
	Purchase	Count	12	0	12
		Expected Count	6.5	5.5	12.0
Total		Count	12	10	22
		Expected Count	12.0	10.0	22.0

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(n_{ij} - n_{ij}^e)^2}{n_{ij}^e} = \frac{(0-5.5)^2}{5.5} + \frac{(12-6.5)^2}{6.5} + \frac{(10-4.5)^2}{4.5} + \frac{(0-5.5)^2}{5.5} = 22$$

Part 2: Perfect association with n=44 observations

			Gender		
			Female	Male	Total
Purchase	No Purchase	Count	0	20	20
		Expected Count	10.9	9.1	20.0
	Purchase	Count	24	0	24
		Expected Count	13.1	10.9	24.0
Total		Count	24	20	44
		Expected Count	24.0	20.0	44.0

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(n_{ij} - n_{ij}^e)^2}{n_{ij}^e} = \frac{(0-10.9)^2}{10.9} + \frac{(24-13.1)^2}{13.1} + \frac{(20-9.1)^2}{9.1} + \frac{(0-10.9)^2}{10.9} = 44$$

Fig. 4.5 Chi-square values based on different sets of observations

Using this formula,[1] the phi coefficient assumes a value from zero to one. If the coefficient has the value of zero, there is no association between the variables. If it has the value of one, the association is perfect.

[1] Some software programmes calculate the phi coefficient for a 2×2 table (four-field scheme) in such a way that phi can assume negative values. This has to do with the arrangement of the rows and columns in the table. In these programmes, a value of (− 1) equals an association strength of (+ 1), and (− 0.6) that of (+ 0.6), etc.

If the contingency table consists of more than two rows and two columns, the phi coefficient will produce values greater than one. Consider a table with three rows and three columns and a table with five rows and four columns. Here too there are perfect associations, as every row possesses values only within a column and every row can be assigned to a specific column (see Fig. 4.6).

Part 1: Perfect association in a 3x3 contingency table

			Purchase			Total
			No Purchase	Frequent Purchase	Constant Purchase	
Customer group	A Customer	Count	0	0	10	10
		Expected Count	3.3	3.3	3.3	10.0
	B Customer	Count	0	10	0	10
		Expected Count	3.3	3.3	3.3	10.0
	C Customer	Count	10	0	0	10
		Expected Count	3.3	3.3	3.3	10.0
	Total	Count	10	10	10	30
		Expected Count	10.0	10.0	10.0	30.0

$$\varphi = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{60}{30}} = \sqrt{2} = 1.4$$

Part 2: Perfect association in a 4x5 contingency table

			Purchase				Total
			No	Infre-quent	Fre-quent	Constant	
Customer Group	A Customer	Count	0	0	10	0	10
		Expected Count	4.0	2.0	2.0	2.0	10.0
	B Customer	Count	0	10	0	0	10
		Expected Count	4.0	2.0	2.0	2.0	10.0
	C Customer	Count	10	0	0	0	10
		Expected Count	4.0	2.0	2.0	2.0	10.0
	D Customer	Count	10	0	0	0	10
		Expected Count	4.0	2.0	2.0	2.0	10.0
	E Customer	Count	0	0	0	10	10
		Expected Count	4.0	2.0	2.0	2.0	10.0
	Total Customer	Count	20	10	10	10	50
		Expected Count	20.0	10.0	10.0	10.0	50.0

$$\varphi = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{150}{50}} = \sqrt{3} = 1.73$$

Fig. 4.6 The phi coefficient in tables with various numbers of rows and columns

As these tables show, the number of rows and columns determines the phi coefficient's maximum value. The reason is that the highest obtainable value for the chi-square rises as the number of rows and columns increases. The maximum value of phi is the square root of the minimum number of rows and columns in a contingency table minus one:

$$\varphi_{\max} = \sqrt{\min(\text{Number of rows}, \text{Number of columns}) - 1} \geq 1 \tag{4.13}$$

In practice, therefore, the phi coefficient should only be used when comparing 2×2 contingency tables.

4.2.4 The Contingency Coefficient

This is why some statisticians suggest using the contingency coefficient instead. It is calculated as follows:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \in [0; 1[\tag{4.14}$$

Like the phi coefficient, the contingency coefficient assumes the value of zero when there is no association between the variables. Unlike the phi coefficient, however, the contingency coefficient never assumes a value larger than one. The disadvantage of the contingency coefficient is that C never assumes the value of one under perfect association. Let us look at the contingency tables in Fig. 4.7.

Although both tables show a perfect association, the contingency coefficient does not have the value of $C = 1$.

The more rows and columns a table has, the closer the contingency coefficient comes to one in case of perfect association. But a table would have to have many rows and columns before the coefficient came anywhere close to one, even under perfect association. The maximal reachable value can be calculated as follows:

$$C_{\max} = \sqrt{\frac{\min(k, m) - 1}{\min(k, m)}} = \sqrt{1 - \frac{1}{\min(k, m)}} \tag{4.15}$$

The value for k equals the number of columns and m the number of rows. The formula below yields a standardized contingency coefficient between zero and one:

$$C_{\text{korr}} = \sqrt{\frac{\chi^2}{\chi^2 + n}} \cdot \sqrt{\frac{\min(k, m)}{\min(k, m) - 1}} = \sqrt{\frac{\chi^2}{\chi^2 + n}} \cdot \frac{1}{\sqrt{1 - \frac{1}{\min(k,m)}}} \in [0; 1] \tag{4.16}$$

Part 1: Perfect association in a 2x2 contingency table

			Gender		
			Female	Male	Total
Purchase	No Purchase	Count	0	10	10
		Expected Count	5.5	4.5	10.0
	Purchase	Count	12	0	12
		Expected Count	6.5	5.5	12.0
Total		Count	12	10	22
		Expected Count	12.0	10.0	22.0

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{22}{22 + 22}} = \sqrt{\frac{1}{2}} = \sqrt{0.5} = 0.71$$

Part 2: Perfect association in a 3x3 contingency table

			Purchase			
			No Purchase	Frequent Purchase	Constant Purchase	Total
Customer Group	A Customer	Count	0	0	10	10
		Expected Count	3.3	3.3	3.3	10.0
	B Customer	Count	0	10	0	10
		Expected Count	3.3	3.3	3.3	10.0
	C Customer	Count	10	0	0	10
		Expected Count	3.3	3.3	3.3	10.0
	Total	Count	10	10	10	30
		Expected Count	10.0	10.0	10.0	30.0

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{60}{60 + 30}} = \sqrt{\frac{2}{3}} = 0.82$$

Fig. 4.7 The contingency coefficient in tables with various numbers of rows and columns

4.2.5 Cramer's V

One measure that is independent of the size of the contingency table is Cramer's V. It always assumes a value between zero (no association) and one (perfect association) and is therefore in practice one of the most helpful measures of association between two nominal or ordinal variables. Its calculation is an extension of the phi coefficient:

$$\text{Cramer}'\text{s V} = \sqrt{\frac{\chi^2}{n \cdot (\min(k, m) - 1)}} = \varphi \cdot \sqrt{\frac{1}{\min(k, m) - 1}} \in [0; 1] \qquad (4.17)$$

The value for n equals the number of observation, k the number of columns, and m the number of rows. The values from the tables in Fig. 4.7 produce the following calculation:

1. $$\text{Cramer}'\text{s V} = \sqrt{\frac{\chi^2}{n \cdot (\min(k, m) - 1)}} = \sqrt{\frac{22}{22 \cdot (2 - 1)}} = 1 \qquad (4.18)$$

2. $$\text{Cramer}'\text{s V} = \sqrt{\frac{\chi^2}{n \cdot (\min(k, m) - 1)}} = \sqrt{\frac{60}{30 \cdot (3 - 1)}} = 1 \qquad (4.19)$$

We have yet to identify which values stand for *weak*, *moderate*, or *strong* associations. Some authors define the following ranges:

- Cramer's V $\in$ [0.00; 0.10[$\rightarrow$ no association
- Cramer's V $\in$ [0.10; 0.30[$\rightarrow$ weak association
- Cramer's V $\in$ [0.30; 0.60[$\rightarrow$ moderate association
- Cramer's V $\in$ [0.60; 1.00[$\rightarrow$ strong association.

4.2.6 Nominal Associations with SPSS

Everyone knows the story of the Titanic. It's a tale of technological arrogance, human error, and social hierarchy. On April 10, 1912, the Titanic embarked on its maiden cruise, from Southampton, England to New York. Engineers at the time considered the giant steamer unsinkable on account of its state-of-the-art technology and sheer size. Yet on April 14 the ship struck an iceberg and sank around 2:15 am the next day. Of the 2201 passengers, only 710 survived.

Say we want to examine the frequent claim that most of the survivors were from first class and most of the victims were from third class. To start we need the information in the Titanic dataset *Chapter4_titanic.sav* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_4), including the variables

gender (child, male, female), *class* (first, second, third, and crew), and *survival* (yes, no) for each passenger.[2]

To use SPSS to generate a crosstab and calculate the nominal measures of association, begin by opening the crosstab window. Select *Analyze* → *Descriptive Statistics* → *Crosstabs*…. Now select the row and column variables whose association you want to analyse. For our example we must select *survival* as the row variable and *class* as the column variable (see Fig. 4.8). Next click on *cells*… to open a cell window. There you can select the desired contingency table calculations (see Fig. 4.8: The cell display). The association measure can be selected under *statistics*… (see Fig. 4.8: Statistics). Click *OK* to generate the tables in Figs. 4.9 and 4.10.

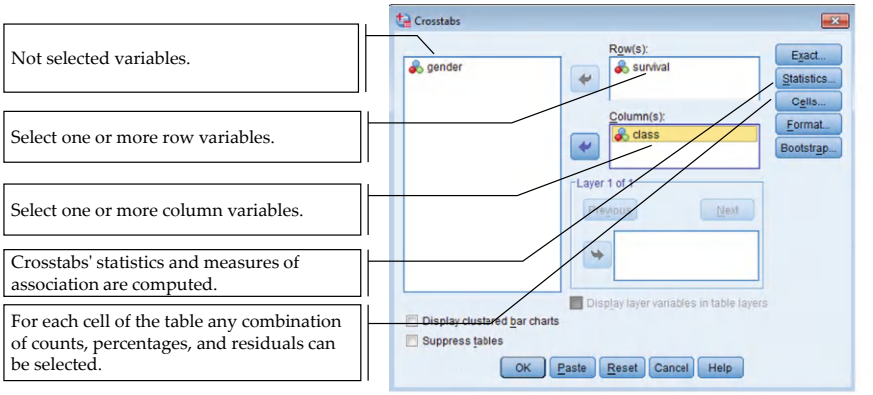
Consider the contingency table in Fig. 4.9, which categorizes survivors by class. Did all the passengers have the same chances of survival?

We see that more passengers in third class (528) lost their lives than passengers in first class (123). But since more passengers were in third class (706 vs. 325), this is no surprise, even if everyone had the same chances of survival. But when we consider the relative frequencies, we see that 32.3% of passengers survived the catastrophe, with 62.2% from first class and only 25.2% from third class. These figures indicate that the 32.3% survival rate is distributed asymmetrically: the larger the asymmetry, the stronger the relationship between class and survival.

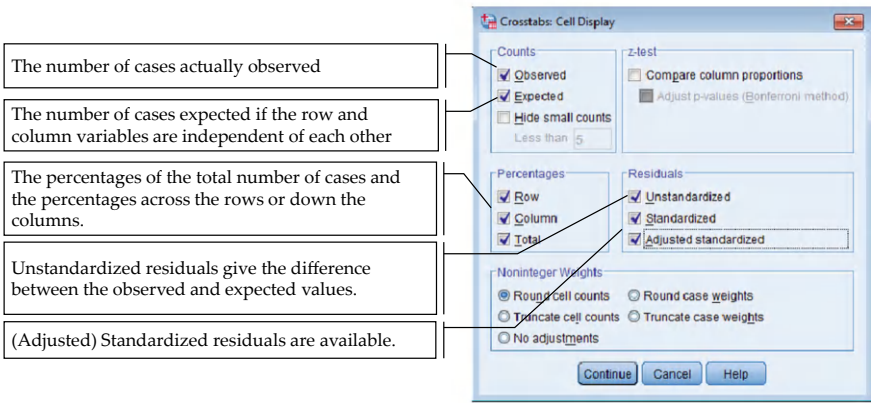
If first-class passengers survived at the average rate, only $32.3\% \cdot 325 \approx 105$ would have made it. This is the expected count under statistical independence. If third-class passengers survived at the average rate, only $67.7\% \cdot 706 \approx 478$ would have died, not 528.

As we saw in the previous sections, the differences between the expected counts and the actual absolute frequency give us a general idea about the relationship between the variables. For closer analysis, however, the differences must be standardized by dividing them by the root of the expected counts (std. residual). The square of the standardized values yields the chi-square for each cell. Positive values for the standardized residuals express an above-average (empirical) frequency in relation to the expected frequency; negative values express a below-average (empirical) frequency in relation to the expected frequency. First-class passengers have a survival value of 9.5, third-class passengers − 3.3—above-average and below-average rates, respectively. Because all standardized residuals are a long way from zero, we can assume there is some form of association.

[2] The data in *Chapter4_Titanic.sav* (SPSS), *Chapter4_Titanic.dta* (Stata), *Chapter4_Titanic.RData* (R), and *Chapter4_Titanic.xls* (Excel) contain figures on the number of persons on board and the number of victims (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_4). The data is taken from the British Board of Trade Inquiry Report (1990), Report on the Loss of the Titanic' (S.S.), Gloucester (reprint).



Cell Display



Statistics

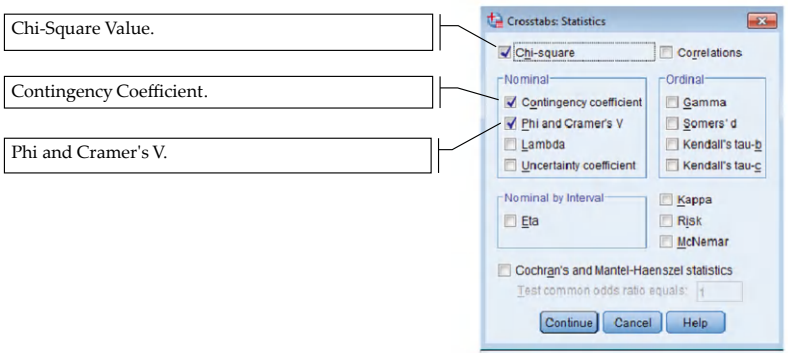


Fig. 4.8 Crosstabs and nominal associations with SPSS (Titanic)

			class				Total
			Crew	First	Second	Third	
survival	Alive	Count	212	202	118	178	710
		Expected Count	285.5	104.8	91.9	227.7	710.0
		% within survival	29.9%	28.5%	16.6%	25.1%	100.0%
		% within class	24.0%	62.2%	41.4%	25.2%	32.3%
		% of Total	9.6%	9.2%	5.4%	8.1%	32.3%
		Residual	-73.5	97.2	26.1	-49.7	
		Std. Residual	-4.3	9.5	2.7	-3.3	
		Adjusted Residual	-6.8	12.5	3.5	-4.9	
	Dead	Count	673	123	167	528	1491
		Expected Count	599.5	220.2	193.1	478.3	1491.0
		% within survival	45.1%	8.2%	11.2%	35.4%	100.0%
		% within class	76.0%	37.8%	58.6%	74.8%	67.7%
		% of Total	30.6%	5.6%	7.6%	24.0%	67.7%
		Residual	73.5	-97.2	-26.1	49.7	
		Std. Residual	3.0	-6.5	-1.9	2.3	
		Adjusted Residual	6.8	-12.5	-3.5	4.9	
Total		Count	885	325	285	706	2201
		Expected Count	885.0	325.0	285.0	706.0	2201.0
		% within survival	40.2%	14.8%	12.9%	32.1%	100.0%
		% within class	100.0%	100.0%	100.0%	100.0%	100.0%
		% of Total	40.2%	14.8%	12.9%	32.1%	100.0%

Fig. 4.9 From raw data to computer-calculated crosstab (Titanic)

The association is confirmed by the relatively high chi-square value and the relatively high measure of association (see Fig. 4.10). The application of the phi coefficient is permitted here—a 4×2 table—as $2 \times k$ or $m \times 2$ tables always yield identical values for Cramer's V and phi. Cramer's V (0.292) indicates an association just shy of moderate, but, as is always the case with Cramer's V, whether the association is the one supposed—in our example, a higher survival rate among first-class passengers than among third-class passengers and not the other way around—must be verified by comparing standardized residuals between actual and the expected frequencies.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	187.793[a]	3	.000
N of Valid Cases	2201		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 91.94.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.292	.000
	Cramer's V	.292	.000
	Contingency Coefficient	.280	.000
N of Valid Cases		2201	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Fig. 4.10 Computer printout of chi-square and nominal measures of association

4.2.7 Nominal Associations with Stata

To analyse nominal associations with Stata, follow a similar approach. Open the dataset *Chapter4_titanic.dta* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_4). Select *Statistics* → *Summaries, tables, and tests* → *Tables* → *Two-way tables with measures of association* to open the window in Fig. 4.11.

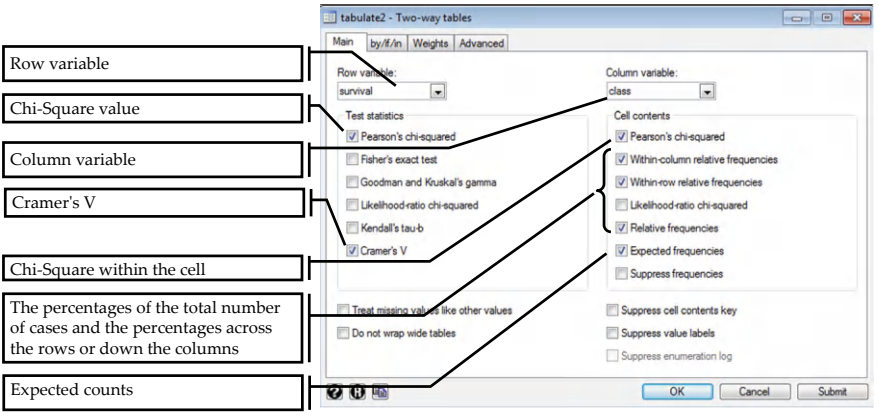


Fig. 4.11 Crosstabs and nominal measures of association with Stata (Titanic)

The rows, columns, and calculations must be selected for each variable. The left side displays the measures of association; the right side shows the cell statistics of the contingency table. Click on *OK* or *Submit* to perform the Stata calculation.[3] The results can now be interpreted as in the SPSS example (see Sect. 4.2.6).

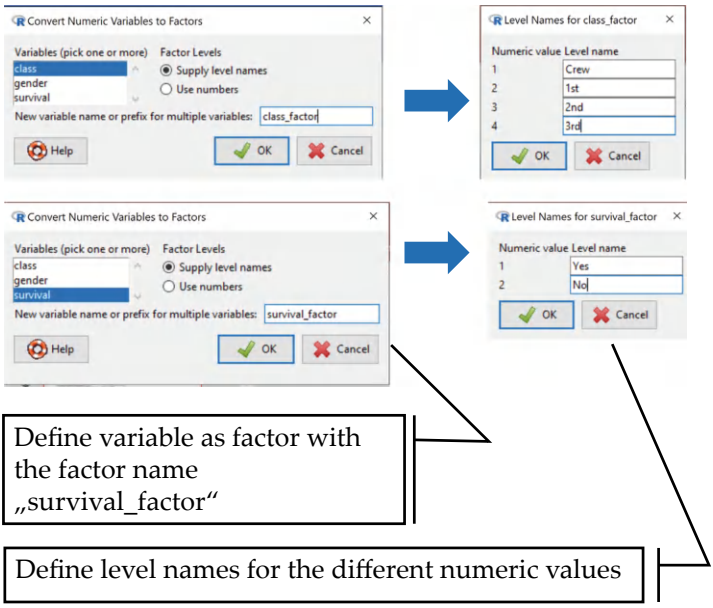
4.2.8 Nominal Associations with R and R Commander

When using statistical methods with R, it is important to ensure that the data are in the correct data format. As with many other similar statistical analyses involving categorical variables, crosstabs in R require categorical data to be in what is known as factor format. Therefore, the data types should be checked in advance (step 1) and, if necessary, converted to factor format (step 2).

- Step 1: Check the data structure: Open the R Commander by typing *library(Rcmdr)* and open the dataset *Chapter4_titanic.RData* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_4) by clicking on *Data → Load data set…* or by using the syntax *load("~/[path]/Chapter4_titanic.RData").* Look at the structure of the data to see if the variables are of an appropriate type (e.g. factors for categorical data). The *str()* function provides a compact and comprehensive summary of the structure of an R dataset, including data types. Type the command *str(Titanic)* in the script window in order to do so. Select the command and click the *Submit* button or press *Ctrl-R* to run the script. The following output will appear in the R (Studio) output window:

```
Rcmdr > str(Titanic)
 'table' num [1:4, 1:2, 1:2, 1:2] 0 0 35 0 0 0 17 0 118 154 …
 - attr(*, "dimnames") = List of 4
  ..$ Class    : chr [1:4] "1st" "2nd" "3rd" "Crew"
  ..$ Sex      : chr [1:2] "Male" "Female"
  ..$ Age      : chr [1:2] "Child" "Adult"
  ..$ Survived: chr [1:2] "No" "Yes"
```

[3] Syntax command: tabulate class survived, cchi2 cell chi2 clrchi2 column expected row V.



Syntax:
titanic <- within(titanic, {class_factor <- factor(class, labels=c('Crew','1st','2nd','3rd'))})
titanic <- within(titanic, {survival_factor <- factor(survival, labels=c('YES','NO'))})

Fig. 4.12 Convert variables to factors with R Commander (Titanic)

The categorical variables are not defined in factor but in character (chr) format and therefore need to be converted.

- Step 2: Convert variables to factors (if necessary): Go to *Data* → *Manage variables in active data set* → *Convert numeric variables to factors…* Select the variables you need to convert and specify the levels if necessary (see Fig. 4.12). Click *OK* to accept the changes. The syntax shown in Fig. 4.12 may of course also be used.

The R Commander is now ready to create crosstabs. Go to *Statistics* → *Contingency tables* → *Two-way table…* In the dialogue boxes, select *survival_factor* as the row variable and *class_factor* as the column variable. Select the desired cell percentages and hypothesis tests options in the Statistics window. Click *OK* to generate the crosstab and the chi-square test results (see Fig. 4.13).

It is even easier to generate a crosstab using the *descr* package. It produces crosstabs similar to the style of typical standard software such as Stata or SPSS.

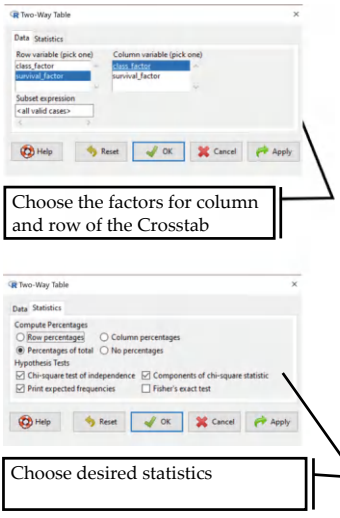
```
# Install and load descr package
        install.packages("descr")
```

```
        library(descr)
# Use CrossTable function from descr
        CrossTable(titanic$survival_factor, titanic$class_factor, chisq = TRUE, expected
= TRUE, sresid = TRUE, format = "SPSS")
```

The association measures Cramer's V and Contingency Coefficient can be calculated using the R Commander script window. Type the following commands in the script window:

```
# Install and load vcd package
        install.packages("vcd")
        library(vcd)
# Calculate Cramer's V and Contingency Coefficient
        titanic_table ← xtabs(Freq ~ Survived + Class, data = Titanic)
        assocstats(titanic_table)
```

Select the commands and click *Submit* or press *Ctrl-R*. See Fig. 4.14 for the resulting crosstabs and nominal measures of association with the R Commander. The results can now be interpreted as in the SPSS example (see Sect. 4.2.6).




```
Frequency table:
                 class_factor
survival_factor Crew 1st 2nd 3rd
            No   212 202 118 178
            Yes  673 123 167 528
Total percentages:
      Crew  1st  2nd  3rd Total
No     9.6  9.2  5.4  8.1  32.3
Yes   30.6  5.6  7.6 24.0  67.7
Total 40.2 14.8 12.9 32.1 100.0
Column percentages:
                 class_factor
survival_factor Crew   1st   2nd   3rd
          Yes     24  62.2  41.4  25.2
          No      76  37.8  58.6  74.8
          Total  100 100.0 100.0 100.0
          Count  885 325.0 285.0 706.0

             Pearson's Chi-squared test
data:  .Table
X-squared = 187.79, df = 3, p-value < 2.2e-16

Expected counts:
                 class_factor
survival_factor      Crew      1st       2nd       3rd
            No   285.4839 104.8387  91.93548 227.7419
            Yes 599.5161 220.1613 193.06452 478.2581
Rcmdr> library(vcd)
Rcmdr> titanic_table <- xtabs(Freq ~ Survived + Class, data = Titanic)
Rcmdr> assocstats(titanic_table)
                    X^2 df P(> X^2)
Likelihood Ratio 180.9  3        0
Pearson          190.4  3        0
Phi-Coefficient   : NA
Contingency Coeff.: 0.28
Cramer's V        : 0.29
```

Fig. 4.13 Crosstabs and nominal measures of association with R Commander (Titanic)

```
Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|            Std Residual |
|-------------------------|

=============================================================================
                                titanic$class_factor
titanic$survival_factor      Crew      1st      2nd      3rd    Total
-----------------------------------------------------------------------------
YES                           212      202      118      178      710
                           285.5    104.8     91.9    227.7
                          18.915   90.046    7.390   10.864
                           29.9%    28.5%    16.6%    25.1%   32.3%
                           24.0%    62.2%    41.4%    25.2%
                            9.6%     9.2%     5.4%     8.1%
                          -4.349    9.489    2.718   -3.296
-----------------------------------------------------------------------------
NO                            673      123      167      528     1491
                           599.5    220.2    193.1    478.3
                           9.007   42.879    3.519    5.173
                           45.1%     8.2%    11.2%    35.4%   67.7%
                           76.0%    37.8%    58.6%    74.8%
                           30.6%     5.6%     7.6%    24.0%
                           3.001   -6.548   -1.876    2.275
-----------------------------------------------------------------------------
Total                         885      325      285      706     2201
                            40.2%    14.8%    12.9%    32.1%
=============================================================================

Statistics for All Table Factors

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 = 187.7932      d.f. = 3      p <2e-16

        Minimum expected frequency: 91.93548

Rcmdr> library(vcd)
Rcmdr> titanic_table <- xtabs(Freq ~ Survived + Class, data = Titanic)
Rcmdr> assocstats(titanic_table)
                    X^2 df P(> X^2)
Likelihood Ratio 180.9  3        0
Pearson          190.4  3        0
Phi-Coefficient   : NA
Contingency Coeff.: 0.280
Cramer's V        : 0.292
```

Fig. 4.14 Crosstabs and nominal measures of association using R syntax (Titanic)

4.2.9 Nominal Associations with Excel

The calculation of crosstabs and related parameters (chi-square, phi, contingency coefficient, Cramer's V) with Excel is tricky compared with professional statistics packages. One of its main drawbacks is the shortage of pre-programmed functions for contingency tables.

Here is a brief sketch of how to perform these functions in Excel if needed. Open the dataset *Chapter4_titanic.xlsx* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_4). First select the (conditional) actual frequencies for each cell as in Fig. 4.15. The pivot table function can be helpful. Select the commands *Insert* and *Pivot Table* to open the *Create Pivot Table*. Then choose *Select a table or a range* and mark the location of the raw data. Click *OK* to store the pivot table in a *New Worksheet*. Drag the variables *survived* and *class* from the field list and drop them in the *Drop Row Fields Here* and *Drop Column Fields Here*. This generates a crosstab without conditional absolute frequencies. These can be added by dragging one of the variables from the field list to the field $\sum$ *values*. Then click on the variable in the field and select *Value Field Settings*… and the option *count* in the dialogue box. This generates a crosstab with the actual

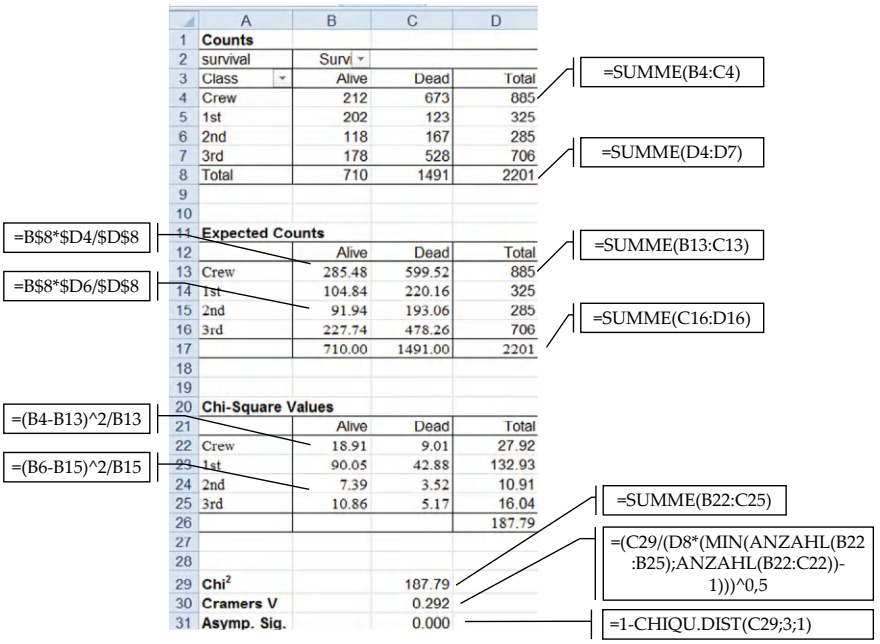


Fig. 4.15 Crosstabs and nominal measures of association with Excel (Titanic)

absolute frequencies. To update the crosstab when changes are made in the raw data, move the cursor over a cell and select *Options* and *Refresh* on the *PivotTable* tab. You can then calculate the expected frequencies using the given formula (row sum multiplied by the column sum divided by the total sum; see the second table in Fig. 4.15). In a new table we can calculate the individual chi-squares for each cell (see the third table in Fig. 4.15). The sum of these chi-squares equals the total chi-square value. From this, we can calculate Cramer's V. The formulas in Fig. 4.15 provide an example.

4.3 Association Between Two Metric Variables

In the previous sections, we explored how to measure the association between two nominal or ordinal variables. This section presents methods for determining the strength of association between two metric variables. As before, we begin with a simple example.

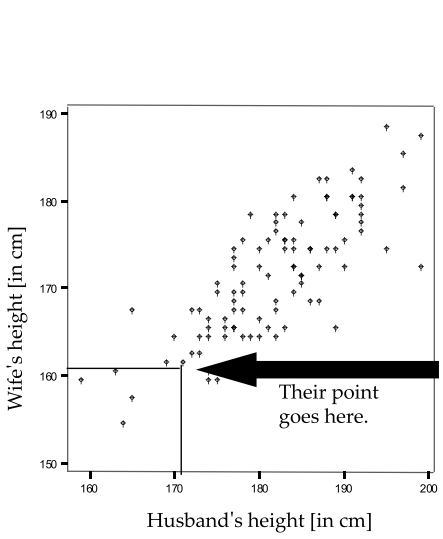
4.3.1 The Scatterplot

Officials performing civil marriage ceremonies frequently observe that brides and grooms tend to be of similar height. Taller men generally marry taller women and vice versa. One official decides to verify this impression by recording the heights of 100 couples. How can he tell whether there's an actual association, and, if so, its strength?

One way to get a sense for the strength of association between two metric variables is to create a so-called scatterplot. The first step is to plot the variables. In our example the groom heights follow the *x*-axis and the bride heights follow the *y*-axis. Each pair forms a single data point in the coordinate system. The first couple (observation 12: "Lesley and Joseph") is represented by the coordinate with the values 171.3 for the groom and 161.0 for the bride. Plotting all the observed pairs results in a cloud of points or scatterplot (see Fig. 4.16).

This scatterplot permits us to say several things about the association between the heights of marrying couples. It turns out that there is indeed a positive association: taller males tend to marry taller females and shorter males to tend marry shorter females. Moreover, the association appears to be nearly linear, with the occasional deviation.

All in all, a scatterplot expresses three aspects of the association between two metric variables. Figure 4.17 provides some examples.

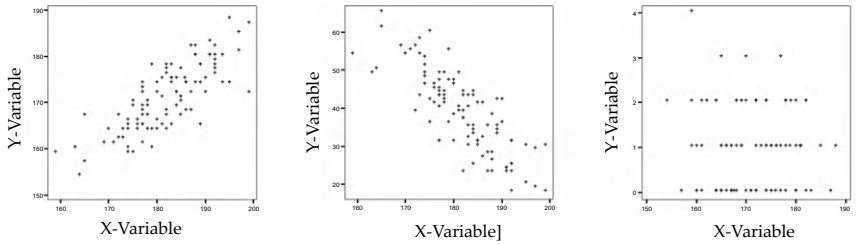


	names	wheight	hheight
1	John and Judy	1590	1809
2	Carl and Kathryn	1560	1841
3	Craig and Jackie	1620	1659
4	Larry and Susan	1540	1779
5	Scott and Susan	1420	1616
6	John and Margaret	1660	1695
7	Stanley and Patricia	1610	1730
8	David and Lisa	1635	1753
9	Robert and Cathy	1580	1740
10	Larry and Karen	1610	1685
11	Steven and Candice	1590	1735
12	Joseph and Lesley	1610	1713
13	Eric and Ethel	1700	1736

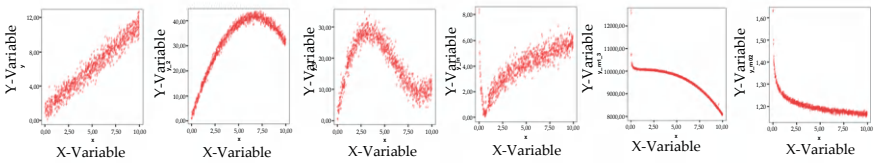
Observation 12:
Joseph (171.3 cm) and Lesley (161.0 cm)

Fig. 4.16 The scatterplot

1. The **direction** of the relationship



2. The **form** of the relationship



3. The **strength** of the relationship

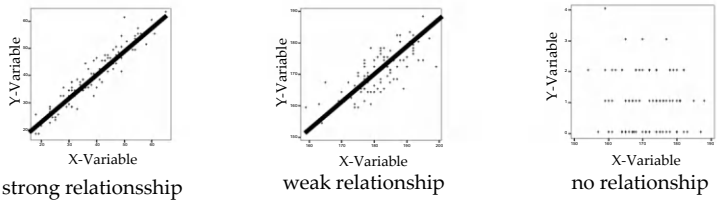


Fig. 4.17 Aspects of association expressed by the scatterplot

1. *The direction of the relationship.* Relationships can be positive, negative, or non-existent. A relationship is positive when the values of the x and y variables increase simultaneously. A relationship is negative when the y variable decreases and the x variable increases. A relationship is non-existent when no patterns can be discerned in the cloud of points, i.e. when x values produce both small and large y values.
2. *The form of the relationship.* The form of a relationship can be linear or non-linear.
3. *The strength of the relationship.* The strength of a relationship is measured by the proximity of the data points along a line. The closer they are, the stronger the relationship.

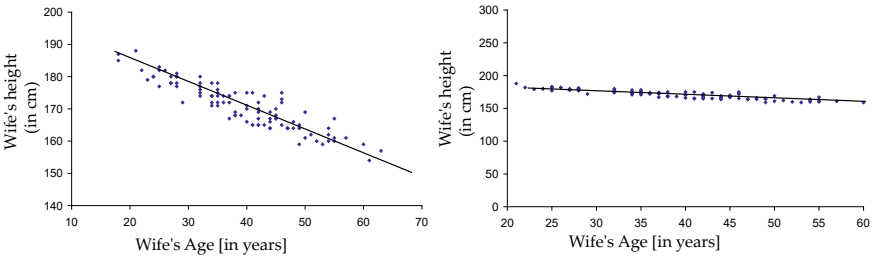


Fig. 4.18 Different representations of the same data (3)

There are many software tools that make creating scatterplots easy.[4] But their interpretation requires care. Figure 4.18 provides an illustrative example. It presents the relationship between female age and height in two ways.

The data used for each scatterplot are identical. In the left diagram in Fig. 4.18, the *y*-axis is scaled between 140 and 200 cm and the *x*-axis between 10 and 70. In the diagram to the right height is scaled between zero and 300 cm and age between 20 and 60. But if you compare the diagrams, your first instinct would be to see a negative relationship in the left diagram, as the line through the cloud of data points appears to be steeper than the line in the second diagram. Moreover, the relationship in the left diagram seems weaker than that in the second diagram, for the observation points scatter at greater distances to the line. A mere change of scale can reinforce or weaken the impression left by the scatterplot. This opens the door to manipulation.

We thus need a measure that gives us an unadulterated idea of the relationship between two metric variables, one that provides us with information about the direction (positive or negative) and the strength of the relationship independent of the unit of measure for the variables. A measure like this is referred to as a correlation coefficient.

4.3.2 The Bravais–Pearson Correlation Coefficient

In many statistics books, the authors make reference to a single type of correlation coefficient; in reality, however, there is more than just one. The *Bravais–Pearson* correlation coefficient measures the strength of a linear relationship. Spearman's correlation coefficient or *Kendall's tau coefficient* (and its variations) measure the

[4] In Excel, mark the columns (i.e. the variables) and use the diagram assistant (under *Insert* and *Charts*) to select the *scatterplot* option. After indicating the chart title and other diagram options (see *Chart Tools*), you can generate a scatterplot. *SPSS* is also straightforward. Select *Graphs* → *Chart Builder* → *Scatter/Dot*, pick one of the scatter options and then drag the variables in question and drop them at the axes. In Stata, select *Graphics* → *Twoway Graph* → *Create* → *Scatter*. In the window define the variables of the *x*- and *y*-axes. The syntax is: scatter variable_x variable_y.

strength of a monotonic relationship as well as the association between two ordinal variables. The *point-biserial correlation coefficient* determines the relationship between a dichotomous and a metric variable.

Let's begin with the Bravais–Pearson correlation coefficient, often referred to as the *product-moment correlation coefficient* or *Pearson's correlation coefficient*. This coefficient was the result of work by the French physicist Auguste Bravais (1811–1863) and the British mathematician Karl Pearson (1857–1936). It defines an absolute measure that can assume values between $r = (-1)$ and $r = (+1)$. The coefficient takes the value of $(+1)$ when two metric variables have a perfect linear and positive relationship (i.e. all observed values lie along a rising linear slope). It takes the value of (-1) when two metric variables have a perfect linear and negative relationship (i.e. all observed values lie along a falling linear slope). The closer it is to zero, the more the value pairs diverge from a perfect linear relationship.

To derive Pearson's correlation coefficient, first we must determine *covariance.* We already learned about variance in our discussion of univariate statistics. We defined it as the measure of the squared average deviation of all observation points. When two variables are involved, we speak of covariance, which is the measure of the deviation between each value pair from the bivariate centroid in a scatterplot. To understand covariance better, let us consider again the scatterplot for the heights of marrying couples. Consider Fig. 4.19.

In this figure a line is drawn through the mean groom height ($\bar{x} = 181.6$ cm) and the mean bride height ($\bar{y} = 170.9$ cm). The point where they intersect is the bivariate centroid for an average couple, where groom and bride are each of

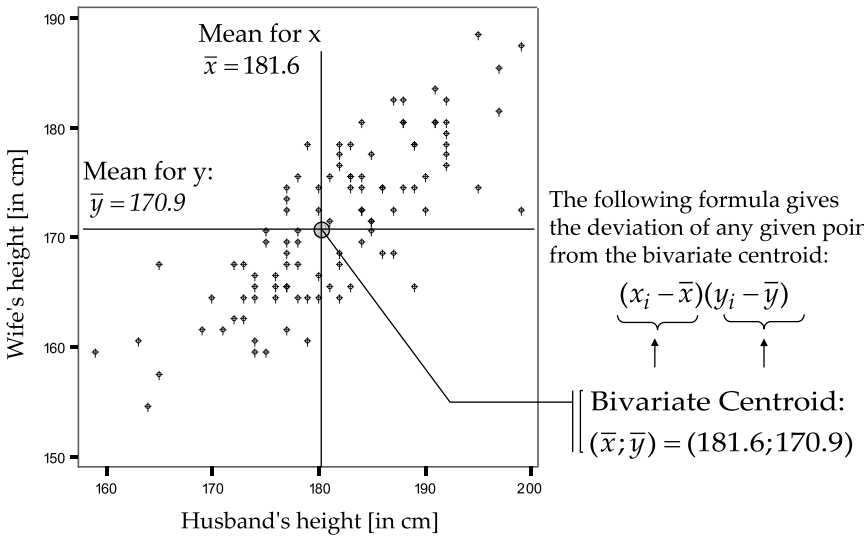


Fig. 4.19 Relationship of heights in married couples

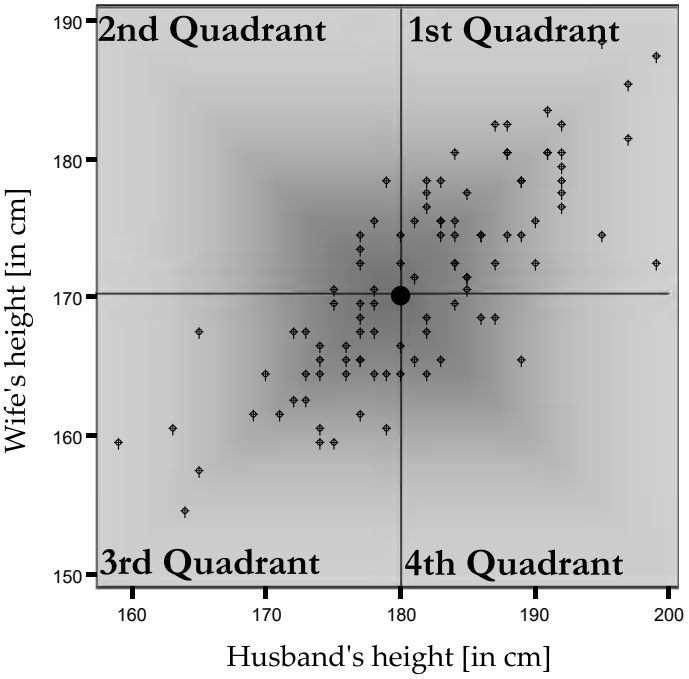


Fig. 4.20 Four-quadrant system

average height. The value pair of the bivariate centroid then becomes the centre of a new coordinate system with four quadrants (see Fig. 4.20).

All points in quadrant one involve marriages between men and women of above-average heights. When the values of quadrant one are entered into the equation $(x_i - \overline{x}) \cdot (y_i - \overline{y})$, the results are always positive. All points in quadrant three involve marriages between men and women of below-average heights. Here too, values fed into the equation $(x_i - \overline{x}) \cdot (y_i - \overline{y})$ produce positive results, as the product of two negative values is always positive.

All the data points in quadrant one and three are located at positive intervals to the bivariate centroid, with intervals being measured by the product $(x_i - \overline{x}) \cdot (y_i - \overline{y})$. This makes sense: the cloud of points formed by the data has a positive slope.

Quadrant number two contains data from taller-than-average women who married shorter-than-average men, while quadrant number four contains data from shorter-than-average women who married taller-than-average men. For these observations, the product of $(x_i - \overline{x}) \cdot (y_i - \overline{y})$ is always negative, which means that their intervals to the bivariate centroid are negative as well. All observed pairs in these quadrants form a cloud of points with a negative slope.

When calculating the strength of the relationship between heights, the important thing is the magnitude of the sum of the positive intervals in quadrants one and

three compared with the sum of the negative intervals in quadrants two and four. The larger the sum of the intervals in quadrants one and three, the larger the positive intervals to the bivariate centroid. The sum of positive and negative intervals in this example produces a positive value, which indicates a positive association between groom height and bride height. If the intervals in quadrants one and three are similar to those in quadrants two and four, the negative and positive intervals to the bivariate centroid cancel each other out and produce a value close to zero. In this case, there is no relationship between the variables, which is to say, there are almost as many taller-than-average (resp. shorter-than-average) grooms marrying taller-than-average (resp. shorter-than-average) brides as taller-than-average (resp. shorter-than average) brides marrying shorter-than-average (resp. taller-than-average) grooms. The last case to consider is when there are relatively large total deviations in quadrants two and four. In this case, there are many negative intervals and few positive deviations from the bivariate centroid, which produces in sum a negative value. The relationship between the variables *groom height* and *bride height* is hence negative.

As should be clear, the sum of intervals between the data points and the bivariate centroid offers an initial measure of the relationship between the variables. Dividing this sum by the number of observations yields the *average deviation from the bivariate centroid*, also known as covariance:

$$\operatorname{cov}(x; y) = S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy} \tag{4.20}$$

If covariance is positive, then the relationship between two metric variables may be positive. If the covariance is negative, then the relationship may be negative. If the covariance is zero or close to it, there tends to be no linear relationship between the variables. Hence, all that need interest us with covariance at this point is its algebraic sign.

If we briefly recall the sections on nominal variables, we'll remember that the χ^2 coefficient assumes the value zero when no association exists, and tends to climb as the strength of the relationship increases. We'll also remember an unfortunate feature of the χ^2 coefficient: its value tends to rise with the size of the sample and with the number of rows and columns in the contingency table. A similar problem applies to covariance. It can indicate the general direction of a relationship (positive or negative) but its size depends on the measurement units being used. This problem can be avoided by dividing by the standard deviation of variables x and y. The result is called Pearson's correlation coefficient.

$$r = \frac{S_{xy}}{S_x S_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right) \cdot \left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2\right)}} \quad \text{with} \quad -1 \le r \le +1 \tag{4.21}$$

The values of Pearson's correlation coefficient always lie between $r = (-1)$ and $r = (+1)$. The closer the correlation coefficient is to $r = (+1)$, the stronger the linear positive relationship is between the variables. If all data points lie along an upwards sloping line, the correlation coefficient assumes the exact value $r = (+1)$. If all data points lie along a downward-sloping line, the correlation coefficient assumes the exact value $r = (-1)$. If no linear relationship between the variables can be discerned, then the correlation coefficient has the value of $r = 0$ (or thereabouts). At what point does the correlation coefficient indicate a linear relationship? Researchers commonly draw the following distinctions:

$\|r\| < 0.5$	Weak linear association
$0.5 \leq \|r\| < 0.8$	Moderate Linear Association
$\|r\| \geq 0.8$	Strong linear association

4.4 Relationships Between Ordinal Variables

Sometimes the conditions for using Pearson's correlation coefficient are not met. For instance, what do we do when one or both variables have an ordinal scale instead of a metric scale? What do we do when the relation is not linear but monotonic? Let's look at some practical examples:

- Despite strongly linear-trending datasets, outliers can produce a low Pearson's correlation coefficient. Figure 4.21 illustrates this case. It juxtaposes the advertising expenditures of a firm with the market share of the advertised product. Both clouds of points are, except for one case, completely identical. In part 1 there is a very strong linear relationship between advertising expenditures and market share: $r = 0.96$. But, as part 2 shows, if you shift one point to the right, the correlation coefficient shrinks to $r = 0.68$. Pearson's correlation coefficient is, therefore, very sensitive to outliers, and this restricts its reliability. What we want is a more robust measure of association.

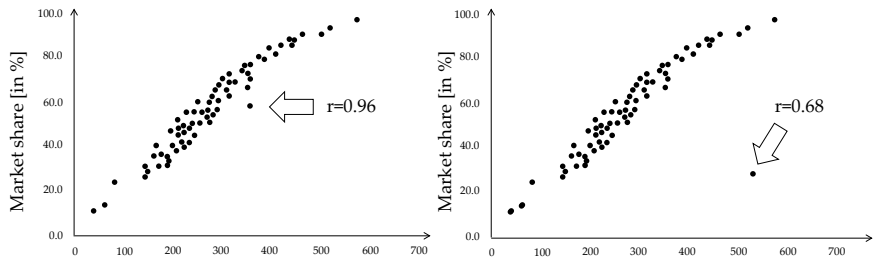


Fig. 4.21 Pearson's correlation coefficient with outliers

- Figure 4.22 displays an excerpt of a survey that asked people to rate the design of a wine bottle and indicate how much they'd pay for it on a five-point scale. Because the variables are not metrically scaled, we cannot use Pearson's coefficient to calculate correlation.
- The survey found a non-linear relationship between the respondents' ratings and willingness-to-pay, as shown in Fig. 4.23. Due to this non-linearity, we can expect the Pearson's correlation coefficient to be low. The relationship shown in the figure is nevertheless monotonic: as the rating increases, and the rate of increase changes, so too does the price respondents are willing to pay. With a linear relationship, the rates of change are constant; we need a measure of association that can assess the strength of monotonic relationships as well.

Question 8: How do you rate the design of the wine bottle on a scale from 1 (poor) to 5 (excellent)?

poor	☐	☐	☐	☐	☐	excellent
	1	2	3	4	5	

Question 9: How much would you pay for this bottle of wine?

☐	☐	☐	☐	☐
€5 or less	€5.01–10	€10.01–15	€15.01–20	€20.01–25

Fig. 4.22 Wine bottle design survey

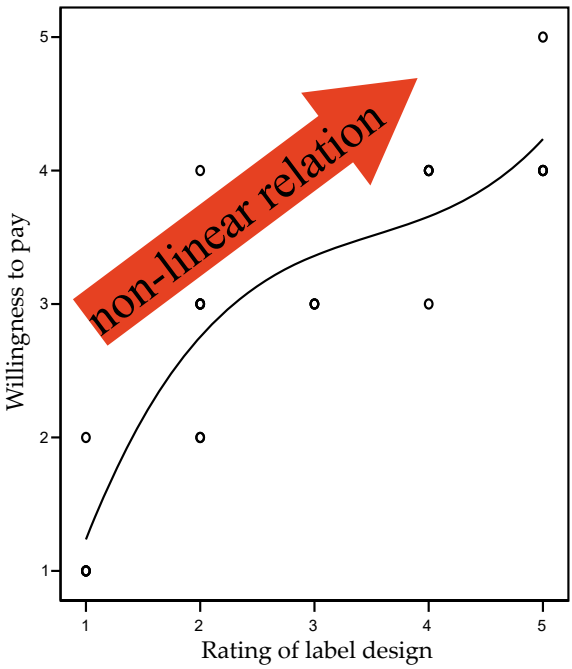


Fig. 4.23 Non-linear relationship between two variables

Fortunately, there are two options: *Spearman's rho (ρ)* and *Kendall's tau (τ)*. Either of these can be used when the conditions for using Pearson's correlation coefficient—i.e. metric scales and linear relationships—are not fulfilled and the dataset contains ordinal variables or monotonic metric relationships.

4.4.1 Spearman's Rank Correlation Coefficient (Spearman's Rho)

Spearman's rank correlation coefficient describes a *monotonic* relationship between ranked variables. The coefficient can assume values between $\rho = (-1)$ and $\rho = (+1)$. It has a value of $\rho = (+1)$ when two paired ordinal or metric variables have a perfect monotonic and positive relationship, i.e. when all observed values lie on a curve whose slope increases constantly but at various rates, as can be seen in Fig. 4.23. By contrast, the coefficient assumes a value of $\rho = (-1)$ when there is a perfect negative monotonic relationship between two variables (i.e. when all observed values lie along a curve whose slope decreases constantly but at various degrees). The more the value of the coefficient approaches zero, the less the value pairs share a perfect monotonic relationship.

The basic idea of Spearman's rho is to create a ranking order for each dataset and then to measure the difference between the ranks of each observation. Spearman's rho treats the ranking orders as cardinal scales by assuming that the distances between them are equidistant. From a theoretical perspective, this is an impermissible assumption. We'll have a closer look at this issue later. To better understand Spearman's rho, let's look at an example.

Imagine you conduct the survey in Fig. 4.22. You ask 25 persons to rate the design of a wine bottle and say how much they'd be willing to pay for it on a five-point scale. You code the results and enter them into a computer as it is presented in Fig. 4.24.

First the dataset is sorted by the value size of one of both variables. In Fig. 4.24 this has already been done for the bottle design rating (variable: *bottle*). The next step is to replace the values of the variable with their rankings. Twenty-five ranks are given, one for each respondent, as in a competition with twenty-five contestants. Each receives a rank, starting at 1st place and ending at 25th place.

Five survey respondents rated the bottle design as poor, and were assigned the value one in the ranking order. Each of these respondent values share 1st place, as each indicated the lowest trait value. What do we do when ranks are tied, i.e. when observations share the same trait values?

The first solution is to use the approach found in athletic competitions. For instance, when three competitors tie for first in the Olympics, each receives a gold medal. Silver and bronze medals are not awarded, and the next placed finisher is ranked 4th. Proceeding analogously, we can assign each observation of *poor* a rank of 1. But as we have already seen multiple times, statistics is first and foremost a

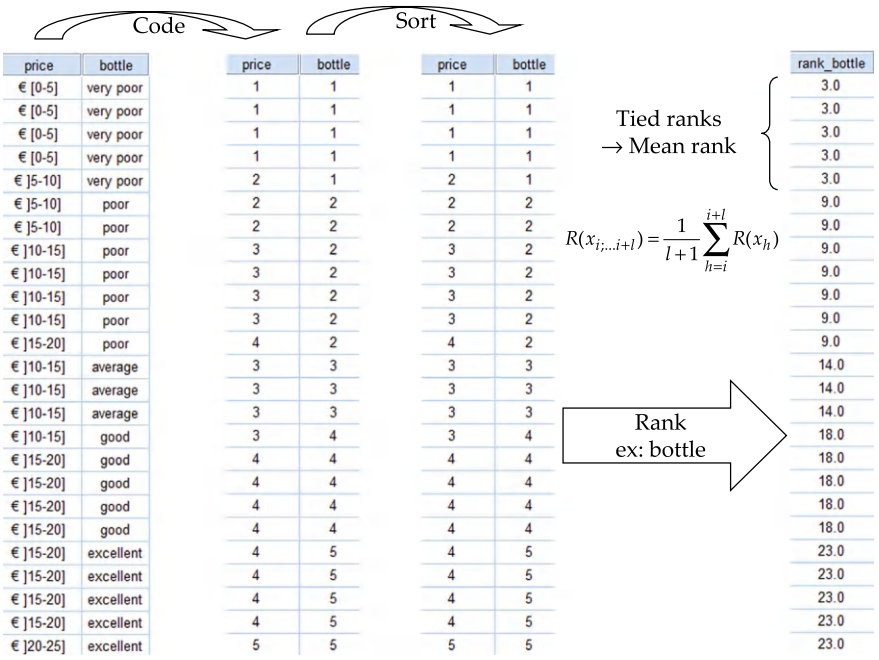


Fig. 4.24 Data for survey on wine bottle design

discipline of averages. In the case of a three-way tie for first, therefore, statistics must determine a *mean rank*. To do this, we award each top-three finisher 1/3 gold (1st place), 1/3 silver (2nd place) silver, and 1/3 bronze (3rd place):

$$1/3 \cdot 1 + 1/3 \cdot 2 + 1/3 \cdot 3 = 1/3 \cdot (1 + 2 + 3) = 2 \tag{4.22}$$

Why use the mean rank approach in statistics? The reason is simple. Assume there are eight contestants in a race, each with a different finishing time. Adding up their place ranks we get $1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 = 36$. Now assume that of the eight contestants, three tie for first. If we use the so-called Olympic solution, the sum of their ranks is 32 $(1 + 1 + 1 + 4 + 5 + 6 + 7 + 8)$. Using the mean rank approach, the sum of their ranks remains $36 = (2 + 2 + 2 + 3 + 4 + 5 + 6 + 7 + 8)$.

Now let's consider the wine bottle design survey from the vantage point of mean rank. Five respondents rated the design as poor $(= 1)$. Using the mean rank approach, each observation receives a three, as

$$1/5 \cdot (1 + 2 + 3 + 4 + 5) = 3. \tag{4.23}$$

Seven respondents rated the design as fair (= 2), occupying places six through twelve in the ranking order. Using the mean rank approach here, each observation receives a 9, as

$$1/7 \cdot (6 + 7 + 8 + 9 + 10 + 11 + 12) = 9. \tag{4.24}$$

We can proceed analogously for the other trait values:

- value of trait 3: $1/3 \cdot (13 + 14 + 15) = 14$
- value of trait 4: $1/5 \cdot (16 + 17 + 18 + 19 + 20) = 18$
- value of trait 5: $1/5 \cdot (21 + 22 + 23 + 24 + 25) = 23$.

The data on willingness-to-pay must be ranked the same way. After sorting the variables and assigning ranks using the method above, we obtain the results in Fig. 4.25, which includes the rating dataset as well.

Now we apply the product-moment correlation coefficient, but instead of Pearson's coefficient, in this case we use Spearman's. Accordingly, we must replace

y_i	x_i	$R(y_i)$	$R(x_i)$	$R(y_i)$-ØR(y)	$R(x_i)$-ØR(x)	$[R(y_i)$-ØR(y)]* [R(xi)-ØR(x)]	$(R(y_i)$-ØR(y))^2	$(R(x_i)$-ØR(x)^2	d^2
1	1	2.5	3.0	-10.5	-10.0	105.0	110.3	100.0	0.3
1	1	2.5	3.0	-10.5	-10.0	105.0	110.3	100.0	0.3
1	1	2.5	3.0	-10.5	-10.0	105.0	110.3	100.0	0.3
1	1	2.5	3.0	-10.5	-10.0	105.0	110.3	100.0	0.3
2	1	6.0	3.0	-7.0	-10.0	70.0	49.0	100.0	9.0
2	2	6.0	9.0	-7.0	-4.0	28.0	49.0	16.0	9.0
2	2	6.0	9.0	-7.0	-4.0	28.0	49.0	16.0	9.0
3	2	11.5	9.0	-1.5	-4.0	6.0	2.3	16.0	6.3
3	3	11.5	14.0	-1.5	1.0	-1.5	2.3	1.0	6.3
3	4	11.5	18.0	-1.5	5.0	-7.5	2.3	25.0	42.3
3	2	11.5	9.0	-1.5	-4.0	6.0	2.3	16.0	6.3
3	3	11.5	14.0	-1.5	1.0	-1.5	2.3	1.0	6.3
3	2	11.5	9.0	-1.5	-4.0	6.0	2.3	16.0	6.3
3	2	11.5	9.0	-1.5	-4.0	6.0	2.3	16.0	6.3
3	3	11.5	14.0	-1.5	1.0	-1.5	2.3	1.0	6.3
4	2	20.0	9.0	7.0	-4.0	-28.0	49.0	16.0	121.0
4	4	20.0	18.0	7.0	5.0	35.0	49.0	25.0	4.0
4	4	20.0	18.0	7.0	5.0	35.0	49.0	25.0	4.0
4	4	20.0	18.0	7.0	5.0	35.0	49.0	25.0	4.0
4	4	20.0	18.0	7.0	5.0	35.0	49.0	25.0	4.0
4	5	20.0	23.0	7.0	10.0	70.0	49.0	100.0	9.0
4	5	20.0	23.0	7.0	10.0	70.0	49.0	100.0	9.0
4	5	20.0	23.0	7.0	10.0	70.0	49.0	100.0	9.0
4	5	20.0	23.0	7.0	10.0	70.0	49.0	100.0	9.0
5	5	25.0	23.0	12.0	10.0	120.0	144.0	100.0	4.0
Sum		325.0	325.0	0.0	0.0	1070.0	1191.0	1240.0	291.0
Mean		13.0	13.0	0.0	0.0	42.8	47.6	49.6	11.6

Fig. 4.25 Rankings from the wine bottle design survey

the original values for x and y with the rankings $R(x)$ and $R(y)$, and the original mean $\bar{x}$ and $\bar{y}$ with the mean rank $\overline{R(x)}$ and $\overline{R(y)}$:

$$\rho = \frac{S_{xy}}{S_x S_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i) - \overline{R(x)}\right) \cdot \left(R(y_i) - \overline{R(y)}\right)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i) - \overline{R(x)}\right)^2\right) \cdot \left(\frac{1}{n}\sum_{i=1}^{n}\left(R(y_i) - \overline{R(y)}\right)^2\right)}} \tag{4.25}$$

When we plug the above data into this formula, we get the following results:

- $\overline{R(x)} = \overline{R(y)} = \frac{1}{25}(1 + 2 + 3 + \cdots + 25) = \frac{1}{n} \cdot \frac{n \cdot (n+1)}{2} = \frac{1}{25} \cdot \frac{25 \cdot (25+1)}{2} = 13$
- $\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i) - \overline{R(x)}\right)^2 = \frac{1}{25}\left((3-13)^2 + \cdots + (23-13)^2\right) = \frac{1240}{25} = 49.6$
- $\frac{1}{n}\sum_{i=1}^{n}\left(R(y_i) - \overline{R(y)}\right)^2 = \frac{1}{25}\left((2.5-13)^2 + \cdots + (25-13)^2\right) = \frac{1191}{25} = 47.6$
- $\sum_{i=1}^{n}\left(R(x_i) - \overline{R(x)}\right)\left(R(y_i) - \overline{R(y)}\right) = ((3-13)(2.5-13)) + \cdots + ((23-13)(20-13)) = 42.8$

These, in turn, produce:

$$\begin{aligned}\rho &= \frac{\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i) - \overline{R(x)}\right) \cdot \left(R(y_i) - \overline{R(y)}\right)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i) - \overline{R(x)}\right)^2\right) \cdot \left(\frac{1}{n}\sum_{i=1}^{n}\left(R(y_i) - \overline{R(y)}\right)^2\right)}} \\ &= \frac{42.8}{\sqrt{49.6 \cdot 47.6}} = 0.880\end{aligned} \tag{4.26}$$

Calculating this formula by hand is time consuming. There is a short-hand version of the formula is frequently used:

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)} \quad \text{with} \quad d_i = (R(x_i) - R(y_i)) \tag{4.27}$$

We first calculate the difference between ranks for each value pair. In our wine bottle survey, the first row contains $d_1 = (2.5 - 3.0) = (-0.5)$. We then square and sum all the differences (see column d^2 in Fig. 4.25). This produces the following:

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot 291}{25 \cdot (25^2 - 1)} = \frac{1746}{15{,}600} = 0.888 \tag{4.28}$$

There is a slight deviation between this result ($\rho = 0.888$) and that of the full-length version of the formula ($\rho = 0.880$). The reason is that, strictly speaking, the simplified version may only be used when there are no tied ranks, which in our sample is not the case.

Some books on statistics say that the shortened formula produces only minor distortions compared with the full-length formula, provided the share of tied ranks are less than 20%. Results close to 20% should thus be interpreted with caution when using this method. Alternatively, you may use the following corrected formula (Bortz et al., 2000, p. 418).

$$\rho_{\text{corrected}} = \frac{2 \cdot \left(\frac{N^3-N}{12} - N\right) - T - U - \sum_{i=1}^{n} d_i^2}{2 \cdot \sqrt{\left(\frac{N^3-N}{12} - T\right) \cdot \left(\frac{N^3-N}{12} - U\right)}} \quad \text{with} \tag{4.29}$$

- T as the length of b tied ranks among x variables:

$$T = \frac{\sum_{i=1}^{b} \left(t_i^3 - t_i\right)}{12}, \tag{4.30}$$

  where t_i equals the number of tied ranks in the ith of b groups for the tied ranks of the x variables.
- U as the length of c tied ranks of y variables:

$$U = \frac{\sum_{i=1}^{c} \left(u_i^3 - u_i\right)}{12}, \tag{4.31}$$

  where u_i equals the number of tied ranks in the ith of c groups for the tied ranks of the y variables.

Of course, hardly anyone today calculates rank correlations by hand. Due to the importance of ordinal scales in social and economic research, Spearman's rank correlation has been implemented in all major statistics software packages. Nevertheless, Spearman's rank correlation has a very serious theoretical limitation: since it is calculated based on the differences between ranks and mean ranks, one must be able to show that consecutive ranks for the trait under investigation are equidistant from each other. With ordinal variables, this is hard to prove. For this reason, other rank correlation coefficients have come into use recently, especially those in the Kendall's tau (τ) coefficient family.

4.4.2 Kendall's Tau (τ)

Unlike Spearman's rank correlation, Kendall's τ does without the assumption of equidistant intervals between two consecutive ranks. It is derived from information permitted for ordinal variables. Kendall's τ thus places fewer demands on the data than Spearman's correlation does.

Two short examples serve to illustrate the basic idea of Kendall's τ. Let us assume a perfect positive monotonic relationship between variables x and y, as shown in Fig. 4.26.

As with Spearman's rank correlation, we first assign the variables x and y the ranks $R(x)$ and $R(y)$. The dataset is then sorted by the size of either $R(x)$ or $R(y)$. The ranking order ordered by size serves as the *anchor column*. The ranks in the anchor column are always ordered from smallest to largest. In Fig. 4.26 the anchor column is $R(x)$. The other ranking order—$R(y)$ in our example—serves as the *reference column*. If a perfect positive and monotonic association is present, the reference column is automatically ordered from smallest to largest, too. With a perfect negative and monotonic association, the reference column is automatically ordered from largest to smallest. Deviations from these extremes correspond to deviations from the monotonic association.

Kendall's τ uses this information and identifies the share of *rank disarray* in the reference column. The share of rank disarray is the percentage of cases in which the reference column deviates from the ranking order of the anchor column.

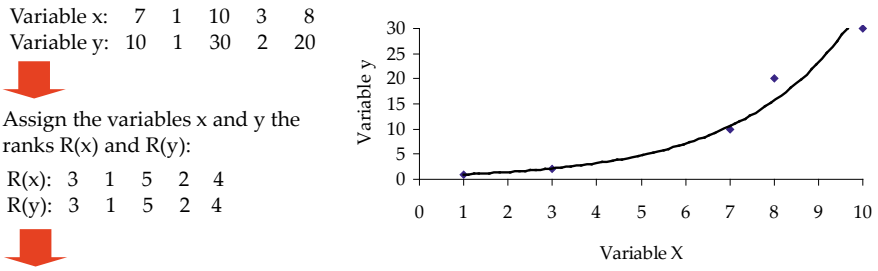


Fig. 4.26 Kendall's τ and a perfect positive monotonic association

First we compare all rank combinations in the reference column, beginning with the first value. If the rank of the first entry is smaller than the entry it is compared with, we have a *concordant pair*. If it is larger, it is called a *discordant pair*. Since in our example all reference ranks (2, 3, 4, 5) are larger than the first (1), we have $P = 4$ concordant pairs and no ($I = 0$) discordant pairs. Next we compare the second rank (2) of the reference column with the subsequent ranks (3, 4, 5) of the same row by size. A comparison with the first rank was already performed in the first step. This gives us three concordant pairs and no discordant pairs. We repeat this procedure with the other ranks in the reference column. Once all possible comparisons have been performed

$$-\text{in our example there are ten;}\quad \frac{n \cdot (n-1)}{2} = \frac{5 \cdot (5-1)}{2} = 10- \tag{4.32}$$

we determine the surplus of concordant pairs (P) to discordant pairs (I). In our example the surplus is ten: $(P - I) = (10 - 0) = 10$. In ten of ten comparisons, the reference column follows the increasing ranking order exactly—indication of a perfect positive and monotonic association. This finds expression in the formula for Kendall's τ_α:

$$\begin{aligned}\tau_a &= \frac{\text{No. of concordant pairs} - \text{No. of discordant pairs}}{n \cdot (n-1)/2} \\ &= \frac{P - I}{n \cdot (n-1)/2} = \frac{10 - 0}{10} = 1\end{aligned} \tag{4.33}$$

If the association was perfectly negative and monotonic, there would have been ten discordant pairs and no concordant pairs. For Kendall's τ_α we arrive at the following:

$$\tau_a = \frac{P - I}{n \cdot (n-1)/2} = \frac{0 - 10}{10} = (-1) \tag{4.34}$$

As with Spearman's rank correlation coefficient, the values of Kendall's τ_α lie between $\tau_\alpha = (-1)$ and $\tau_\alpha = (+1)$. If two paired ordinal or metric traits possess a perfect monotonic and positive association (i.e. if all values lie on a curve that rises constantly but at varying rates), the measure assumes the value $\tau_\alpha = (+1)$. By contrast, if there is a perfect negative monotonic association (i.e. if all values lie on a slope that falls constantly but at varying rates), it takes the value $\tau_\alpha = (-1)$. The more the value of the coefficient approaches $\tau_\alpha = 0$, the more the value pair deviates from a perfect monotonic association. This is because in such cases the ordering of the reference column is neither wholly positive nor wholly negative, resulting in both concordant pairs and discordant pairs. If there are an equal number of concordant pairs and discordant pairs, Kendall's τ_α assumes a value of $\tau_\alpha = 0$, as shown in Fig. 4.27:

$$\tau_a = \frac{P - I}{n \cdot (n-1)/2} = \frac{5 - 5}{10} = 0 \tag{4.35}$$

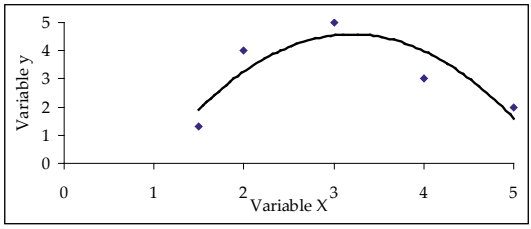
Variable x: 2 1.5 3 4 5
Variable y: 4 1.3 5 3 2



Assign the variables x and y the ranks R(x) and R(y):

R(x): 2 1 3 4 5
R(y): 4 1 5 3 2





The dataset is then sorted by size of either R(x) or R(y). R(x) in our case:

Anchor column (R(x)): 1 2 3 4 5
Reference column (R(y)): 1 4 5 3 2



Compare all rank combinations in the references column, beginning with the first value:

$R(y_1)$- $R(y_2)$ ⇨(+)
$R(y_1)$- $R(y_3)$ ⇨(+)
$R(y_1)$- $R(y_4)$ ⇨(+)
$R(y_1)$- $R(y_5)$ ⇨(+)

$R(y_2)$- $R(y_3)$ ⇨(+)
$R(y_2)$- $R(y_4)$ ⇨(-)
$R(y_2)$- $R(y_5)$ ⇨(-)

$R(y_3)$- $R(y_4)$ ⇨(-)
$R(y_3)$- $R(y_5)$ ⇨(-)

$R(y_4)$- $R(y_5)$ ⇨(-)

(+): Concordant pair; (-): Discordant pair

Fig. 4.27 Kendall's τ for a non-existent monotonic association

The simple formula Kendall's τ_a assumes that no tied ranks are present. If tied ranks are present, the corrected formula *Kendall's* τ_b should be used:

$$\tau_b = \frac{P - I}{\sqrt{\left(\frac{n\cdot(n-1)}{2} - T\right)\left(\frac{n\cdot(n-1)}{2} - U\right)}} \tag{4.36}$$

where

- T is the length of the b tied ranks of x variables:

$$T = \frac{\sum_{i=1}^{b} t_i(t_i - 1)}{2}, \tag{4.37}$$

  and t_i is the number of tied ranks in the ith of b groups of tied ranks for the x variables.
- U is the length of c tied ranks of the y variables:

$$U = \frac{\sum_{i=1}^{c} u_i(u_i - 1)}{2}, \tag{4.38}$$

  and u_i is the number of tied ranks in the ith of c groups of tied ranks for the y variables.

i	y	x	R(y)	R(x)	concordant pairs	discordant pairs
1	1	1	2.5	3.0	20	0
2	1	1	2.5	3.0	20	0
3	1	1	2.5	3.0	20	0
4	1	1	2.5	3.0	20	0
5	2	1	6.0	3.0	18	0
6	2	2	6.0	9.0	13	0
7	2	2	6.0	9.0	13	0
8	3	2	11.5	9.0	9	0
9	3	3	11.5	14.0	9	1
10	3	4	11.5	18.0	9	1
11	3	2	11.5	9.0	9	0
12	3	3	11.5	14.0	9	1
13	3	2	11.5	9.0	9	0
14	3	2	11.5	9.0	9	0
15	3	3	11.5	14.0	9	1
16	4	2	20.0	9.0	1	0
17	4	4	20.0	18.0	1	0
18	4	4	20.0	18.0	1	0
19	4	4	20.0	18.0	1	0
20	4	4	20.0	18.0	1	0
21	4	5	20.0	23.0	0	0
22	4	5	20.0	23.0	0	0
23	4	5	20.0	23.0	0	0
24	4	5	20.0	23.0	0	0
25	5	5	25.0	23.0	0	0
	Sum		325.0	325.0	197	4
	Mean		13.0	13.0		

Tied ranks in the anchor column R(y)

Tied ranks in the anchor column R(y)

Tied ranks in the anchor column R(y)

The existing order of the reference column R(x) – 3.0, 9.0, and 9.0 – is only one possible variation. The calculation of Kendall's τ_b assumes that tied ranks in the anchor column can lead concordant pairs and discordant pairs in the reference column to be overlooked.

Tied ranks in the anchor column R(y)

Fig. 4.28 Kendall's τ for tied ranks

The more tied ranks that are present in a dataset, the smaller the value of Kendall's τ_α compared with Kendall's τ_b. The practical application of this very complicated formula can be illustrated using our wine bottle survey (see Fig. 4.28).

After assigning ranks to the datasets *willingness-to-pay* (y) and *bottle design* (x), the rankings are ordered in accordance with the anchor column $R(y)$. Tied ranks are present for both ranking orders. For each of the first four ranks of the reference column $R(x)$—all with a value of 3.0—there are 20 concordant pairs and no discordant pairs, as 20 of the 25 observed values are larger than three. The fifth observation of the reference column $R(x)$ also has the value of 3.0. Here too, each of the 20 subsequent observations is larger than 3.0. Based on this information, we would expect 20 concordant pairs as well, but in reality there are only 18. Why?

The reason has to do with the tied ranks in the anchor column $R(y)$. Observations five to seven display a rank of 6.0 for all $R(y)$. The existing order of the reference column $R(x)$—3.0, 9.0, and 9.0—is only one possible variation; the sequence could also be 9.0, 9.0, and 3.0. Here too, the anchor column would be correctly ordered from smallest to largest. The calculation of Kendall's τ_b assumes that tied ranks in the anchor column can lead concordant pairs and discordant pairs in the reference column to be overlooked. For observation number five there are only 18 concordant pairs—all observation values between eight and 25. We proceed the same way with observation number eight. For observations eight to 15 there are eight tied ranks for the anchor column, whose grouping would be

random. Possible concordant pairs and discordant pairs are only considered for observations 16–25. For observation nine there are nine concordant pairs and one discordant pair.

This results in 197 concordant pairs and only four discordant pairs, so that:

$$\tau_b = \frac{197 - 4}{\sqrt{\left(\frac{25\cdot(25-1)}{2} - 73\right)\left(\frac{25\cdot(25-1)}{2} - 54\right)}} = 0.817 \tag{4.39}$$

and

- $T = \frac{\sum_{i=1}^{b} t_i(t_i-1)}{2} = \frac{4\cdot(4-1)+3\cdot(3-1)+8\cdot(8-1)+9\cdot(9-1)}{2} = 73$
- $U = \frac{\sum_{i=1}^{b} u_i(u_i-1)}{2} = \frac{5\cdot(5-1)+7\cdot(7-1)+3\cdot(3-1)+5\cdot(5-1)+5\cdot(5-1)}{2} = 54.$

Kendall's τ_b can also be calculated from a square contingency table. The datasets from our wine bottle survey can be inserted into the square contingency table in Fig. 4.29. The observations in the contingency table's rows and columns represent the value pairs subjected to the anchor column/reference column procedure.

We derive the number of concordant pairs by comparing all existing rank combinations in the reference column $R(x)$. This produces the following calculation:

$$\begin{aligned} P = {} & 4 \cdot (2 + 4 + 1 + 3 + 1 + 4 + 4 + 1) + 1 \cdot (4 + 1 + 3 + 1 + 4 + 4 + 1) \\ & + 2 \cdot (3 + 1 + 4 + 4 + 1) + 4 \cdot (4 + 4 + 1) + 3 \cdot (4 + 4 + 1) \\ & + 1 \cdot (4 + 1) + 1 \cdot 1 + 4 \cdot 1 = 197 \end{aligned} \tag{4.40}$$

For discordant pairs, the reverse applies:

$$I = 4 \cdot 0 + 1 \cdot 0 + 2 \cdot 0 + 4 \cdot 0 + 3 \cdot 0 + 1 \cdot 0 + 1 \cdot (3 + 1) + 4 \cdot 0 = 4 \tag{4.41}$$

Kendall's τ_b can now be derived from the above formula. If Kendall's τ_b is derived from a *non-square contingency table*, the values $\tau_b = (+\ 1)$ and $\tau_b = (-$

R(y)

R(x)	2.5	6.0	11.5	20.0	25.0	Total
3.0	4	1				5
9.0		2	4	1		7
14.0			3			3
18.0			1	4		5
23.0				4	1	5
Total	4	3	8	9	1	25

Fig. 4.29 Deriving Kendall's τ_b from a contingency table

1) can never be reached, even if the association is perfectly monotonic. Instead we must calculate *Kendall's* τ_c:

$$\tau_c = \frac{2 \cdot \min[\#\text{rows};\ \#\text{columns}] \cdot (P - I)}{(\min[\#\text{rows};\ \#\text{columns}] - 1) \cdot n^2} \tag{4.42}$$

The example from Fig. 4.29 yields the following calculation:

$$\tau_c = \frac{2 \cdot \min[5; 5] \cdot (197 - 4)}{(\min[5; 5] - 1) \cdot 25^2} = \frac{2 \cdot 5 \cdot (193)}{(5 - 1) \cdot 25^2} = 0.772 \tag{4.43}$$

4.5 Measuring the Association Between Two Variables with Different Scales

In previous sections, we discussed the measures of association between two nominal, two ordinal, and two metric variables. But what about the association between two variables with different scales? For instance, how can we measure the association between the nominally scaled variable *gender* and the metrically scaled variable *age*. Below I briefly discuss some examples.

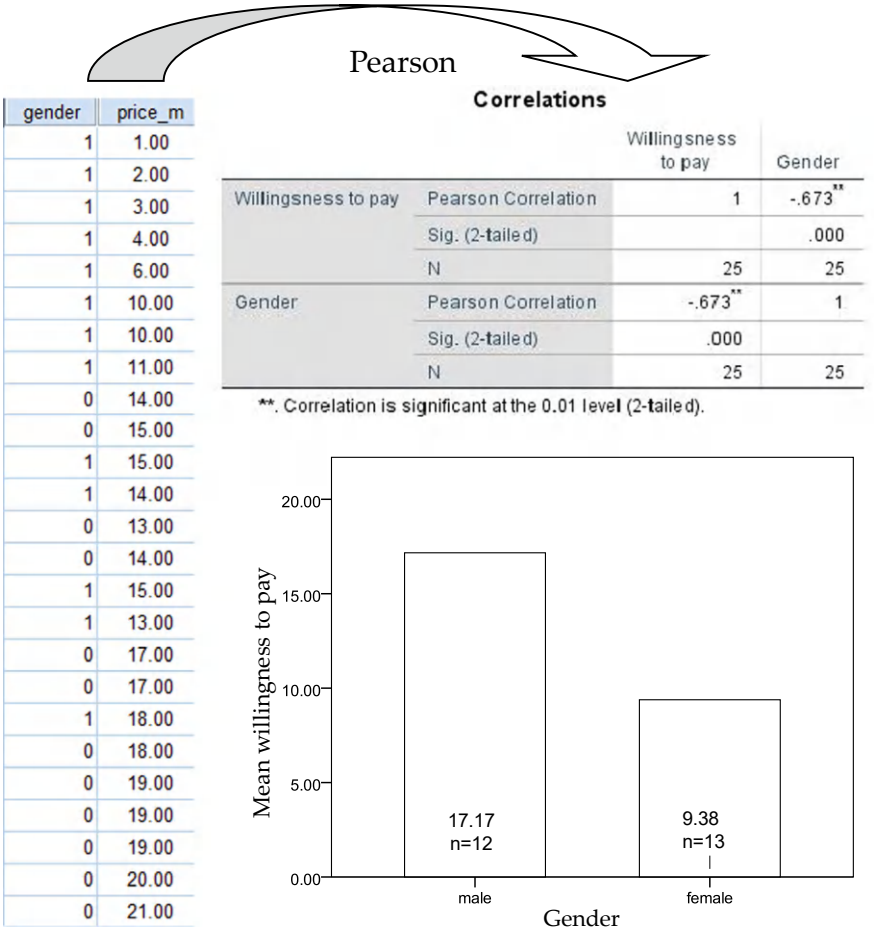
4.5.1 Measuring the Association Between Nominal and Metric Variables

There is no commonly applied measure of correlation for nominal and metric variables. The following alternatives are recommended:

- In practice, statisticians usually apply *statistical tests* (see Sect. 9) to assess differences between nominal groups with regard to metric variables.
- It is also possible to convert metric variables into ordinal variables via classification and then use an appropriate method such as Cramer's V. But this method is fairly uncommon in practice.
- Another seldom used approach is the *point-biserial correlation* (r_{pb}). It measures the association between a dichotomous variable (a special case of a nominal scale with only two possible outcomes) and a metric variable.

Let's discuss the last case in more detail using our wine bottle survey. Imagine that the survey asks respondents to indicate their gender and how much they'd pay in whole euro amounts. *Willingness-to-pay* is now a metric variable (*price_m*) and gender is a dichotomous variable (*gender*)—zero for *male* and one for *female*. The results are shown in Fig. 4.30.

Ordering mean values by gender, we discover that on average male respondents are willing to pay €17.17 and female respondents are willing to play €9.38.



gender	price_m
1	1.00
1	2.00
1	3.00
1	4.00
1	6.00
1	10.00
1	10.00
1	11.00
0	14.00
0	15.00
1	15.00
1	14.00
0	13.00
0	14.00
1	15.00
1	13.00
0	17.00
0	17.00
1	18.00
0	18.00
0	19.00
0	19.00
0	19.00
0	20.00
0	21.00

Correlations

		Willingsness to pay	Gender
Willingsness to pay	Pearson Correlation	1	-.673**
	Sig. (2-tailed)		.000
	N	25	25
Gender	Pearson Correlation	-.673**	1
	Sig. (2-tailed)	.000	
	N	25	25

**. Correlation is significant at the 0.01 level (2-tailed).

Fig. 4.30 Point-biserial correlation

Willingness-to-pay is thus higher on average with men than with women. Can we infer from these results an association between gender and willingness-to-pay?

The point-biserial correlation can be used to determine the strength of association in cases like these. This approach assumes that Pearson's correlation can be used to measure the association between a dichotomous variable and a metric variable. This surprising assumption is possible because variables coded as either zero or one can be regarded metrically. Applied to our case: if the value of the variable *gender* is one, the more female the respondent is. If the value of the variable *gender* is zero, the more male the respondent is. Using Pearson's correlation for both variables, we get a correlation coefficient between $r_{pb} = (-1)$ and $r_{pb} = (+1)$.

The lower limit $r_{pb} = (-1)$ means that all respondents coded as zero (male) have higher values with the metric variable (willingness-to-pay) than respondents coded as one (female). By contrast, a point-biserial correlation of $r_{pb} = (+1)$ means that all respondents coded as zero (male) have lower values with metric variables (willingness-to-pay) than respondents coded as one (female). The more frequently higher and lower values appear mixed in the metric variable (willingness-to-pay), the less we can infer the value of the metric variable from gender, and vice versa, and the closer the point-biserial correlation approaches the value $r_{pb} = 0$.

Of course, the formula for Pearson's correlation can be used to calculate the point-biserial correlation. This formula can be simplified as follows:

$$r_{pb} = \frac{\bar{y}_1 - \bar{y}_0}{S_y}\sqrt{\frac{n_0 \cdot n_1}{n^2}}, \tag{4.44}$$

where

- n_0: number of observations with the value $x = 0$ of the dichotomous trait
- n_1: number of observations with the value $x = 1$ of the dichotomous trait
- n: total sample size $n_0 + n_1$
- $\bar{y}_0$: mean of metric variables (y) for the cases $x = 0$
- $\bar{y}_1$: mean of metric variables (y) for the cases $x = 1$
- S_y: standard deviation of the metric variable (y).

For our example, this results in the following:

$$r_{pb} = \frac{\bar{y}_1 - \bar{y}_0}{S_y}\sqrt{\frac{n_0 \cdot n_1}{n^2}} = \frac{9.38 - 17.17}{5.8}\sqrt{\frac{12 \cdot 13}{25^2}} = (-0.67) \tag{4.45}$$

The negative point-biserial correlation indicates that the respondents whose dichotomous variable value is one (female) show a lower willingness-to-pay than the respondents whose dichotomous variable value is zero (male).

The point-biserial correlation is usually applied only when a variable contains a true dichotomy. A true dichotomy occurs when a variable possesses only two possible values, such as *male* or *female*. By contrast, if a metric variable is dichotomized—for example, if two age groups are produced from metric age data—and the variable is distributed normally, the point-biserial correlation underestimates the actual association between the observed variables (see Bowers, 1972).

4.5.2 Measuring the Association Between Nominal and Ordinal Variables

Cramer's V is a common tool for measuring the strength of association between a nominal and an ordinal variable, provided the number of possible outcomes for the ordinal variable is not too large. Statistical tests (Mann–Whitney or Kruskal–Wallis, see Sects. 9.4.2 and 9.5.2) are frequently used in empirical practice, as it's usually less about the association between (nominal) groups with regard to ordinal variables than about their distinctions.

In the special case of a dichotomous nominal variable, we can also use a *biserial rank correlation.* When there are no tied ranks, association can be calculated as follows (Glass, 1966):

$$r_{\text{bisR}} = \frac{2}{n} \cdot \left(\overline{R(y_1)} - \overline{R(y_0)} \right), \tag{4.46}$$

where

- n: total sample size $n_0 + n_1$
- $\overline{R(y_0)}$: mean rank for nominal cases $x = 0$ of ordinal variables (y)
- $\overline{R(y_1)}$: mean rank for nominal cases $x = 1$ of ordinal variables (y).

4.5.3 Association Between Ordinal and Metric Variables

Janson and Vegelius (1982) made some proposals for just such a measure of correlation, but these parameters have never been of much importance for researchers or practitioners. This is mostly because the simplified approaches for using Spearman's correlation coefficient or Kendall's τ are more than adequate. There are two such approaches:

1. Classify the metric variable and convert it into an ordinal scale. This produces two ordinal variables whose monotonic association can be determined by Spearman's correlation or Kendall's τ.
2. Subject the observations from the metric variable unclassed to the usual rank assignment. This also produces two ordinal ranking orders.

To illustrate, let us turn again to the wine bottle survey but change it somewhat: instead of a five-point score (ordinal scale), the 25 respondents now indicate their willingness-to-pay in euros (metric scale). We obtain the results shown in Fig. 4.31.

Here the metrically scaled willingness-to-pay variable (*price_m*) is converted into a ranking order (*rprice*). This eliminates information about the interval between one person's willingness-to-pay and another's, but it preserves their ranking. The conversion of the metric dataset into a ranking order replaces a higher scale (metric) with a lower scale (ordinal). The price is relatively small—we can

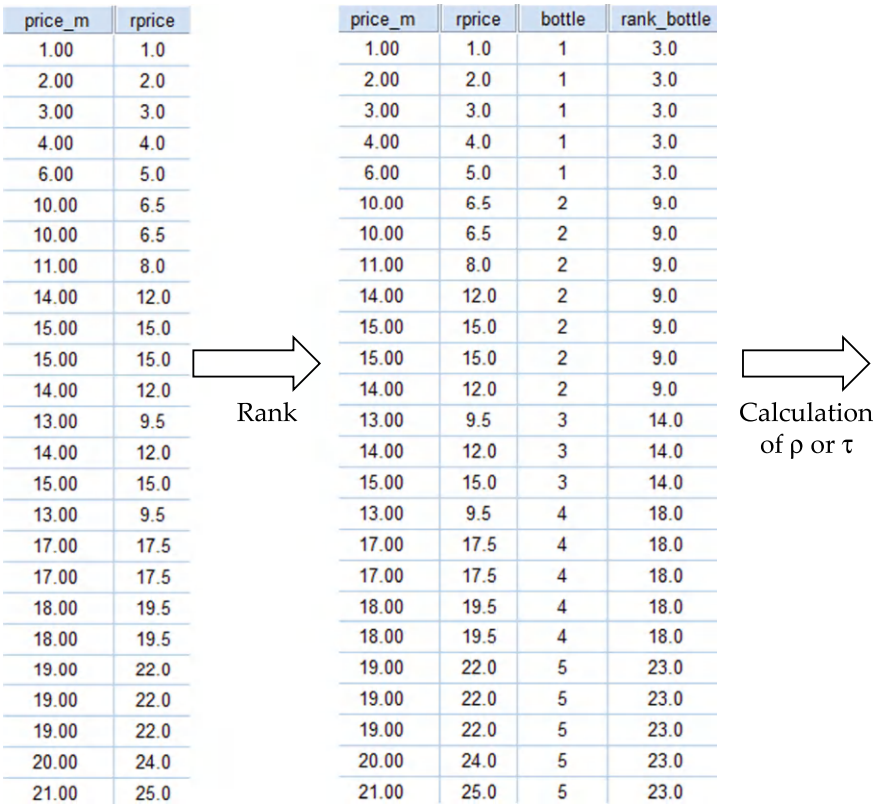


Fig. 4.31 Association between two ordinal and metric variables

make statements only about the monotonic association—which explains the failure of other coefficients proposed to measure the association between ordinal and metric variables.

4.6 Calculating Correlation with a Computer

When using R, SPSS or Stata to calculate ρ and τ, rank assignment occurs automatically, sparing us the extra step, and the original metric or ordinal variables can be entered directly. With Excel we need to calculate variable rank before proceeding.

4.6.1 Calculating Correlation with R and R Commander

In R Commander, calculate Pearson's correlation by selecting *Statistics* → *Summaries* → *Correlation test*… to open the respective dialogue box. Select

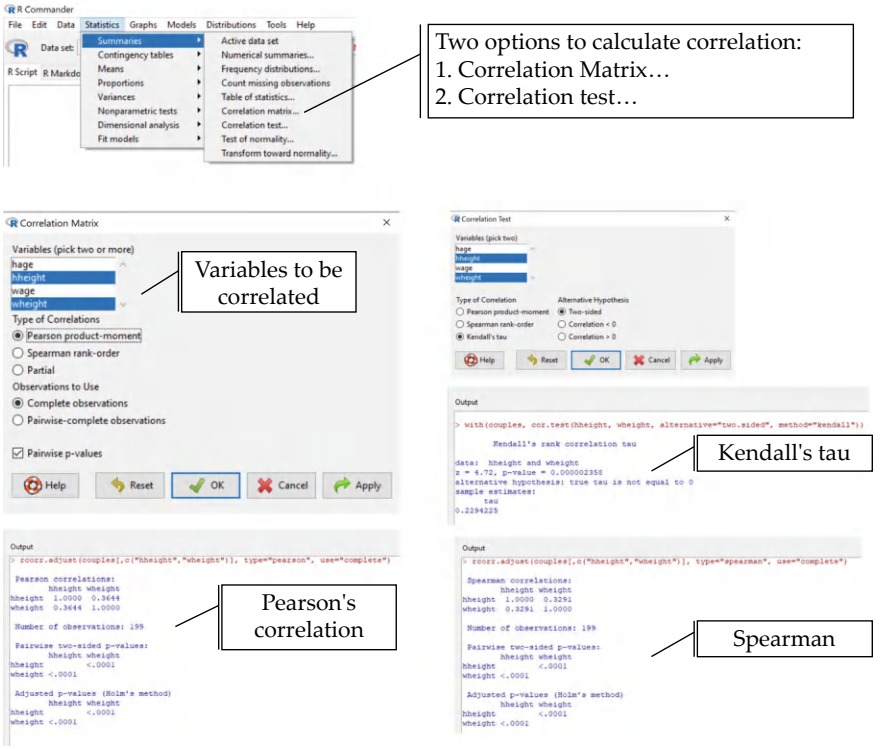


Fig. 4.32 Calculating correlation with R and R Commander

the two variables to be correlated and use *Pearson product-moment* to calculate the linear correlation between two metric variables or use *Kendall's* τ_b or *Spearman rank-order* when determining the monotonic correlation between two metric or ordinal variables (see Fig. 4.32).

In the example of the heights of couples getting married, we select the variables *husband's height* (hheight) and *wife's height* (wheight). The results are shown in Fig. 4.32. Pearson's correlation has the value $r = 0.364$, Kendall's τ_b has the value $\tau_b = 0.229$, and Spearman's correlation has the value $\rho = 0.329$.

4.6.2 Calculating Correlation with SPSS

In SPSS, calculate Pearson's correlation by selecting *Analyze* → *Correlate* → *Bivariate…* to open the *Bivariate Correlations* dialogue box. Before selecting the desired correlation (Pearson, Kendall's τ_b, or Spearman), we need to think about the scale of the variables to be correlated. Use the Pearson correlation when calculating the linear association between two metric variables. Use Kendall's τ_b or Spearman's correlation when determining the monotonic

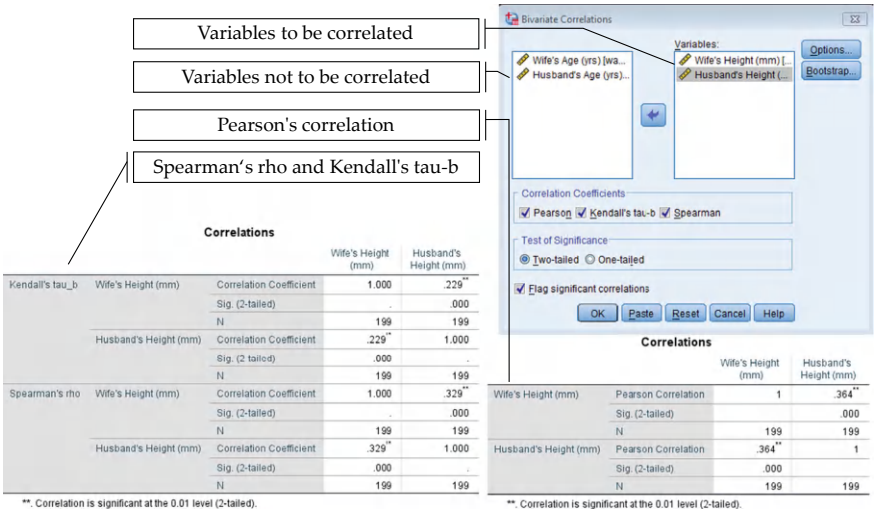


Fig. 4.33 Calculating correlation with SPSS

association between two metric or ordinal variables. Mark the variables to be correlated and click the middle arrow to move them to the field *variables* (see Fig. 4.33). Then click *OK* to carry out the calculation.

In the example of the heights of couples getting married, we select the variables *husband's height* (hheight) and *wife's height* (wheight). The results are shown in Fig. 4.33. Pearson's correlation has the value $r = 0.364$, Kendall's τ_b has the value $\tau_b = 0.229$, and Spearman's correlation has the value $\rho = 0.329$.

4.6.3 Calculating Correlation with Stata

Unlike SPSS, the command windows for calculating the three correlation coefficients in Stata are not in the same place. Select *Statistics* → *Summaries, tables, and tests* → *Summary and descriptive statistics* → *Correlations and covariances* to open the dialogue box for calculating Pearson's correlation. For Spearman's rank correlation or Kendall's rank correlation, select *Statistics* → *Summaries, tables, and tests* → *Nonparametric tests of hypotheses*, and then choose the desired correlation coefficient.

In the first input line (*Variables [leave empty for all]*) enter the variables to be correlated. In our example these are the heights of the grooms (hheight) and brides (wheight). This information is sufficient for calculating Pearson's correlation coefficient. Click *OK* or *Submit* to execute the Stata command (see Fig. 4.34).[5]

[5] Syntax command: *correlate hheight wheight*.

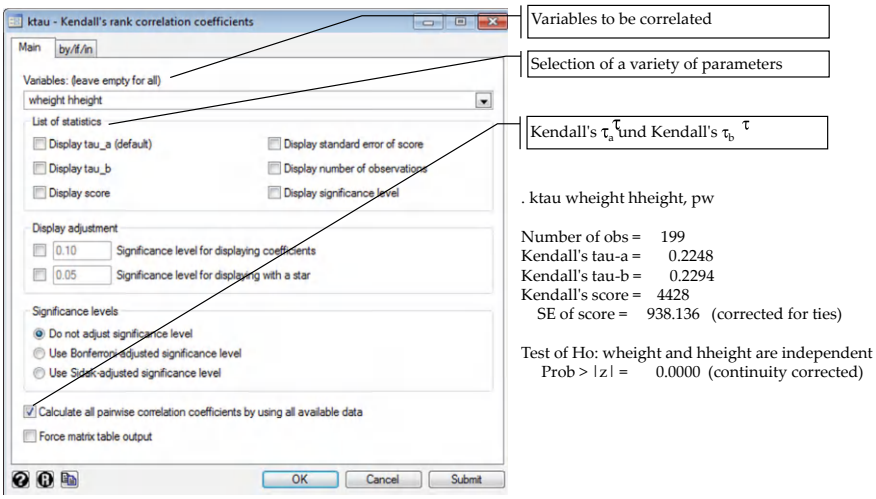


Fig. 4.34 Calculating correlation with Stata (Kendall's τ)

In the dialogue box for calculating Spearman's correlation or Kendall's τ you can also select a variety of parameters under the submenu *List of statistics*. It is recommended, however, that all Kendall and Spearman coefficients be calculated using the command *Calculate all pairwise correlation coefficients by using all available data*. Click *OK* or *Submit* to execute the Stata command.[6] For Kendall's τ, we get the values $\tau_a = 0.225$ and $\tau_b = 0.229$. Spearman's correlation can be calculated in the same manner.

4.6.4 Calculating Correlation with Excel

Excel has a pre-programmed function for calculating Pearson's correlation coefficient. To use it, move the cursor over the cell whose correlation coefficient is to be calculated and mark it. Then go to *Formulas* and *Insert Function* to select the category *Statistical* and the function *Correl*. Enter the array of both variables into the fields Matrix1 and Matrix2, respectively. For our wedding ceremony, the height data for grooms goes in cell D2:D202 and the height data for brides goes in cell C2:C202. The correlation result updates automatically whenever the original data in the predefined cells are changed.

In Excel, Spearman's correlation must be calculated manually, which requires some extra effort. First we assign ranks to the variables, converting the original metric data into ranked sets. In Sect. 4.4.1, we learned that Spearman's correlation

[6] Syntax command for Kendall's tau: *ktau hheight wheight, pw*. Syntax command for Spearman's rho: *ktau hheight wheight, pw*.

is a Pearson's correlation with ranked variables. Excel possesses a rank function (*RANK*) but the calculation is not based on mean ranks. Whenever ranks are tied, Excel assigns the lowest rank to each. This is the "Olympic solution" discussed above. To determine average ranks for tied ranks, use the following correction factor:

$$[\text{Count(Field)} + 1 - \text{RANK(Cell; Field; 0)} - \text{RANK(Count; Field; 1)}]/2 \quad (4.47)$$

Field describes the arrays containing the values of the two variables (e.g. A2:B12). This correction factor must be added to every tied rank:

$$\text{RANK(Cell; Field; 1)} + \text{Correction factor} \quad (4.48)$$

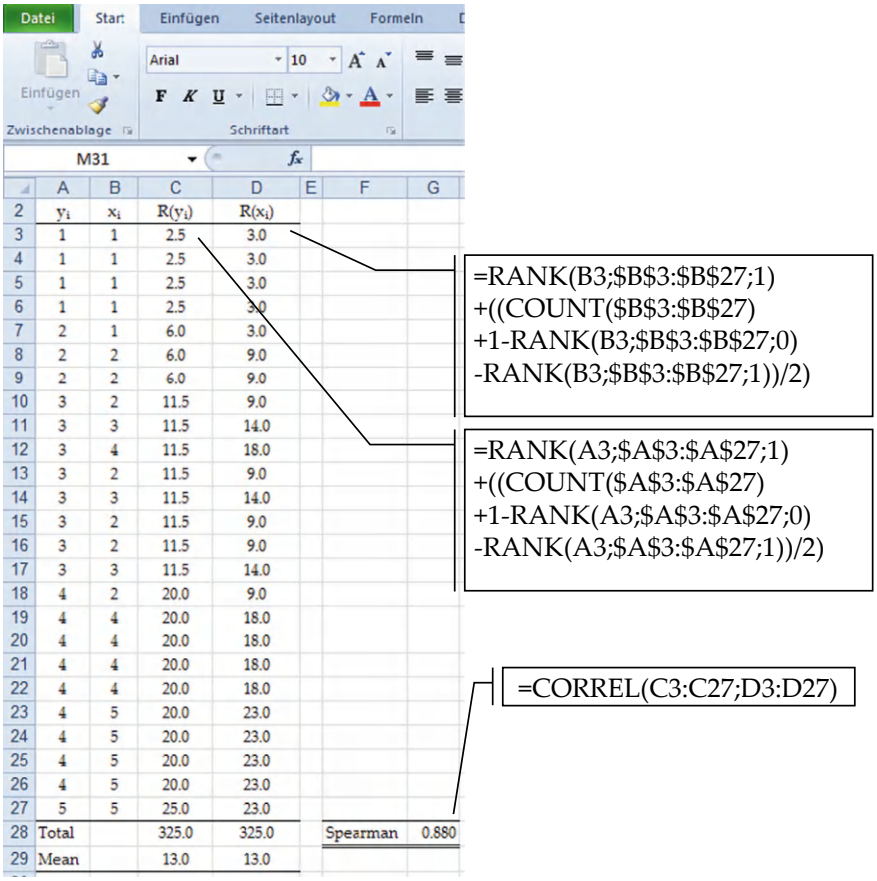
The Excel formula for the correlation coefficient can be applied to the corrected ranks *Correl(Array1; Array2)*. Figure 4.35 shows once again how to calculate Spearman's correlation with Excel.

Calculating Kendall's τ for larger datasets is laborious with Excel. The command = *COUNTIF(field; condition)* can be used to help count concordant pairs and discordant pairs, but the condition must be entered for each row (observation) separately, which is why standard Excel commands should not be used for calculating Kendall's τ. Fortunately, add-ins can be purchased for Excel that makes Kendall's τ easier to calculate.

4.7 Spurious Correlations

Correlation is a statistical method that provides information about a linear or monotonic relationship between two measured variables. If the value of the correlation coefficient is $r = 0$ or thereabouts, we can usually assume that no linear association exists. If the correlation coefficient is relatively large, we can assume the variables are related in some way, but we may not necessarily assume that they are connected by an inherent, or causal, link. There are many events whose association produces a large correlation coefficient but where it would be absurd to conclude that the one caused the other. Here are some examples:

- It is discovered that pastors' salaries and alcohol prices have correlated for many years. Does this mean that the more pastors make, the more they spend on alcohol?
- Researchers in Sweden examined the human birth rate and the stork population over a long period and determined that the two strongly and positively correlate. Does this mean that newborns are delivered by storks?
- The odds of surviving one's first heart attack are many times higher in smokers than in non-smokers. Is smoking good for your health?
- In postwar Germany there was a strong correlation between orange imports and deaths. Are oranges bad for your health?



	A	B	C	D	E	F	G
2	y_i	x_i	$R(y_i)$	$R(x_i)$			
3	1	1	2.5	3.0			
4	1	1	2.5	3.0			
5	1	1	2.5	3.0			
6	1	1	2.5	3.0			
7	2	1	6.0	3.0			
8	2	2	6.0	9.0			
9	2	2	6.0	9.0			
10	3	2	11.5	9.0			
11	3	3	11.5	14.0			
12	3	4	11.5	18.0			
13	3	2	11.5	9.0			
14	3	3	11.5	14.0			
15	3	2	11.5	9.0			
16	3	2	11.5	9.0			
17	3	3	11.5	14.0			
18	4	2	20.0	9.0			
19	4	4	20.0	18.0			
20	4	4	20.0	18.0			
21	4	4	20.0	18.0			
22	4	4	20.0	18.0			
23	4	5	20.0	23.0			
24	4	5	20.0	23.0			
25	4	5	20.0	23.0			
26	4	5	20.0	23.0			
27	5	5	25.0	23.0			
28	Total		325.0	325.0		Spearman	0.880
29	Mean		13.0	13.0			

Fig. 4.35 Spearman's correlation with Excel

- The likelihood of dying in bed is larger than the likelihood of being killed in a car or plane crash. Are beds dangerous?
- Researchers find a positive correlation between body size and alcohol consumption. Are all tall people drinkers?

Demagogues and propagandists love to exploit fallacies such as these, supporting their arguments with the statement "statistics show". Those trained in statistics know better: correlation does not always imply causation. A correlation without causation is called a *spurious correlation.*

Why do spurious correlations occur? Sometimes correlation occurs by accident. These accidental correlations are often referred to as *nonsense correlations.*

But spurious correlations do not always result by accident. Frequently, two variables correlate because of a third variable that influences each (see Fig. 4.36). In this case one speaks of a *common cause.* The correlation between the stork

population and the number of newborns is one example. Data collection on human birth rates and stork populations in Sweden began in the early twentieth century. Over the next 100 years rural society became increasingly industrialized and cities grew. This development displaced the stork population to more rural areas. At the same time, families living in the newly urbanized areas had fewer children, while those in rural areas continued to have many children. The result: cities saw fewer births and fewer storks, and the countryside saw more births and more storks. Hence, industrialization served as the common cause for the correlation between storks and newborns. A common cause is also behind the correlation of alcohol prices and pastor salaries: inflation over the years caused both wages and prices to rise.

Another reason for spurious correlation is the influence of *mediator variables.* This happens when a variable *A* correlates with a variable *B* and variable *A* influences variable *B* via a mediator variable. Consider the correlation between height and alcohol consumption. As it turns out, it depends on the gender of the users. Men show a higher level of alcohol consumption, and men are on average taller than women. Height, therefore, is the mediator variable through which the variable *gender* influences the variable *alcohol consumption.*

Likewise, the association between time in bed and mortality rate arises only because people who spend more time in bed are more likely to have a serious

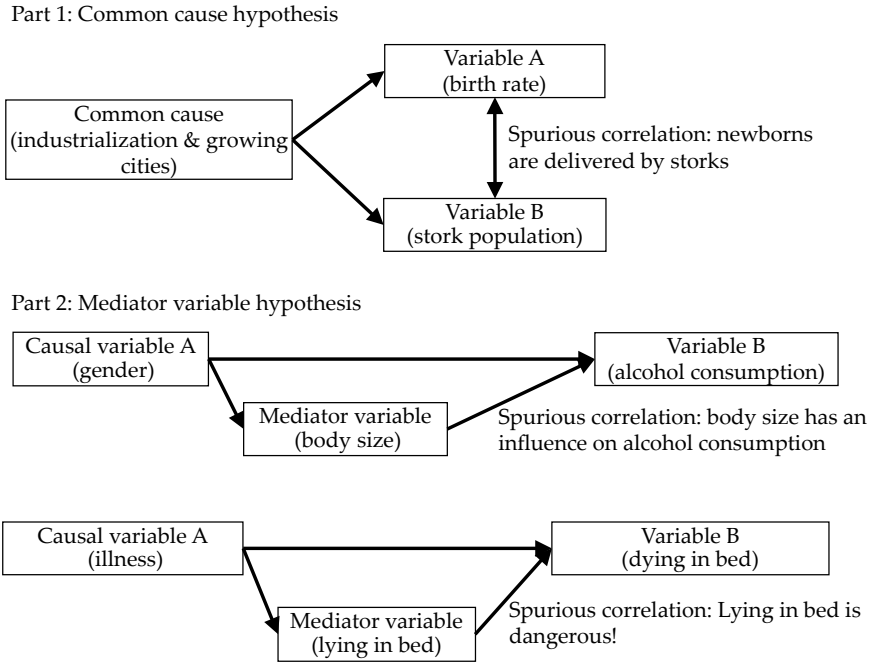


Fig. 4.36 Reasons for spurious correlations

illness, and people with a serious illness are more likely to die. In this way, serious illness influences mortality rate via the mediator variable *time in bed*.

Finally, smokers survive their first heart attack more frequently than non-smokers because smokers usually have their first heart attack at a much younger age. Here, the actual causal variable for the likelihood of survival is age.

4.7.1 Partial Correlation

If researchers suspect a spurious correlation while analysing data, they must adjust the results accordingly. For instance, when a common cause is involved, the correlation between variables A and B must be cleansed of the influence of the common cause variables. The true correlation between mediator variables and variable B is only expressed when one removes the effects of possible causal variables beforehand. We'll look at how to do this using an example from economics.

The owner of petrol station called SPARAL wants to know whether there is an association between the price of high-octane fuel and market share. So he correlates the price of high-octane petrol with the market share for 27 days. He determines a correlation coefficient of $r_{yz} = (-\ 0.723)$. This represents a strong negative correlation, and it makes sense economically: the higher the price, the less the market share, and vice versa. Next the SPARAL owner wants to know how the prices at the JETY station down the street influence his market share. So he examines the association between the price of JETY high-octane petrol and the SPARAL market share. He finds a correlation of $r_{xy} = (-\ 0.664)$. Unlike the last correlation, this one doesn't make economic sense: the higher the competitor's price for high-octane fuel, the lower the market share of his product SPARAL. What can the reason be for this unexpected direction of association?

Now, petrol prices are mainly shaped by crude oil prices (in addition to oligopolistic skimming by petrol stations on weekends and holidays). If the prices for crude oil sink, the market expects a price reduction, and petrol prices decline. In the reverse case, increased crude oil prices lead to higher prices at the pump.

In our example, the crude oil market serves as the common cause for the price association between JETY and SPARAL. This applies both for the correlations described above and for the strong correlation coefficient—$r_{xz} = 0.902$—between high-octane fuels at JETY and SPARAL. Both petrol stations increase (or sink) their prices almost simultaneously based on the crude oil market. The correlations are represented graphically in Fig. 4.37.

For the SPARAL petrol station owner, however, a crucial remains: What is the magnitude of the association between the competitor's high-octane fuel prices and his own market share? To answer this question, we must first remove—or control for—the effect caused by SPARAL's high-octane fuel price, i.e. the SPARAL price along with related developments on the crude oil market. This allows us to isolate the effect of the competitor's price on SPARAL's market share. How great is the correlation between the variables x (JETY price) and the variable y (SPARAL market share) if the variable z (SPARAL price) is eliminated?

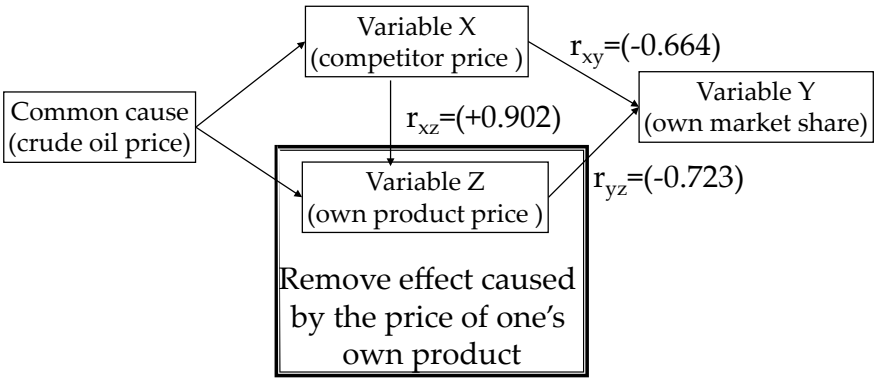


Fig. 4.37 High-octane fuel and market share: an example of spurious correlation

One speaks in such cases of a partial correlation between the variables x and y, with the effect of a variable z removed. It can be calculated as follows:

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{\left(1 - r_{xz}^2\right) \cdot \left(1 - r_{yz}^2\right)}} = \frac{-0.664 - (0.902 \cdot (-0.723))}{\sqrt{\left((1 - 0.902^2) \cdot \left(1 - (-0.723)^2\right)\right)}} = -0.04 \tag{4.49}$$

This equation produces a partial correlation effect of $r_{xy.z} = -0.04$, which indicates no association between the price for JETY high-octane fuel and the market share of SPARAL. Hence, the attendant has no need to worry about the effect of JETY's prices on his market share—the effect is close to zero.

4.7.2 Partial Correlations with R and R Commander

To calculate a partial correlation using the R Commander, select *Statistics → Summaries → Correlation matrix…* This opens the *Correlation Matrix* dialogue box. Select all variables to be correlated (gross price of the own product, market share of the own product, and net price of the competitor) under *Variables (pick two or more)* and click *Partial* under *Type of correlations*. This produces the partial correlation coefficients in Fig. 4.38.

4.7.3 Partial Correlations with SPSS

To calculate a partial correlation with SPSS, select *Analyze → Correlate → Partial*. This opens the *Partial Correlations* dialogue box. Enter the variable to be checked (SPARAL price for high-octane and SPARAL market share) under *Variables*. This produces the partial correlation coefficient in Fig. 4.39.

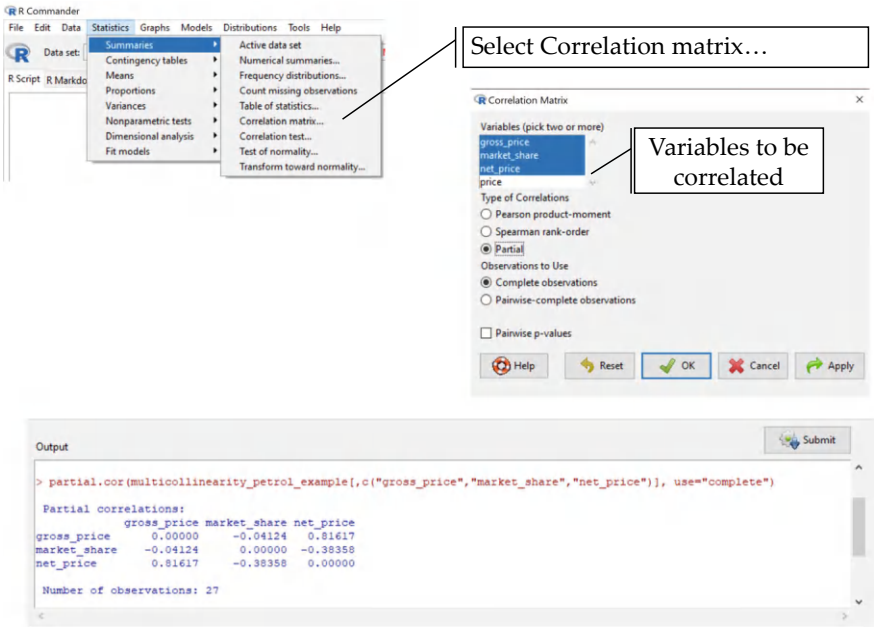
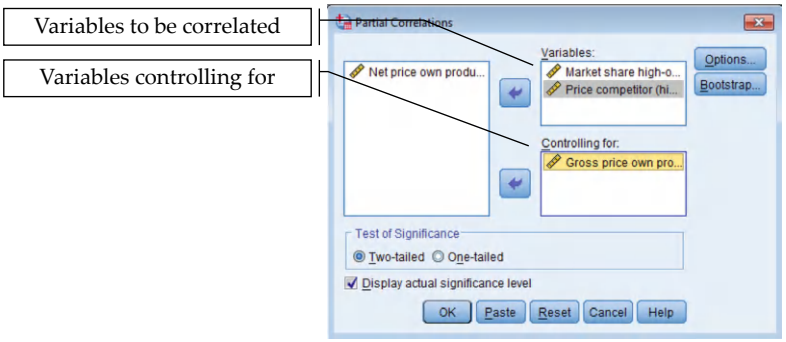


Fig. 4.38 Partial correlation with R and R Commander (high-octane petrol)



Effect of the competitor's price on SPARAL's market share controlled for SPARAL's own price:

Correlations

Control variables		Market share, high-octane petrol	Competitor price (JETY high-octane petrol)
Gross price of own product (SPARAL high-octane petrol)	Market share, high-octane petrol	1.000	-.041
	Competitor price (JETY high-octane petrol)	-.041	1.000

Fig. 4.39 Partial correlation with SPSS (high-octane petrol)

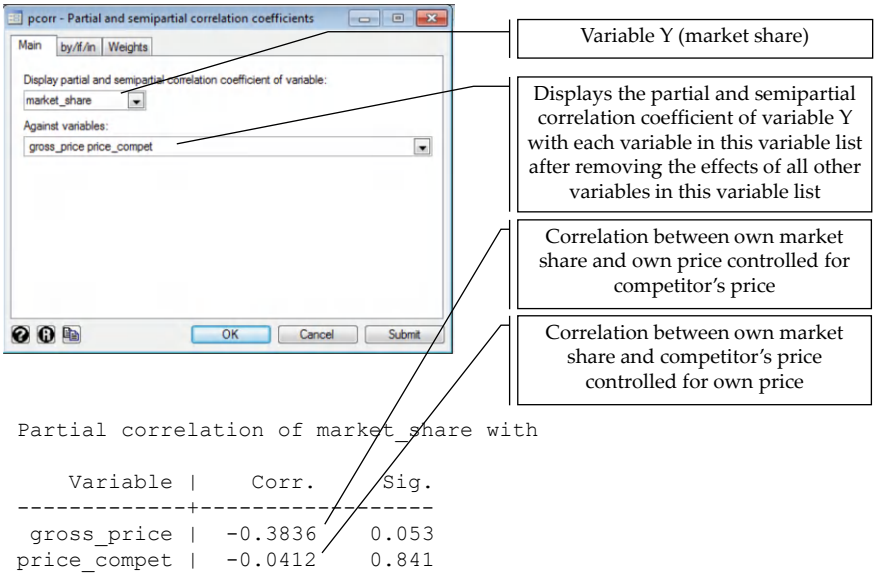


Fig. 4.40 Partial correlation with Stata (high-octane petrol)

4.7.4 Partial Correlations with Stata

The analysis can be performed with Stata in a similar manner. Select *Statistics → Summaries, tables, and tests → Summary and descriptive statistics → Partial correlations* to open the *Partial correlations coefficient* dialogue box.

In the first input line (*Display partial correlation coefficient of variable:*) enter the y variable. In the second input line (*Against variables:*) enter the x and z variables (and others if needed). Click *OK* or *Submit* to execute the Stata command.[7] When checked for the JETY price, the correlation coefficient for the association between the price of SPARAL and the market share of SPARAL is $r_{yz.x} = (-0.3836)$. With the effect of SPARAL's price removed, the correlation coefficient for the association between JETY and the market share of SPARAL is $r_{xy.z} = (-0.0412)$ (see Fig. 4.40).

4.7.5 Partial Correlation with Excel

Excel has no pre-programmed functions for calculating partial correlations. To perform them, you have to programme them yourself. First calculate the correlations between all variables (r_{xy}, r_{xz}, r_{yz}) with the CORREL command. Then use the

[7] Syntax: pcorr market_share gross_price price_compet.

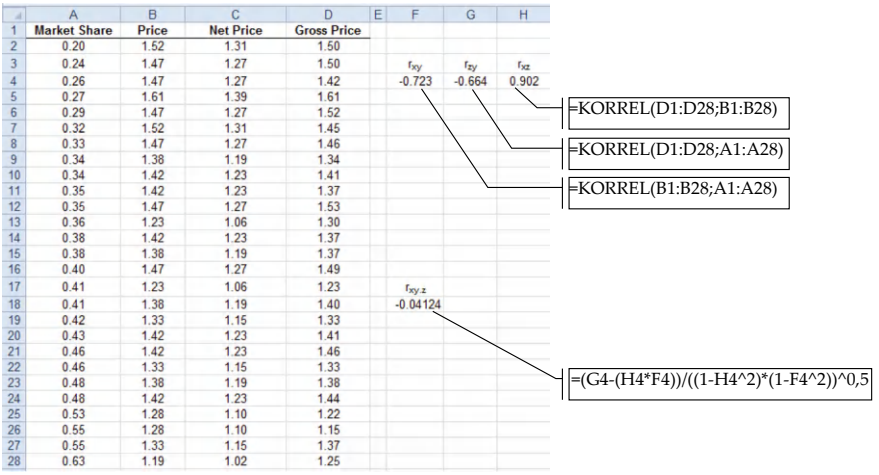


Fig. 4.41 Partial correlation with Excel (high-octane petrol)

formula

$$r_{xy.z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{\left(1 - r_{xz}^2\right) \cdot \left(1 - r_{yz}^2\right)}} \tag{4.50}$$

to programme the partial correlation coefficient. Figure 4.41 provides some examples.

4.8 Chapter Exercises

Exercise 1
One-hundred customers were randomly selected for an experiment measuring the effect of music on the amount of money people spend in a supermarket. One half of the customers shopped on days when no background music was played. The other half shopped on days accompanied by music and advertisements. Each customer was assigned to one of three groups—high, moderate, or low—based on how much he or she spent.

(a) Your hard drive crashes and you lose all your data. Fortunately, you manage to reconstruct the survey results for 100 observations from your notes. The relative frequency is $f(x = 2 \mid y = 3) = 0.5$ and the absolute frequency is $h(y = 1) = 35$. Based on this information, fill in the missing cells below.

	High amount spent ($y = 1$)	Moderate amount spent ($y = 2$)	Low amount spent ($y = 3$)	Sum (X)
With music ($x = 1$)	30			
W/o music ($x = 2$)		20		
Sum (Y)			40	

(b) After reconstructing the data, you decide to increase your sample size by surveying 300 additional customers. This leaves you with the following contingency table. Fill in the marginal frequencies and the expected counts under statistical independence. In the parentheses provide the expected counts given the actual number of observations.

		High ($y = 1$)	Moderate ($y = 2$)	Low ($y = 3$)	Sum (X)
With music ($x = 1$)	Count (expected count)	130 (____)	30 (____)	50 (____)	
Without music ($x = 2$)	Count (expected count)	40 (____)	20 (____)	130 (____)	
Sum (Y)	Count				

(c) Determine the chi-square value.
(d) Calculate Cramer's V.

Exercise 2
You are given the task of sampling the household size of customers at a grocery store and the number of bananas they buy.

(a) You collect 150 observations. The relative frequency is $f(x = 4 \mid y = 2) = 1/18$ and the absolute frequency is $h(x = 2 \mid y = 3) = 30$. Based on this information, fill in the missing cells below.

	1 person ($y = 1$)	2 persons ($y = 2$)	≥ 3 persons ($y = 3$)	Sum (x)
0 bananas ($x = 1$)	20	30		60
1 bananas ($x = 2$)		20		55
2 bananas ($x = 3$)			20	27
≥ 3 bananas ($x = 4$)				
Sum (y)	33	54		

(b) The data you collect yields the following contingency table. Fill in the marginal frequencies and the expected counts under statistical independence. In the parentheses provide the expected counts given the actual number of observations.

	1 person ($y = 1$)	2 persons ($y = 2$)	≥ 3 persons ($y = 3$)	Sum (x)
0 bananas ($x = 1$)	40 (____)	0 (____)	40 (____)	
1 banana ($x = 2$)	103 (____)	15 (____)	87 (____)	
2 bananas ($x = 3$)	5 (____)	0 (____)	3 (____)	
≥ 3 bananas ($x = 4$)	2 (____)	0 (____)	5 (____)	
Sum (y)				

(c) Determine the chi-square value.
(d) Calculate Cramer's V.
(e) Why doesn't it make sense to calculate phi in this case?

Exercise 3
A company measures customer satisfaction in three regions, producing the following crosstab:

			Region			Total
			Region 1	Region 2	Region 3	
Customer satisfaction	Excellent	Count	13	0	2	15
		Expected count	6.1	5.5	3.5	15.0
		% within customer satisfaction	86.7	0.0	13.3	100.0
		% within region	61.9	0.0	16.7	28.8
		% of total	25.0	0.0	3.8	28.8
	Average	Count	0	10	10	20
		Expected count	8.1	7.3	4.6	20.0
		% within customer satisfaction	0.0	50.0	50.0	100.0
		% within region	0.0	52.6	83.3	38.5
		% of total	0.0	19.2	19.2	38.5
	Poor	Count	8	9	0	17
		Expected count	6.9	6.2	3.9	17.0
		% within customer satisfaction	47.1	52.9	0.0	100.0
		% within region	38.1	47.4	0.0	32.7
		% of total	15.4	17.3	0.0	32.7
Total		Count	21	19	12	52
		Expected count	21.0	19.0	12.0	52.0
		% within customer satisfaction	40.4	36.5	23.1	100.0

(continued)

(continued)

		Region			Total
		Region 1	Region 2	Region 3	
	% within region	100.0	100.0	100.0	100.0
	% of total	40.4	36.5	23.1	100.0

Chi-square tests			
	Value	df	Asymp. sig. (2-sided)
Pearson chi-square	34.767[a]	4	0.000
Likelihood ratio	48.519	4	0.000
Linear-by-linear association	0.569	1	0.451
N of valid cases	52		

[a] 3 cells (33.3%) have an expected count less than 5. The minimum expected count is 3.46

Symmetric measures			
		Value	Approx. sig.
Nominal by nominal	Phi	0.818	0.000
	Cramer's V	0.578	0.000
	Contingency coefficient	0.633	0.000
N of valid cases		52	

[a] Not assuming the null hypothesis
[b] Using the asymptotic standard error assuming the null hypothesis

(a) What percentage of respondents answering "good" come from region 3?
(b) Interpret the strength of the association and assess the suitability of the phi coefficient, Cramer's V, and the contingency coefficient for solving the problem. Discuss possible problems when using the permitted measures of association and indicate regions with above-average numbers of satisfied or dissatisfied respondents.

Exercise 4

(a) Based on the data in Exercise 5 in Chap. 3 you conjecture that price is the decisive variable determining sales. Use a scatterplot to verify this hypothesis.
(b) Determine the standard deviation for price and the covariance between price and quantity of sales.
(c) Determine the strength of the linear metric association between item price and quantity of sales per country.
(d) Determine Spearman's rank correlation coefficient.
(e) Use a scatterplot to interpret your results from (c) and (d).

Exercise 5
A PISA study assesses the performance of students in 14 German states. The variables *scientific literacy* (x) and *reading comprehension* (y) yield the following information:

- $\overline{x}^2 = 3.20$
- $\sum_{i=1}^{n} (x_i - \overline{x})^2 = 3042.36$
- $\sum_{i=1}^{n} y_i = -309$
- $\sum_{i=1}^{n} y_i^2 = 10{,}545$
- $\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = 2987.81$

(a) What is the (unweighted) mean value for reading comprehension?
(b) What is the empirical standard deviation for reading comprehension?
(c) What is the variation coefficient for reading comprehension?
(d) Determine the empirical variance for scientific literacy.
(e) Determine the covariance between the variables x and y.
(f) Determine the strength of the linear metric association between reading comprehension and scientific literacy.
(g) Determine the rank correlation coefficient under the assumption that the sum of the squared rank differences for statistical series is 54.

Exercise 6
You want to find out whether there is an association between the quantity of customer purchases (y) and customer income € (x). For 715 customers, you calculate a covariance between income and purchase quantity of $S_{XY} = 2.4$ for 715.

(a) What does covariance tell us about trait association?
(b) Calculate Pearson's correlation coefficient assuming that:

$$\sum_{i=1}^{n} (x_i - \overline{x})^2 = 22{,}500 \quad \text{and} \quad \sum_{i=1}^{n} (y_i - \overline{y})^2 = 17{,}000$$

(c) Explain the association between the traits based on the calculated correlation coefficient.

Exercise 7
The newspaper *Stupid Times* published a study on the connection between the number of books people have read (x) and the serious colds they have had. The study—which relied on a mere five observations—produced the following data:

Observation	1	2	3	4	5
$(x_i - \bar{x})(y_i - \bar{y})$	203.4	847.4	9329.4	4703.4	− 225.6

The standard deviation of books read is 432.9; the standard deviation of serious colds is 7.5.

(a) Calculate Pearson's correlation coefficient. What conclusion is *Stupid Times* likely to have drawn?
(b) Explain what a spurious correlation is theoretically.
(c) Based on your understanding of a spurious correlation, how do you interpret the result in (a)?

Exercise 8
For a particular brand of potato chips, a market research institute determines a high correlation coefficient—$r = (-0.7383)$—between sales and price. Accidentally, they also discover a weak association—$r = (+0.3347)$—between potato chips sales and toilet paper price.

(a) How should we interpret the correlation coefficient of $r = 0.3347$ for potato chip sales and toilet paper price?
(b) Calculate the partial correlation coefficient between potato chip sales and toilet paper price to the nearest thousandth and controlled for the potato chip price. The correlation between toilet paper price and potato chip price is $r = (-0.4624)$.
(c) How should we interpret the results?

Exercise 9
Researchers investigated the market share of a product called *Funny* in a variety of retail stores. A few stores ran advertisements during certain weeks. Researchers assemble the following data:

	Store promotion		Statistic
Market share for Funny	No	Mean	0.3688
		Std. deviation	0.0943
	Yes	Mean	0.4090
		Std. deviation	0.0963

The standard deviation of all observations for the variable *Market Share FUNNY* is 0.095. Is there an association between advertising (1 = advertising; 0 = no advertising) and (metric) market share achieved? Identify the appropriate measure of association.

4.9 Exercise Solutions

Solution 1

(a)

	High amount spent ($y = 1$)	Moderate amount spent ($y = 2$)	Low amount spent ($y = 3$)	Sum (X)
With music ($x = 1$)	30	5	20	55
W/o music ($x = 2$)	5	20	20	45
Sum (Y)	35	25	40	100

(b)

		High ($y = 1$)	Moderate ($y = 2$)	Low ($y = 3$)	Sum (X)
With music ($x = 1$)	Count (expected counts)	130 (89.25)	30 (26.25)	50 (94.50)	210
W/o music ($x = 2$)	Count (expected counts)	40 (80.75)	20 (23.75)	130 (85.50)	190
Sum (Y)	Count	170	50	180	400

(c) $\chi^2 = \frac{(130-89.25)^2}{89.25} + \frac{(30-26.25)^2}{26.25} + \cdots + \frac{(130-85.5)^2}{85.5} = 84.41$

(d) $V = \sqrt{\frac{\chi^2}{n \cdot (\text{Min}(\text{number of columns, number of rows}) - 1)}} = \sqrt{\frac{84.41}{400 \cdot 1}} = 0.46$

Solution 2

(a)

	1 person ($y = 1$)	2 persons ($y = 2$)	≥ 3 persons ($y = 3$)	Sum (x)
0 bananas ($x = 1$)	20	30	10	60
1 banana ($x = 2$)	5	20	30	55
2 bananas ($x = 3$)	6	1	20	27
≥ 3 bananas ($x = 4$)	2	3	3	8
Sum (y)	33	54	63	150

(b)

	1 person ($y = 1$)	2 persons ($y = 2$)	≥ 3 persons ($y = 3$)	Sum (x)
0 bananas ($x = 1$)	40 (40)	0 (4)	40 (36)	80

(continued)

(continued)

	1 person ($y = 1$)	2 persons ($y = 2$)	≥ 3 persons ($y = 3$)	Sum (x)
1 bananas ($x = 2$)	103 (102.5)	15 (10.25)	87 (92.25)	205
2 bananas ($x = 3$)	5 (4)	0 (0.4)	3 (3.6)	8
≥ 3 bananas ($x = 4$)	2 (3.5)	0 (0.35)	5 (3.15)	7
Sum (y)	150	15	135	300

(c) $\chi^2 = 9.77$. If the last three rows are added together due to their sparseness, we get: $\chi^2 = 0 + 4 + 0.44 + 0 + 1.45 + 0.16 = 6.06$.

(d) $V = \sqrt{\frac{\chi^2}{n \cdot (\text{Min(number of columns, number of rows)} - 1)}} = \sqrt{\frac{9.77}{300 \cdot 2}} = 0.1276$. If the last three rows are added together due to their sparseness, we get: $V = \sqrt{\frac{6.06}{300 \cdot 1}} = 0.142$

(e) Phi is only permitted with two rows or two columns.

Solution 3

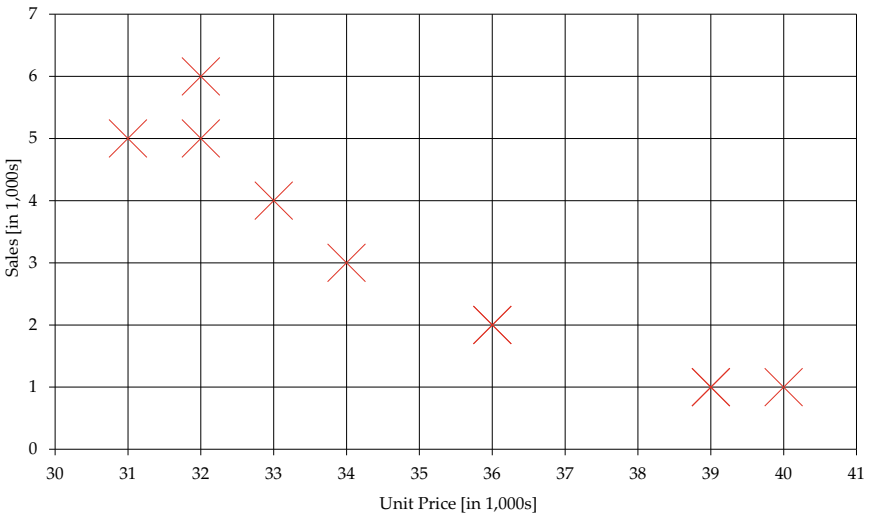
(a) $f(\text{Region} = \text{Region3} \mid \text{assessment} = \text{good}) = 2/15 \cdot 100\% = 13.3\%$.

(b)

- Phi is unsuited, as the contingency table has more than two rows/columns.
- The contingency coefficient is unsuited, as it only applies when the tables have many rows/columns.
- Cramer's V can be interpreted as follows: $V = 0.578$. This indicates a moderate association.
- The assessment *good* has a greater-than-average frequency in region 1 (expected count = 6.1; actual count = 13); a lower-than-average frequency in region 2 (expected count = 5.5; actual count = 0); a lower-than-average frequency in region 3 (expected count = 3.5; actual count = 2). The assessment *fair* has a greater-than-average frequency in region 2 (expected count = 7.3; actual count = 10); a greater-than-average frequency in region 3 (expected count = 4.6; actual count = 10). The assessment *poor* has a greater-than-average frequency in region 1 (expected count = 6.9; actual count = 8).
- Another aspect to note is that many cells are unoccupied. One can thus ask whether a table smaller than 3×3 should be used (i.e. 2×2; 2×3; 3×2).

Solution 4

(a) Y: Sales; X: Price [in 1000s]



(b)

Country	Sales [in 1000s]	Unit price [in 1000s]	Sales^2 [in 1000s]	Unit price^2 [in 1000s]	Sales·Price	R (Sales)	R (Price)	d_i	d_i^2
1	6	32	36	1024.00	192.00	10	2.5	7.5	56.25
2	4	33	16	1089.00	132.00	7	4	3	9
3	3	34	9	1156.00	102.00	6	5	1	1
4	5	32	25	1024.00	160.00	8.5	2.5	6	36
5	2	36	4	1296.00	72.00	4.5	6.5	− 2	4
6	2	36	4	1296.00	72.00	4.5	6.5	− 2	4
7	5	31	25	961.00	155.00	8.5	1	7.5	56.25
8	1	39	1	1521.00	39.00	2	8.5	− 6.5	42.25
9	1	40	1	1600.00	40.00	2	10	− 8	64
10	1	39	1	1521.00	39.00	2	8.5	− 6.5	42.25
Sum	30	352	122	12,488.00	1003.00	55	55	0	315
Mean	3.0	35.2	12.2	1248.80	100.30	5.5	5.5	0.0	31.5

Unit price [in 1000s of MUs]:

$$\overline{x} = \frac{1}{10}(32 + 33 + 34 + \cdots + 39) = 35.2$$

$$S_{\text{emp}} = \sqrt{\frac{(x_i - \overline{x})^2}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2} = \sqrt{\frac{1}{10}12{,}488 - 35.2^2} = \sqrt{9.76} = 3.12$$

Sales:

$$\overline{y} = \frac{1}{10}(6 + 4 + 3 + \cdots + 1) = 3.0$$

$$S_{\text{emp}} = \sqrt{\frac{(y_i - \overline{y})^2}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2} = \sqrt{\frac{1}{10}122 - 3^2} = \sqrt{3.2} = 1.79$$

Covariance:

$$\begin{aligned} S_{xy} &= \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i - \overline{x} \cdot \overline{y} = \frac{1}{10}(6 \cdot 32 + \cdots + 1 \cdot 39) - 35.2 \cdot 3 \\ &= 100.3 - 105.6 = (-5.3) \end{aligned}$$

(c) $r = \frac{S_{xy}}{S_x S_y} = \frac{-5.3}{1.79 \cdot 3.12} = (-0.95)$.

(d) $\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2-1)} = 1 - \frac{6 \cdot (7.5^2 + 3^2 + \cdots + (-6.5^2))}{10 \cdot (10^2-1)} = 1 - \frac{6 \cdot 315}{10 \cdot (10^2-1)} = (-0.909)$.
When this coefficient is calculated with the full formula, we get: $\rho = (-0.962)$.
The reason is because of the large number of rank ties.

(e) Negative monotonic association.

Solution 5

(a) $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{-309}{14} = (-22.07)$

(b) $S_{\text{emp}} = \sqrt{\frac{\sum_{i=1}^{n} y_i^2}{n} - \overline{y}^2} = \sqrt{\frac{10{,}545}{14} - 22.07^2} = \sqrt{266.129} = 16.31$

(c) Coefficient of Variation $\frac{S_{\text{emp}}}{|\overline{y}|} = \frac{16.31}{|-22.07|} = 0.74$

(d) $S_{\text{emp}}^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n} = \frac{3042.36}{14} = 217.31$

(e) $S_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n} = 213.42$

(f) $r = \frac{S_{xy}}{S_x \cdot S_y} = 0.89$

(g) $\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2-1)} = 1 - \frac{6 \cdot 54}{14 \cdot (14^2-1)} = 0.88$.

Solution 6

(a) The covariance only gives the direction of a possible association.
(b) $r = \frac{2.4}{\sqrt{\frac{22,500}{715} \cdot \frac{17,000}{715}}} = \frac{2.4}{5.61 \cdot 4.88} = 0.0877$
(c) No linear association.

Solution 7

(a) Using the table we can calculate the following: $\frac{1}{5}\sum_{i=1}^{5}(x_i - \overline{x})(y_i - \overline{y}) = 2971.6$. Pearson's correlation is then: $r = \frac{2971.6}{432.96 \cdot 7.49} = 0.916$. The *Stupid Times* will conclude that reading books is unhealthy, because the linear association is large between colds and books read.
(b) With a spurious correlation, a third (hidden) variable has an effect on the variables under investigation. It ultimately explains the relationship associated by the high coefficient of correlation.
(c) A spurious correlation exists. The background (common cause) variable is age. As age increases, people on average read more books and have more colds. If we limit ourselves to one age class, there is probably no correlation between colds had and books read.

Solution 8

(a) The higher the price for toilet paper, the higher the sales for potato chips.
(b) The formula for the partial coefficient of correlation is: $r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{\left(1 - r_{xz}^2\right) \cdot \left(1 - r_{yz}^2\right)}}$.

In the example the variable x equals potato chip sales, variable y potato chip price, and variable z toilet paper price. Other variable assignments are also possible without changing the final result. We are looking for $r_{xz.y}$. The formula for the partial correlation coefficient should then be modified as follows:

$$r_{xz.y} = \frac{r_{xz} - r_{xy}r_{zy}}{\sqrt{\left(1 - r_{xy}^2\right) \cdot \left(1 - r_{zy}^2\right)}}$$
$$= \frac{0.3347 - ((-0.7383) \cdot (-0.4624))}{\sqrt{\left(1 - (-0.7383)^2\right) \cdot \left(1 - (-0.4624)^2\right)}} = (-0.011)$$

(c) The association in (a) is a spurious correlation. In reality there is no association between toilet paper price and potato chip sales.

Solution 9

$$r_{\text{pb}} = \frac{\overline{y}_1 - \overline{y}_0}{S_y}\sqrt{\frac{n_0 \cdot n_1}{n^2}} = \frac{0.41 - 0.37}{0.095}\sqrt{\frac{2427 \cdot 21,753}{24,180^2}} = 0.127$$

References

Bortz, J., Lienert, G. A., & Boehnke, K. (2000). *Verteilungsfreie Methoden der Biostatistik* (2nd ed.). Springer.

Bowers, J. (1972). A note on comparing r-biserial and r-point biserial. *Educational and Psychological Measurement, 32*, 771–775.

British Board of Trade. (1990). *Report on the loss of the 'Titanic' (S.S.). British Board of Trade Inquiry Report (reprint).* Allan Sutton Publishing.

Glass, G. V. (1966). Note on rank-biserial correlation. *Educational and Psychological Measurement, 26*, 623–631.

Janson, S., & Vegelius, J. (1982). Correlation coefficients for more than one scale type. *Multivariate Behaviorial Research, 17*, 271–284.

Classical Measurement Theory

5

The term *descriptive statistics* refers to all techniques used to obtain information based on the description of data from a population. The calculation of figures and parameters as well as the generation of graphics and tables are just some of the methods and techniques used in descriptive statistics. *Inferential statistics*, sometimes referred to as *inductive statistics*, did not develop until much later. It uses samples to make conclusions, or inferences, about a population. The term *descriptive statistics* refers to all techniques used to obtain information based on the description of data from a population. The calculation of figures and parameters as well as the generation of graphics and tables are just some of the methods and techniques used in descriptive statistics. *Inferential statistics*, sometimes referred to as *inductive statistics*, did not develop until much later. It uses samples to make conclusions, or inferences, about a population. Many of the methods of inferential statistics go back to discoveries made by such thinkers as Jacob Bernoulli (1654–1705), Abraham de Moivre (1667–1754), Thomas Bayes (1702–1761), Pierre-Simon Laplace (1749–1827), Carl Friedrich Gauß (1777–1855), Pafnuty Lvovich Chebyshev (1821–1894), Francis Galton (1822–1911), Ronald A. Fisher (1890–1962), and William Sealy Gosset (1876–1937). Thanks to their work, we no longer have to count and measure each individual within a population, but can instead conduct a smaller, more manageable survey. This comes in handy when a full survey would be too expensive or take too long, or when collecting the data damages the elements under investigation (as with various kinds of material testing, such as wine tasting). But inferential statistics has a price: because the data are collected from a sample, not from a total population, our conclusions about the data

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*,
Springer Texts in Business and Economics,
https://doi.org/10.1007/978-3-031-78070-7_5

carry a certain degree of uncertainty. But inferential statistics can also define the "price" of this uncertainty using margins of error. Classical measurement theory gives us the tools to calculate statistical error.

Classical measurement theory makes two basic assumptions. The first is that the characteristics of the data being studied are stable, allowing for reliable measurements. The second is that the measurements are valid, which is to say, that recorded values correspond with actual values. This does not mean that classical measurement theory rejects the possibility of error altogether; on the contrary. It takes for granted that even the most careful measurements will contain some mistakes. For example, an astronomer who measures the distance between the same two planets on five different evenings is likely to arrive at five (slightly) different values. This can happen for a number of reasons. Light reflecting from nearby may distort the results. Minor nonsystematic errors are another possibility. Anyone who has ever measured the floor area of his or her apartment on more than one occasion and ended up with different totals has experienced such error. The problem is not that measuring tape standards have changed. It is that the tape was read incorrectly, or the width of the baseboard is irregular, or the walls are crooked. Statisticians call minor nonsystematic errors random errors or statistical errors. They occur unexpectedly, but follow certain laws that enable the statistical calculation of the size of the nonsystematic error. If you carry out measurements often enough, the values form a normal distribution, with certain values (the actual values) occurring frequently and divergent values (the nonsystematic errors) occurring less and less frequently the more they deviate from the mean. We will talk about the normal distribution of nonsystematic errors below.

Hence, we can be very certain that the most frequent values are the actual values—unless, of course, we've made a systematic error. Systematic errors are a real problem: inferential statistics cannot calculate them and the extent to which they distort the results is nearly impossible to estimate. What is an example of a systematic error? Say we want to measure the size of an apartment, but instead of measuring from zero, we mistakenly measure from 1, so that every length we measure is one cm longer than it really is. In this example we could have easily seen our error had we carefully checked the measuring tape. But in economic and social research, data from surveys are often distorted by *unnoticed* systematic errors. The reason is that collecting and recording data by means of interviews and questionnaires is a social process shaped by a multitude of exogenous factors. To name just one problem: respondents do not behave passively when being questioned—instead, they develop action and reaction strategies based on their own agendas, interpretations, and the specific circumstances at play. This kind of variability makes exact measurement impossible, as the values vary depending on the subject and the situation.

So while systematic and nonsystematic errors are both undesirable, the consequences of the former are much more serious, which is why special efforts must be made to eliminate their occurrence. One goal of empirical research, therefore, must be to prevent situation-specific systematic distortion. We can start by considering the most common forms of such distortion:

- sampling errors—caused by samples that are not representative of the population (see Sect. 5.1);
- nonsampling errors—caused by response or nonresponse errors (see Sect. 5.2).

5.1 Sources of Sampling Errors

The unease some people have towards statistics often arises from its reliance on samples, not entire populations; and because samples are incomplete, people often assume they cannot be valid. Critics like to cite election forecasts as examples of how political analysts "always get it wrong", and this shouldn't come as a surprise: the field of opinion research became known mostly on account of a famously wrong prediction. In the autumn of 1936, the weekly magazine *The Literary Digest* projected that the U.S. presidential incumbent, Franklin Delano Roosevelt, would be defeated by his Republican challenger, Alf Landon, 57–43%. The forecast was considered highly credible. Not only had *The Literary Digest* correctly predicted every presidential election since 1916; the opinion poll it used for the 1936 forecast was the largest ever conducted, with ten million people surveyed and 2.3 million respondents—a huge sample even by today's standards. But a single pollster dared to challenge the wisdom of *The Literary Digest*: George Gallup. Using a much smaller sample size—50,000 people—he projected a Roosevelt win. As we know today, his prediction was correct. Roosevelt beat Landon with 62% of the vote.

How could this happen? How could Gallup's smaller sample make a better prediction than the millions of questionnaires collected by *The Literary Digest*? Many people intuitively believe that large samples are more accurate than small samples. Since larger samples take into account a larger share of total observations, they must better approximate actual values, it is assumed. But in reality size alone is not decisive for the representativeness of a sample; the key factor is its randomness. And precisely here lay the central mistake of *The Literary Digest*. In gathering data, its analysts relied primarily on registered car owners and telephone users, groups whose members tended not only to be wealthy (in the 1930s only the rich could afford cars and telephones) but also to vote Republican (true for the rich then and now). That is to say, *The Literary Digest* selected a sampling frame—the group to be surveyed—that was not representative of the U.S. electorate. Moreover, whenever political opinion polls are voluntary, the people who take part are often motivated by a desire to express dissatisfaction with the incumbent. This further distorted *The Literary Digest* sample in favour of Republican voters.

Whenever a sample is unrepresentative of a population, it is considered biased. The size of a sample alone is never proof of its representativeness and never suffices to compensate for a sampling error.[1] A sample is only representative when

[1] Still, we can hear and read incorrect statements like this all the time: "The sample size of 250 is large. It *must* be representative".

it is isomorphic to the population being measured. A representative sample corresponds to the distribution of all features of the population relevant to the study. It is a faithful copy of the total population (Berekoven et al., 2009).

Therefore, there are several standard sampling methods used in practice, either alone or in combination (see Fig. 5.1). Some do not ensure isomorphism from the outset but can be made representative retroactively (e.g. by weighting certain groups in the sample). Generally, sampling methods can be grouped into nonprobability and probability approaches. In the former, the decision whether to include an element in the sample is left to the researcher, and the sampling probability of individual observations cannot be shown in advance. This is precisely the reason why certain methods of inferential statistics produce nongeneralizable results. Nevertheless, nonprobability sampling makes sense when an interesting characteristic is to be found only in a few individuals, or individuals are to be included in the sample that are subjectively considered typical or extreme for the population. For example, the effect of a medication on stomach aches in zero gravity can be tested only on a small group, i.e. astronauts. Another example is lead user analysis, which focuses on those customers who acquire products at an early point in the product life cycle.

These nonprobability techniques, also known as convenience sampling or judgemental sampling, make sense for certain research questions but are not suited for general sampling. The first step for creating a miniature cross-section of the total population from the sample is known as quota sampling, which seeks to represent a certain characteristic (say, gender) in proportion to the population. After determining the proportion of the characteristic in the population—for example, 52% women and 48% men—researchers then choose subjects at their discretion while making sure that the characteristic is represented proportionally. But as we saw above, this sort of nonrandom sampling is subjective in nature, and hence should only be used when truly necessary.

Random sampling and probability sampling give more reliable results. With simple random sampling every element of the total population has the same probability of appearing in the sample. This can be achieved by numbering and randomising the elements in a population before selection. Stratified sampling shares some similarities with quota sampling. It begins by assigning elements of a population to different strata based on certain characteristics. Then, once the elements are stratified, it applies random sampling within each group. This is where it parts ways with quota sampling. If the sampling fraction in each strata is proportionate to those of the population, the samples are proportionally allocated; if not, they are disproportionally allocated.

With cluster sampling, selection is based on groups of elements (e.g. regions). The serial application of multiple methods is known as sequential sampling.[2] In the case of probability sampling, elements appear in the sample randomly. This allows the probability of sampling a specific population element to be determined

[2] For more information on sampling methods, see Malhotra et al. (2017, pp. 410ff.).

Nonprobability Sampling	
Researchers decide which members to include in a sample. Probabilities for the elements cannot be shown. Problems: subjective selection; restricted result generalizability; may not be used for inferential statistics.	
Quota sampling	Quota sampling represents the proportions of a certain, important characteristic (e.g. gender) within the population. After these quotas are determined, samples are taken by convenience sampling (see below). Quotas are complicated to construct when populations are heterogeneous.
Convenience sampling	Can be used when a certain characteristic is limited to several elements in the population; the selection of elements suffices to explain most of the variance in the population. Not suited as a general sampling method.
Probability Sampling	
Elements appear in the sample randomly. Probabilities of selecting specific elements can be shown. May be used for inferential statistics.	
Simple random sampling	Every element has the same probability of appearing in the sample. Elements in the population are numbered and randomised before selection.
Stratified sampling	Before sampling, elements of a population are divided into separate subgroups (strata) by specific characteristic. If the sampling fraction in each of the strata is proportional to that of the total population, the samples are proportionately allocated. If not, they are disproportionately allocated.
Cluster sampling	The population is broken down into element groups, or clusters. From this, clusters are selected randomly and their individual elements studied. This can lead to cluster effects, as differences between the elements in a single cluster may be less in a normal random sample.
Sequential sampling	Random selection takes places via serial application of more than one probability sampling method.

Fig. 5.1 Empirical sampling methods

in advance, which in turn enables the application of inferential statistics and the generalizability of sample results to the total population with a specified margin of error.

5.2 Sources of Nonsampling Errors

The road to an unbiased sample is rocky; navigating it successfully requires sound experience, as errors can arise even after careful sampling. Nonsampling errors can be random or systematic. Random nonsampling errors are evenly distributed around the actual value. Because they all share a standard error, the actual value

can be correctly estimated as the average of all values. A systematic nonsampling error biases the results in a certain direction, and has an adverse effect on the results.

Figure 5.2 spotlights possible sources of nonsampling errors. These errors can be minimized by intelligent research design and by careful monitoring for the following problems:

Nonresponse errors: If a survey is unable to reach members of a certain population group because they are not home, or because they have refused to participate, then the group will be underrepresented in the sample.

Response errors: These occur when respondents give false answers, or if correct answers are misunderstood or misanalysed by the questioner. Generally, response errors have three kinds of sources:

1. The first is when respondents are unable to provide a valid answer due to uncertainty, fatigue, boredom, false memories, vague questions, or other circumstances. This is called an inability error. As these examples show, cognitive deficits need not be the cause. It also depends on the difficulty of the question. Who can remember the brand name of a yoghurt consumed four weeks

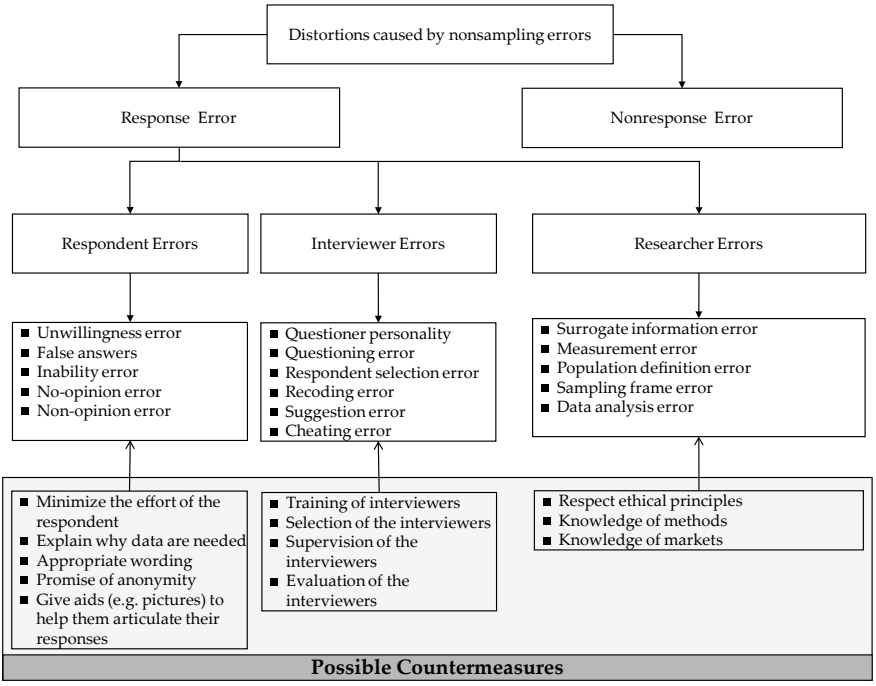


Fig. 5.2 Distortions caused by nonsampling errors. *Source* Based on Malhotra et al. (2017, p. 410). Figure compiled by the author

ago? Faulkenberry and Mason (1978, p. 533) distinguish between two types of answer omissions: *no opinion* (respondents are indecisive about an answer—for example, because of an ambiguous question), and *non-opinion* (respondents have no opinion about a topic). Sometimes, respondents provide false answers to please the questioner or to cover up their own ignorance. Respondents also tend to give answers they perceive as more socially desirable, regardless of whether they are true. This effect is even greater when third parties—such as members of the public—are present during the interview. Unwillingness errors are another possible source of error. This type of error occurs when missing answers are distributed systematically across respondent groups. For instance, the average income in a given population will be erroneously high if low-income groups tend to skip questions asking how much they earn.

2. The questioner can also be a source of sample error. Questioning errors, recording errors (be they erroneous, careless, or selective), or questioner personality (appearance/charisma/conduct) can falsify data. A questioner with a moustache is less likely to hear the statement "I don't like men with moustaches" than a clean-shaven one is. Attempts by questioners to influence respondents or to suggest certain answers can also distort the results. Often, there is a thin line between assistance and suggestion. When respondents must be found with rare characteristics for the sake of quota sampling, questioners are often tempted to fill out questionnaires themselves (leading to a cheating error or respondent selection error).
3. Finally, researchers themselves can influence respondents' answers. How a question is asked—e.g. how it is formulated, the scales offered for responding—can be even more important than what is asked about (measurement error). Moreover, researchers might have defined the population incorrectly (population definition error), adopted an unsuitable sampling method (sampling frame error: e.g. *The Literary Digest's* reliance on telephone users and car owners), or made errors when selecting proxies (surrogate information error). The last type of error occurs when an actual action cannot be observed or can only be observed indirectly (such as a future purchase decision) and for this reason proxies (such as product preferences) must be used instead.

Given the many sources of error in empirical surveys, researchers must remain on the look-out for problems and factor them into data interpretation. One problem we have yet to deal with is sample size. How large should a sample be? Generally, the reliability of a random sample increases with its size. Yet we also have to keep feasibility in mind. Imagine you've just made soup and want to see if it has enough salt. You need to take a sample that will allow you to infer about the soup as a whole. Say you can choose between a toothpick, a spoon, or a ladle for gathering the sample. Which is best? The toothpick provides a too small sample; it won't give you enough of an impression. The ladle provides too much of a sample; it commits you to eating a whole bowl of soup. But the spoon gives us the exact

size we need. Later in this book I will talk more about ideal sample size as it pertains to statistical populations. But let us first turn to the basics of calculating probability.

References

Berekoven, L., Eckert, W., & Ellenrieder, P. (2009). *Marktforschung. Methodische Grundlagen und praktische Anwendungen* (12th ed.). Gabler.

Faulkenberry, G. D., & Mason, R. (1978). Characteristics of nonopinion and no opinion response groups. *Public Opinion Quarterly, 42*, 533–543.

Malhotra, N. K., Nunan, D., & Birks, D. F. (2017). *Marketing research. An applied approach* (5th ed.). Pearson.



Calculating Probability

6

Let me summarize what I have said about samples so far. Attempts to acquire and collect data about all elements in a given population should be avoided when doing it

- … is too expensive;
- … takes too much time;
- … damages the elements under investigation (material testing, wine tasting, etc.); or
- … is impossible to organize.

And surely for many other reasons as well. A sample allows us to make inferences about the characteristics of a given population with a degree of probability. In everyday language, "probably" is used whenever we are uncertain about a future outcome. "Tomorrow it will probably rain", "I'll probably pass the statistics exam", "We'll probably see each other tomorrow"—all of these statements are about future events that may or may not occur. Indeed, we don't even know the likelihood of a given event occurring. Statements such as the following are more specific: "The probability of holding a ticket with six out of 49 matching numbers in the lottery is very low". "The probability that a newborn will weight more than five kilogrammes is very low". In these cases, we at least know the approximate probability (*very low* and *low*, respectively). The most concrete form of indicating the probability of an event is when we use numbers: "The probability of holding a lottery ticket with six out of 49 matching numbers is one to 13,983,916". "The probability that a newborn will weigh more than five kilogrammes is 0.03%" (Schwarze, 2013, p. 12). Things always seem to get more concrete when probability is expressed

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*,
Springer Texts in Business and Economics,
https://doi.org/10.1007/978-3-031-78070-7_6

numerically, such as 100% for certain events and 0% for impossible events. How do we determine such numbers? Before we learn how to calculate probability, we must first introduce a few terms and basic concepts.

6.1 Key Terms for Calculating Probability

Drawing lottery balls or flipping a coin are experiments that can produce two or more results, known as outcomes. Which outcome ultimately occurs cannot be determined in advance. Tossing a die can result in odd numbers or even numbers. These, in turn, can be broken down into elementary events—1, 3, 5 or 2, 4, 6. Elementary events each contain a single outcome and are mutually exclusive. The set of all possible elementary events—represented by the set $\Omega = \{\omega_1, \omega_2, \omega_3, \ldots \omega_m\}$—makes up the sample space (see Fig. 6.1). The sample space when tossing a die consists of the elementary events $\Omega = \{1, 2, 3, 4, 5, 6\}$. Single events can also be combined to form a union of events. For instance, in the die-toss experiment, the numbers three and under can be formed from combinations of the elementary events $\{1, 2, 3\}$. Logically, the combination of k outcomes can be represented as a logical disjunction, as all outcomes belong in either one or the other specified outcome (see Fig. 6.1).

$$\left(A = A_1 \cup A_2 \cup \ldots \cup A_k = \bigcup_{i=1}^{k} A_i\right) \tag{6.1}$$

Two events that occur at the same time are called an intersection of events (see Fig. 6.2).

$$\left(A = A_1 \cap A_2 \cap \ldots \cap A_k = \bigcap_{i=1}^{k} A_i\right) \tag{6.2}$$

This is a logical conjunction: the intersection $A \cap B$ occurs when event B automatically follows the occurrence of event A. Let us return to the die-toss experiment. Consider the set whose numbers are even and less than five. The even numbers are described by the set $A = \{2, 4, 6\}$ and the numbers under five are described by the set $B = \{1, 2, 3, 4\}$. The numbers two and four satisfy the condition of both sets, so that $A \cap B = \{2, 4\}$. If event B does not automatically follow event A, then A and B are mutually exclusive events, in which case $A \cap B = \{\}$.

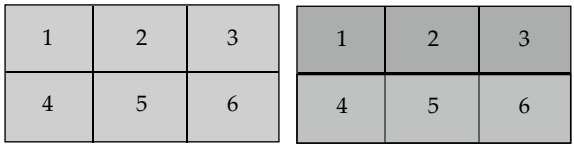


Fig. 6.1 Sample space and combined events when tossing a die

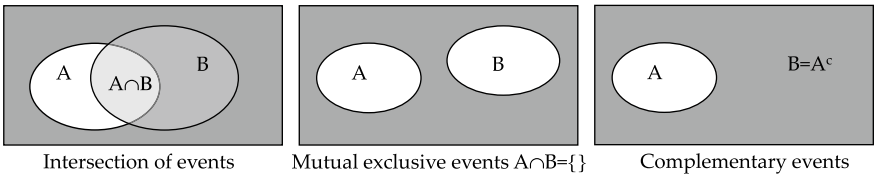


Fig. 6.2 Intersection of events and complementary events

The union of sets does not necessarily equal the total sample space Ω. As the die-toss experiment shows, three and four are mutually exclusive, but together do not describe the sample space of the experiment. This special case of mutually exclusive events describing together the whole sample space is called a complementary event. Here, event B does not automatically follow the occurrence of event A and both events make up the sample space Ω, so that $A \cup B = \Omega$ and $A \cap B = \{\}$. In the die-toss experiment, for instance, the set of even numbers A and the set of odd numbers B are mutually exclusive. Here $(A \cap B = \{\})$ and the union of sets equal the total sample space: $A \cup B = \{2, 4, 6\} \cup \{1, 3, 5\} = \{1, 2, 3, 4, 5, 6\} = \Omega$ (see Fig. 6.2).

Random experiments can be divided into single-stage experiments and multiple-stage experiments. The former consists of only one experiment; the latter, of two or more in succession. Flipping a coin three times in a row is an example of a three-stage random experiment. It can be represented using an event tree. Each path between the root and the nodes and the nodes and the leaves in the event tree depicted by Fig. 6.3 represents a possible outcome.

6.2 Probability Definitions

Until now we have only discussed events and combinations of events. While none of this appears at first to serve the goal of defining probability more concretely, it has helped to expand our understanding. Let us consider the following statements:

1. The probability of winning the lottery with six matching numbers from a range of one to 49 is one to 13,983,816.
2. The probability of a newborn weighing more than five kilogrammes is 0.3%.
3. The probability that other intelligent life exists in the universe is 50%.

Semantically, all three statements are the same: “probability” appears every time. But each statement is governed by a different model for quantifying probability. The models can be assigned to different approaches in the history of probability theory. The first example follows the rules of classical probability theory; the second, statistical probability theory; the third, subjective probability theory.

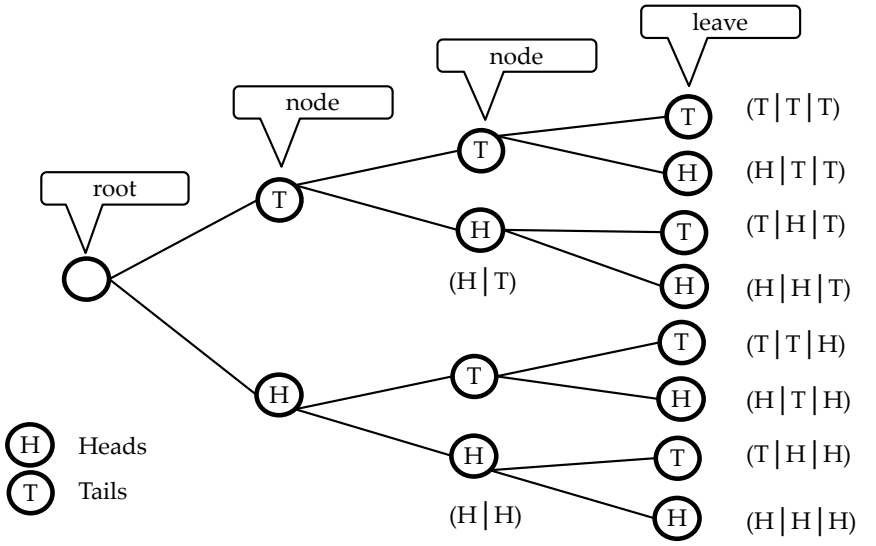


Fig. 6.3 Event tree for a sequence of three coin tosses

Classical probability theory assumes that elementary events are equally probable. The mathematician Jacob Bernoulli (1654–1705) grounded this assumption in what he called the principle of insufficient reason—also called the principle of indifference. This principle states that the probability of any event occurring is the same, provided there is no reason why one event is more probable than another. In the lottery, the combination of numbers (1, 2, 3, 4, 5, 6) is just as probable as any other. Since there are 13,986,816 possible combinations when 49 different numbers can be selected, the odds of holding the winning combination is one to 13,983,816. Using this principle Pierre-Simon Laplace (1749–1827) formulated the classical definition of probability: the ratio of the number of favourable cases to the number of possible cases:

$$P(A) = \frac{\text{number of favorable cases for event } A}{\text{number of possible cases}} = \frac{\text{Number}(A)}{\text{Number}(\Omega)} \tag{6.3}$$

This kind of probability is often referred to as Laplace probability. Consider our die-toss experiment. The probability of throwing an even number on a die is $P(\text{even numbers}) = 3/6 = 1/2$, since there are six possible outcomes and three possible even numbers. In the same vein, if a raffle drum contains 20 winning tickets and 100 tickets in all, the odds of picking a winning ticket are $P(\text{winning ticket}) = 20/100 = 0.2$.

Each of these examples assumes that the elementary events are equally probable, which is to say, every possible outcome possesses the same Laplace probability: $P(j) = 1/n$.[1] In practice, however, this is not always the case. Say an airplane passenger uses the above formula to calculate the probability of a crash. Accordingly, he assigns the value one to the number of unfavourable cases and the value two to the number of all possible cases, so that $\Omega = \{\text{crash; no crash}\}$. The resulting probability of a crash is $P(\text{crash}) = ½ = 0.5$. If this were true, then a plane would crash on average every other flight, and no one would even get near an airport. The reason for the error is that the odds of a plane's crashing are not the same as its not crashing. This is the problem with classical probability theory in its practical application.

The same problem arises with newborn bodyweight, as not every weight is equally probable. To estimate the probability of a baby being born over five kilogrammes, we must rely on the relative frequency of newborns over five kilogrammes. Since the previous share of newborns weighing over five kilogrammes is 0.3%, we assume that the probability of it happening again is 0.3%, all other things being equal.

This assertion is founded on statistical probability theory. Its origins go back to the axiomatic definition of probability provided by Andrey Kolmogorov (1903–1987). In it, every event is assigned a real number $P(A)$ that expresses the chance of it occurring as a functional equation. Statistical probability possesses the following properties:

- The probability is expressed by a non-negative number: $P(A) \geq 0$.
- If two probabilities are mutually exclusive, the sum of probabilities for the union of events is: $P(A \cup B) = P(A) + P(B)$ for $A \cap B = \{\}$.
- The probability of a certain event is 1 (= 100%).

This means that probabilities are not just random; they indicate the chance of an event actually occurring. If an event is impossible, then its probability is $P(A) = 0$; if its occurrence is certain, then its probability is $P(A) = 1$.[2]

But if the events are not equally probable, how can we allocate a probability value to every event? The scientist and mathematician Richard von Mises (1883–1953) based probability on the relative frequencies of a given event. He assumed that the occurrence of a given event is random. Since an endlessly repeated random experiment to determine the true probability is impossible, relative frequency must be used to estimate the value of empirical probability. Take as an example a coin toss with the possible outcomes "heads" or "tails". We flip the coin 100 times to determine the relative frequency of "tails". The results are shown in Fig. 6.4.

[1] Combinatorics can be used to determine the number of all possible elementary events for larger numbers of cases (see Sect. 6.3.2).

[2] $P(A) = 1$ does not necessarily mean that event A is certain. It only indicates that the relative frequency of these events in a large number of n cases is 100%. Similarly, for $P(A) = 0$ the relative frequency of event A in a large number of n cases is $P(A) = 0\%$.



Fig. 6.4 Relative frequency for a coin toss

In our test, the relative frequency is $f(x) = 3/4 = 0.75$ after four tosses; $f(x) = 13/20 = 0.65$ after 20 tosses; $f(x) = 28/70 = 0.54$ after 70 tosses; and $f(x) = 51/100 = 0.51$ after 100 tosses. The more we toss the coin, the more the relative frequency approaches the true probability of it turning up tails. Accordingly, the probability of event A occurring equals the limit value of the relative frequency when carrying out an infinite number of random experiments:

$$P(A) = \lim_{n\to\infty}(f_n(A)) = \lim_{n\to\infty}\left(\sum_{i=1}^{n}\frac{x_i}{n}\right), \quad \text{whereby: } \begin{cases} x_i = 1 \text{ for tails} \\ x_i = 0 \text{ for heads} \end{cases} \tag{6.4}$$

Jacob Bernoulli described the relationship between relative frequency and probability in a series of endlessly repeated experiments as the law of large numbers. An empirical proof for this relationship is impossible, as no one can perform an experiment an endless number of times, and every attempt at providing a mathematical proof has failed so far (Schira, 2016). But the approach provides a good basis for inferential statistics. Its probabilities are not exact, yet with a sufficiently large sample they can be estimated pretty well.

Classical and statistical probability theories are objective approaches. Both are very useful in practice but also have their limits, such as when the probability of an event must be determined before it occurs, or when an experiment can only be performed once. Our third example—the probability of there being other forms of intelligent life in the universe—is such a case. No assumptions can be made about the potential frequency of the event and its probability cannot be determined by repeat experiments. In these cases, we have to use subjective probability theory. The approach, developed by Savage (1917–1971) and De Finetti (1906–1985), individualizes how we understand probability. Here probability is defined as a measure of the trust a rational thinking person has in the occurrence of a certain event (see De Finetti, 2008; Savage, 1954). Each assessment must be based on intuition, expertise, and experience. In the subjective approach, probability expresses

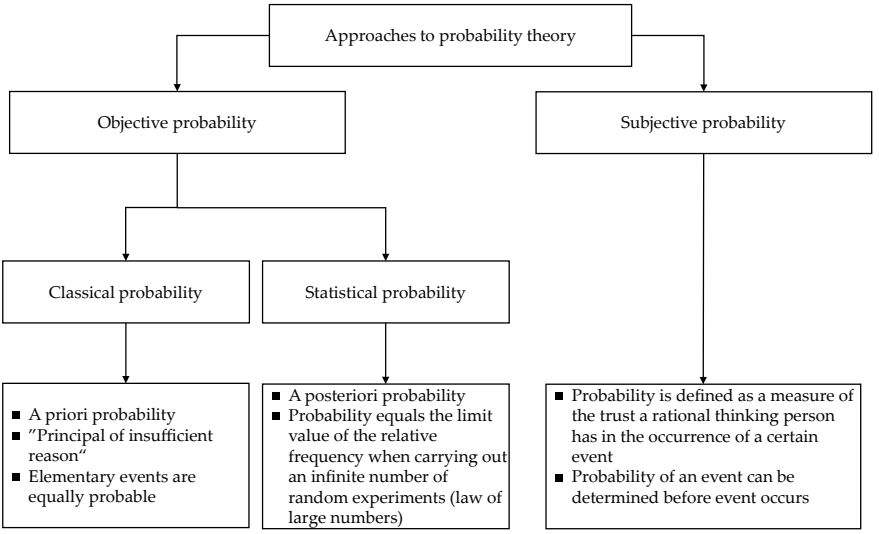


Fig. 6.5 Approaches to probability theory

an individual assessment, summed up by the idea that probability is degree of belief. A rationally thinking person, for instance, would rate the odds of a rolling a six with a normal die as one in six. These are the same odds identified by classical probability theory. But if the same person should decide that the die was loaded, his or her assessment of the probability would change. The odds represent the individual probability of an outcome. Say the person is still ready to make a bet on the next toss if the number six brings him €2 for every €1 he bets. Here the subjective probability for six is $P(X = 6) = 1/(2 + 1) = 1/3$.

Figure 6.5 summarizes the three approaches to statistical probability we discussed in this section. Next we address the basic rules for calculating probability. These apply regardless of the approach we use, holding equally for classical, statistical, and subjective probability theories. For the rules of calculation it is irrelevant how one arrives at the values for individual probabilities; the point is to calculate probability using the values one already has.

6.3 Foundations of Probability Calculus

6.3.1 Probability Tree

Probability trees are a simple way of representing probabilities and using them for calculation. They are derived from event trees and depict the probability of all possible event combinations. The paths are labelled with the probability associated with the node they leave. The sum of the probabilities on all paths must always result in 1. This ensures that all possible events are considered in the probability

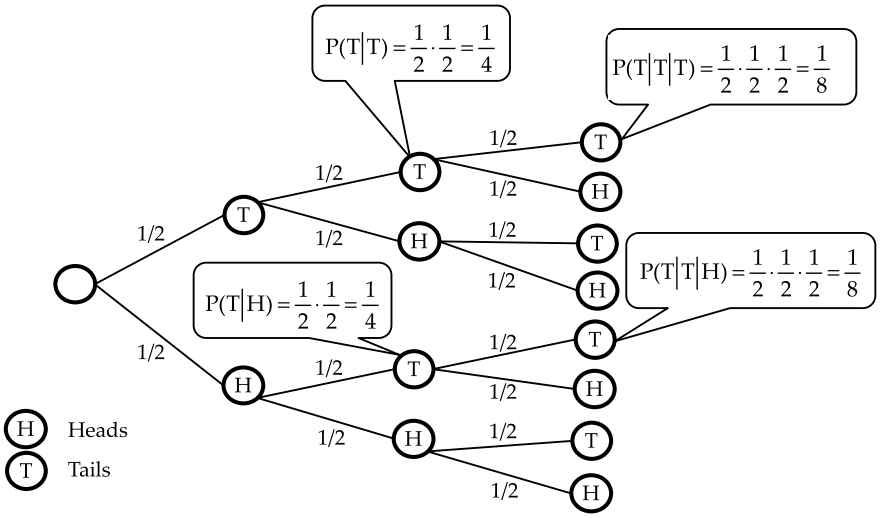


Fig. 6.6 Probability tree for a sequence of three coin tosses

tree. Figure 6.6 provides a sample illustration of a probability tree for a sequence of three coin tosses. The probability of a given sequence of outcomes can be identified by multiplying the probabilities on the path between the root and the leaf (law of multiplication for paths). The outcome sequence Heads, Tails, Heads can thus be expressed as:

$$P(H|T|H) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8} \tag{6.5}$$

The probabilities can also be determined for outcome sequences that do not end in the leaves of the probability tree. The outcome sequence Heads, Tails (with the outcome of the third toss remaining open) amounts to:

$$P(H|T) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} \tag{6.6}$$

If the outcomes consist of several paths in a tree diagram, the probability is the sum of each path probability (law of addition for paths). If we are interested in finding the probability of once throwing heads and twice throwing tails, then we must add together the probabilities of the paths $(T|H|H)$, $(H|T|H)$, and $(H|H|T)$. This yields:

$$P(T|H|H) + P(H|T|H) + P(H|H|T) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8} \tag{6.7}$$

Probability trees are suitable when there are a small number of multistage experiments, a low number of possible outcomes, and the outcome sequence matters. When a large number of possible outcomes exist and the sequence of elements plays no role, combinatorics can help us calculate Laplace probabilities.

6.3.2 Combinatorics

Combinatorics can be used not only to calculate Laplace probabilities but also to select and order elements from a finite set or population. Moreover, they are the basis for statistical—binomial or hypergeometric—distributions. These are all good reasons to turn to the topic of combinatorics in more detail. Hartung (2009, p. 96) writes that combinatorics can answer two types of questions: "How many ways are there of arranging N elements?" and "How many ways are there of selecting k elements from N?" For answering the first question we calculate what are called permutations; for the second question we calculate what are called combinations or variations.

Let us start by calculating permutations. A permutation describes only one possible arrangement of N different elements of a given set, in which each element exists only once or every element can be selected only once. Let us assume, for instance, that a customer has seen three different commercials for a product, but we do not know the order in which he saw them. The sequence "commercial #1, commercial #2, and commercial #3" is only one possible order, as is "commercial #3, commercial #2, and commercial #1". All in all, the commercials can be watched in

$$P_3^3 = 3! = 1 \cdot 2 \cdot 3 = 6 \tag{6.8}$$

different sequences.[3] This is known as a *permutation without repetition.* Every sequence has a Laplace probability of $P = 1/6$. The number of permutations without repetition of N different objects can be calculated as follows:

$$P_N^N = N! \tag{6.9}$$

A *permutation with repetition* exists when at least two elements are identical or cannot be distinguished. In this case, k different groups of elements exist. Let us assume, for instance, that our company has run four commercials. In the case of permutation without repetition there is a total of

$$P_4^4 = 4! = 24 \tag{6.10}$$

[3] The exclamation point in the expression $n!$ describes the mathematical function of the factorial. The factorial of an integer n is the product of all natural numbers equal to or smaller than n: $n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n = \prod_{k=1}^{n} k$.

different possible sequences. Say one commercial has a green background, another a red background, and the remaining two a blue background. There are commercials with $k = 3$ groups of different background colours. We have to ask ourselves how many permutations of background colours exist. Because two backgrounds repeat themselves, there must be fewer permutations than in the case without repetition. The permutations "green, red, blue 1, blue 2" and "green, red, blue 2, blue 1" are identical with regard to background colour and thus count as only one sequence, not as two. To calculate the number of permutations with repetition, we use this formula:

$$P^N_{n_1;\ldots;n_k} = \frac{N!}{n_1!n_2!\ldots n_k!} \tag{6.11}$$

Here the number of group elements N equals the sum of elements in k groups:

$$\sum_{i=1}^{k} n_i = N \tag{6.12}$$

Hence, there are

$$P^4_{1;1;2} = \frac{4!}{1!\,1!\,2!} = 12 \tag{6.13}$$

permutations (with repetition) of different background colours.

What if we are interested in the selection of elements, not in their arrangement? A *variation* is when all that matters is the selection of k elements from a set of N elements and their order. A *combination* is when all that matters is the selection of k elements from a set of N elements. Here too, they can be with or without repetition (see Fig. 6.7).

	Variations (order matters)	**Combinations** (order does not matter)
Sample with repetition	$\tilde{V}^N_n = N^n$ E.g.: Variation of all winners at N rounds of poker	$\tilde{C}^N_n = \binom{N+n-1}{n}$ E.g.: Combination of winners at n rounds of poker
Sample without repetition	$V^N_n = n!\binom{N}{n} = \frac{N!}{(N-n)!}$ E.g.: Variation when sending orders to suppliers	$C^N_n = \binom{N}{n} = \frac{N!}{n!(N-n)!}$ E.g.: Lottery

Fig. 6.7 Combination and variation. *Source* Wewel and Blatter (2019) figure modified slightly

A *combination without repetition* where order does not matter is a lottery drawing. From N balls, each stamped with a unique number, n is drawn. Because the balls are not put back into the machine, the numbers are not repeated. The order of the balls is not important for whether you win. The number of possible combinations can be calculated with the formula:

$$C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!} \tag{6.14}$$

$\binom{N}{n}$ is called the binomial coefficient. For six balls drawn from 49 balls, there are

$$C_6^{49} = \binom{49}{6} = \frac{49!}{6!(49-6)!} = 13{,}983{,}816 \tag{6.15}$$

combinations of different numbers. Somewhat more complicated is the calculation of *combinations with repetitions* where order doesn't matter. This always occurs when elements can be selected multiple times over multiple drawings and the order of the drawing is not important. Assume that four persons play two rounds of poker and you want to know how many different combinations of winners exist. In contrast to a lottery, each player can win ("be drawn") more than once. But you are only interested in how often and not in which game the players win. This combination can be calculated using the formula:

$$\tilde{C}_n^N = \binom{N+n-1}{n} \tag{6.16}$$

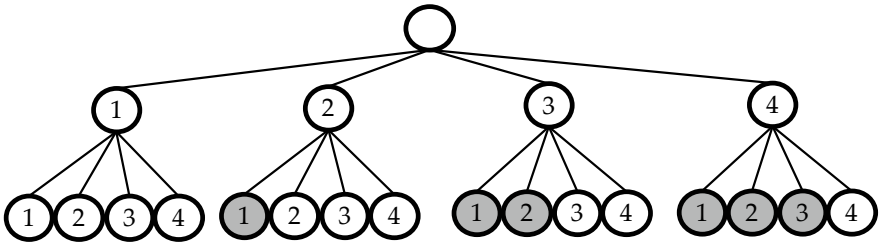
With $N = 4$ players and $n = 2$ rounds of pokers there are ten different combinations of winners that vary only by the number of victories (see Fig. 6.8):

$$\tilde{C}_2^4 = \binom{4+2-1}{2} = \binom{5}{2} = \frac{5!}{2!(5-2)!} = 10 \tag{6.17}$$

If the order of the poker winner matters, we need to calculate variation repetition where order matters. In the first game, four players can win. By the second game there are already $4 \cdot 4 = 16$ different variations. For $n = 2$ rounds and $N = 4$ players, and given

$$\tilde{V}_n^N = N^n \tag{6.18}$$

there is a total of $\tilde{V}_2^4 = 4^2 = 16$ possibilities (see Fig. 6.8).



Note:
Four persons play two rounds of poker. If a player can win more than once and order matters, there are 16 possible outcomes. If a player can win more than once and order *does not* matter, there are 10 possible outcomes. The six gray circles in the figure above indicate the redundant outcomes.

Fig. 6.8 Event tree for winner combinations and variations with four players and two games

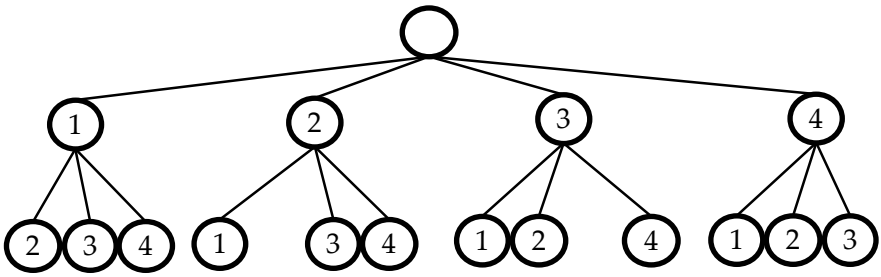
The number of *variations without repetition* when order matters can be calculated with the formula:

$$V_n^N = n!\binom{N}{n} = \frac{N!}{(N-n)!} \tag{6.19}$$

Once again, we can illustrate with poker. Say you want to know how many different sequences of winners exist when you have $n = 2$ rounds of pokers and $N = 4$ players, and every player must leave the table after winning a round (see Fig. 6.9). Using the above formula, we get

$$V_2^4 = \frac{4!}{(4-2)!} = 12 \tag{6.20}$$

Admittedly, it is not common for poker winners to leave the table. But variations with repetition where order matters can be very useful for other practical situations.



Note:
Four persons play two rounds of poker. Winners must leave the table and order matters.

Fig. 6.9 Event tree for winning variations without repetition for four players and two rounds

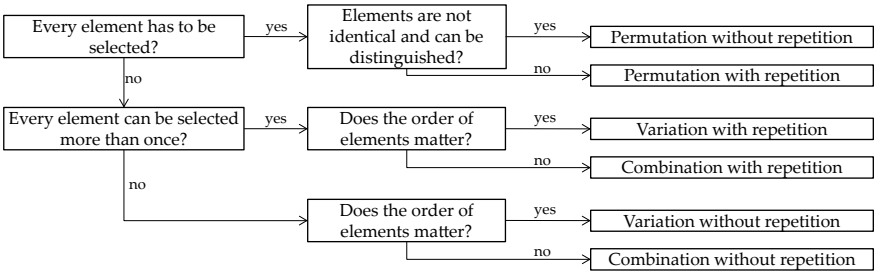


Fig. 6.10 Deciding between permutation, combination, and variation. *Source* Bourier (2018, p. 80). Compiled by the author

For example, say you want to calculate the number of possible sequences of suppliers. How many different possibilities exist when from 20 suppliers three are to be chosen and the first supplier gets 50%, the second 30%, and the third 20% of orders. In this case, succession plays a crucial role for the scope of orders received. In total, we have

$$V_3^{20} = \frac{20!}{(20-3)!} = 6840 \tag{6.21}$$

different variations. Typically, students new to statistics have a difficult time deciding if a given situation calls for a permutation, a combination, or a variation. Bourier (2018) thus recommends using an algorithmic approach, summarized in Fig. 6.10.

6.3.3 The Inclusion–Exclusion Principle for Disjoint Events

If we know the probabilities for individual events and we want to calculate the probability that two or more of them occur, we must use the Inclusion–Exclusion Principle. This involves nothing more than adding up two or more paths in a probability tree. The probability of two mutually exclusive events A and B—which is to say, $P(A \cap B) = \{\}$—equals the sum of individual probabilities for the events A and B, so that $P(A \cup B) = P(A) + P(B)$. Accordingly, for multiple mutually exclusive events, the following formula (addition rule):

$$P\left(\bigcup_{i=1}^{m} A_i\right) = \sum_{i=1}^{m} P(A_i) \tag{6.22}$$

Let us take the simple example of a die-toss experiment. The probability of rolling an even number equals the sum of individual probabilities for 2, 4, and 6. With a normal six-sided die, the individual probabilities are 1/6 each, so that the

total probability of rolling an even number is:

$$P(\text{“even number”}) = P(2) + P(4) + P(6) = 1/6 + 1/6 + 1/6 = 1/2. \quad (6.23)$$

Let us consider another example. A raffle drum contains 1000 tickets, with 10 first prizes and 80 second prizes. The individual prize probabilities are:

$$P(\text{“First Prize”}) = 10/1000 = 1\% \text{ and } P(\text{“Second Prize”}) = 80/1000 = 8\% \quad (6.24)$$

First and second prize tickets cannot be drawn at the same time; they are mutually exclusive and their intersection is empty. The probability of winning one of the prizes is the sum of the probabilities of drawing a first or second prize:

$$P(\text{“First Prize”} \cup \text{“Second Prize”}) = 0.01 + 0.08 = 9\%. \quad (6.25)$$

A special case of mutual exclusive events is that of complementary events. As I discuss above, these are mutually complementing events in some set Ω. Their probabilities also add up to one $P(A \cup B) = P(\Omega) = 1$, so that a probability can always be calculated using the others. This means that:

$$P(A) = 1 - P\left(A^c\right) = 1 - P(B) \quad (6.26)$$

Exploiting this relationship chiefly makes sense when one of the probabilities is complicated to calculate and the other easy. Let us assume as an example a die-toss experiment with two loaded dice, i.e. dice where the odds for rolling each number are not equal. What is the probability of rolling a number less than eleven? Table 6.1 presents all the combinations and their probabilities for the two dice.

To identify $P(\text{number} < 11)$, we add up the individual probabilities for the 33 dice combinations whose sum is smaller than eleven. Since the probability $P(\text{number} \geq 11)$ is complementary, all we need to do is add up the remaining probabilities (marked in grey) and subtract this value from one:

$$\begin{aligned} P(\text{number} < 11) &= 1 - P(\text{number} \geq 11) \\ &= 1 - (6.8027\% + 6.8027\% + 8.1633\%) \approx 78.23\% \end{aligned} \quad (6.27)$$

Table 6.1 Toss probabilities with two loaded dice

	1	2	3	4	5	6
1	0.2268%	0.4535%	0.6803%	0.9070%	1.1338%	1.3605%
2	0.4535%	0.9070%	1.3605%	1.8141%	2.2676%	2.7211%
3	0.6803%	1.3605%	2.0408%	2.7211%	3.4014%	4.0816%
4	0.9070%	1.8141%	2.7211%	3.6281%	4.5351%	5.4422%
5	1.1338%	2.2676%	3.4014%	4.5351%	5.6689%	*6.8027%*
6	1.3605%	2.7211%	4.0816%	5.4422%	*6.8027%*	*8.1633%*

6.3.4 Inclusion–Exclusion Principle for Nondisjoint Events

The addition of probabilities for nonmutually exclusive events—called dependent events—works somewhat differently. In this case, the intersection is a non-empty set that is counted twice when adding up two individual probabilities. Consider again the intersection of events shown in Fig. 6.2. The probability of two nonmutually exclusive events A and B equals the sum of the individual probability of events A and B minus the intersection of the events:

$$P(A \cup B) = P(A) + P(B) - (A \cap B) \tag{6.28}$$

To illustrate, let us use again the die-toss experiment. How large is the probability of rolling an even number or a number smaller than four? The set of numbers smaller than four is $A = \{1, 2, 3\}$; the set of even numbers is $B = \{2, 4, 6\}$. Both sets intersect at two ($A \cap B = \{2\}$) and possess a probability of $P(A \cap B) = 1/6$. According to the Inclusion–Exclusion Principle for nonmutually exclusive events, the probability is:

$$P(A \cup B) = P(A) + P(B) - (A \cap B) = 3/6 + 3/6 - 1/6 = 5/6 \tag{6.29}$$

6.3.5 Conditional Probability

The independence of events is an important area of inquiry. Two events are independent if the occurrence of an event has no influence on the probability of the other occurring. For instance, the result of a lottery drawing last Saturday has no influence on the drawing next Saturday (assuming, that is, that the lottery machine and the 49 balls are in proper working order). Similarly, the probability of one student passing a statistics exam should be independent of the results of the student one desk over. Let us look more closely at another example Table 6.2 (Velleman, 2002, p. 14-1).

Researchers at the Baystate Medical Center in Springfield, Massachusetts studied the birth weights of 189 newborns. A total of 59 newborns (31%) had a lower-than-average weight; 75 (40%) had an average weight; and 55 (21%) had a

Table 6.2 Birth-weight study at Baystate Medical Center

Mothers smokers/nonsmokers	Birth weights			Sum
	High	Middle	Low	
Nonsmokers	42	44	29	115
Smokers	13	31	30	74
Sum	55	75	59	189

Source Velleman (2002, p. 14-1)

higher-than-average weight. Accordingly, a pregnant woman has a 31% of giving birth to an underweight baby, so that $P(\text{low birth weight}) = 59/189 = 31\%$.

The researchers suspect that pregnant women who smoke have a higher-than-average probability of having a baby with a low birth weight than pregnant women who do not smoke. In the study, 74 of the women smoked while pregnant; 30 of them had babies with low birth weights. The probability of giving birth to an underweight baby for pregnant women who smoked while pregnant is $P(\text{low birth weight} \mid \text{smoker}) = 30/74 = 40\%$. This is called conditional probability because a restriction is applied to the persons being observed—women who smoke. The condition is separated formally from the event with a vertical line and expressed as $P(\text{event} \mid \text{condition})$. For the calculation, we divide the probability of the intersection of events ($P(\text{low birth weight} \cap \text{smoker})$) by the probability of the condition $P(\text{smoker})$. For our example, we get:

$$P(A|B) = \frac{P(\text{low birth weight} \cap \text{smoker})}{P(\text{smoker})} = \frac{\frac{30}{189}}{\frac{74}{189}} = \frac{30}{74} = 0.4 \tag{6.30}$$

Generally, then, the following holds true:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{6.31}$$

6.3.6 Independent Events and Law of Multiplication

If birth weight is not influenced by whether or not the mother smokes during pregnancy, then the probability of having an underweight baby would have to be 31% for women who smoke during pregnancy and for those who don't. That is to say, $P(\text{low birth weight} \mid \text{smoker}) = P(\text{low birth weight}) = 0.31$. As we calculated above, however, this is not the case: 40% of women who smoke while pregnant give birth to an underweight infant.

To sum up, we can say that two events are independent from each other when $P(A \mid B) = P(A)$ or when $P(B \mid A) = P(B)$.

$$\text{Since } P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ then}$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = P(A) \tag{6.32}$$

Yet this is only the case when $P(A \cap B) = P(A) \cdot P(B)$. Generally, two events A and B are independent from each when the product of the individual probability equals the probability of the intersection of events. This is also called the

law of multiplication for independent events. If, by contrast, $P(A \mid B) \neq P(A)$ or $P(A \cap B) \neq P(A) \cdot P(B)$, then the two events are dependent. But in any case the general law of multiplication applies:

$$P(A \cap B) = P(A|B) \cdot P(B) \text{ and } P(A \cap B) = P(B|A) \cdot P(A) \tag{6.33}$$

6.3.7 Law of Total Probability

From applying the Inclusion–Exclusion Principle and multiplication we can derive the law of total probability. Let's say we want to know the probability for event A and we only know the probabilities of the intersections of the probability of A with all other possible events B to Z, so that:

$$P(A) = P(A \cap B) + P(A \cap C) + \cdots + P(A \cap Z) \tag{6.34}$$

For the example in Table 6.2, the probability of low birth weight results from adding the probability of the intersections for low birth weight between pregnant women who smoke and those who don't:

$$\begin{aligned} &P(\text{low birth weight}) \\ &\quad = P(\text{low birth weight} \cap \text{smoker}) + P(\text{low birth weight} \cap \text{nonsmoker}) \end{aligned} \tag{6.35}$$

By using the law of multiplication for independent events, we get the general equation:

$$P(A) = P(A|B) \cdot P(B) + P(A|C) \cdot P(C) + \cdots + P(A|Z) \cdot P(Z) \tag{6.36}$$

Applied to the example in Table 6.2 this becomes:

$$\begin{aligned} P(\text{low birth weight}) &= P(\text{low birth weight}|\text{nonsmoker}) \cdot P(\text{nonsmoker}) \\ &\quad + P(\text{low birth weight}|\text{smoker}) \cdot P(\text{smoker}) \\ &= \frac{29}{115} \cdot \frac{115}{189} + \frac{30}{74} \cdot \frac{74}{189} = \frac{59}{189} \end{aligned} \tag{6.37}$$

Of course, we could have looked up this probability in table. But this is not always possible.

Consider a company whose products undergo computer-controlled quality inspection. From past experience, the company knows that the probability of finding a defective product is $P(\text{defect}) = 0.01$. The computer-controlled inspection has a 99% probability ($P(\text{out} \mid \text{defect}) = 99\%$) of detecting a defective product. In $P(\text{out} \mid \text{not defect}) = 0.5\%$ of the cases, however, non-defective products are

mistakenly sorted out. What is the probability that the computer-controlled inspection will sort out a product ($P(\text{out})$)? Using the law of total probability, we can combine the probabilities of a product being mistakenly sorted out and a product correctly sorted out:

$$P(\text{out}) = P(\text{out} \cap \text{defect}) + P(\text{out} \cap \text{not defect}) \tag{6.38}$$

This equals:

$$P(\text{out}) = P(\text{out}|\text{defect}) \cdot P(\text{defect}) + P(\text{out}|\text{not defect}) \cdot P(\text{not defect}) \tag{6.39}$$

The resulting probability for a product being sorted out is:

$$P(\text{out}) = 0.99 \cdot 0.01 + 0.005 \cdot (1 - 0.99) = 1.49\% \tag{6.40}$$

6.3.8 Bayes' Theorem

Now let us look at the products that have been sorted out. In what percentage of cases is the product actually defective ($P(\text{defect} \mid \text{out})$)? This question can be answered using a theorem proposed by Thomas Bayes (1702–1761). The derivation of the relationship is simple. From Sect. 6.3.6 we know that both $P(A \cap B) = P(A|B) \cdot P(B)$ and $P(A \cap B) = P(B|A) \cdot P(A)$ are true. Each equation can be formulated in terms of the other, so that $P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$. This in turn can be translated into Bayes' theorem:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \quad \text{for } P(B) > 0 \tag{6.41}$$

If we express the denominator $P(B)$ using the law of total probability, we arrive at the general formula for Bayes' theorem:

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{P(B)} = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{j=1}^{n} P\left(B|A_j\right) \cdot P\left(A_j\right)} \quad \text{for } \sum_{j=1}^{n} P\left(B|A_j\right) \cdot P\left(A_j\right) > 0 \tag{6.42}$$

For our example, this means that the probability of a correctly sorted product results from:

$$P(\text{defect}|\text{out}) = P(\text{out}|\text{defect}) \cdot \frac{P(\text{defect})}{P(\text{out})} = 0.99 \cdot \frac{0.01}{0.0149} = 66\% \tag{6.43}$$

This means that around every third product sorted out is mistakenly labelled defective.

Conditional probabilities—calculated by Bayes' theorem—are much more important for practice than simple individual probabilities. They point to the particular features of certain groups and detect their relationships. Yet in general discussions, conditions tend to be left out, with probabilities distorted or made universally valid. Krämer (2015) gives a few examples. One is the statistic that two thirds of all male drinkers are married. But the conclusion that married life causes men to hit the bottle is erroneous. The conditional probability of married people in the group of male drinkers ($P(\text{married} \mid \text{drinkers}) = 66\%$) is not identical with the conditional probability of being a male drinker in the group of married people ($P(\text{drinkers} \mid \text{married}) = ?$). How can we find out the probability that a married man or an unmarried man becomes a drinker? To figure out which group—married men or unmarried men—tends more towards alcoholism, we need to compare conditional probabilities.

Bayes' theorem helps solve this kind of problem. The probability that a German selected at random is single is around $P(\text{single}) = 22\%$[4]; the probability of alcoholism is around $P(\text{drinkers}) = 2\%$.[5] From this, we get:

$$\begin{aligned} P(\text{drinkers}|\text{single}) &= \frac{P(\text{single}|\text{drinkers}) \cdot P(\text{drinkers})}{P(\text{single})} \\ &= \frac{0.33 \cdot 0.02}{0.22} = 3.0\% \end{aligned} \tag{6.44}$$

Married men show a lower probability of alcoholism:

$$\begin{aligned} P(\text{drinkers}|\text{married}) &= \frac{P(\text{married}|\text{drinkers}) \cdot P(\text{drinkers})}{P(\text{married})} \\ &= \frac{0.66 \cdot 0.02}{0.78} = 1.7\% \end{aligned} \tag{6.45}$$

This illustrates once again how important it is to observe conditional probability exactly. Bayes' theorem allows us to calculate an unknown conditional probability using a known conditional and two total unconditional probabilities.

6.3.9 Postscript: The Monty Hall Problem

We want to illustrate Bayes' theorem using the well-known Monty Hall problem. This is a puzzle based on the TV show *Let's Make a Deal*, which was broadcast in the United States in the 1960s and 70s. Contestants were shown three closed doors. Behind one of the closed doors was a large prize—usually a car—and behind each

[4] Author's own calculations. Single persons are defined as those who do not live in family-like circumstances (see Statistisches Bundesamt, 2009, p. 38).

[5] Around 1.3 million German citizens are considered alcoholics (see Drogenbeauftragte der Bundesregierung—Bundesministerium für Gesundheit, 2009, p. 38).

of the other doors a conciliation prize, which in some cases turned out to be a goat. After the contestant decided for one of the three doors, Monty Hall, the host, opened one of the two remaining doors. Naturally it was always the door that did not have the main prize. Contestants were then given the choice to decide for the other door or remain with their first choice. Is the probability of winning the main prize higher when contestants choose the other door or when they stick with their initial pick? Or does it even matter? Are the probabilities not equal—fifty-fifty?

The Monty Hall problem has been widely discussed and is yet another reminder of the wide gulf that sometimes exists between intuition and statistical truth. The columnist Marilyn vos Savant was ridiculed for claiming that contestants who change their mind increase their odds of winning, but it turns out her critics were wrong. Indeed, it can be shown that the probability of winning doubles when contestants change their first choice and decide for the other door.

Let's try to solve this problem without math. To repeat: there is a total of three doors. Before the host opens the door, the probability that the main prize is behind one of the doors is $P(\text{main prize}) = 1/3$. Let us assume that contestants decide first for door #1. What does the host do?

- Possibility 1: The main prize is behind door #1. In this case, the host will open either door #2 or door #3, one of the conciliatory prizes. In this case, contestants lose when they decide for another door.
- Possibility 2: The main prize is behind door #2. In this case, the host must open door #3, otherwise he'd reveal the location of the main prize. In this case, contestants win the main prize when they change their minds.
- Possibility 3: The main prize is behind door #3. In this case, the host opens door #2, otherwise he'd reveal the main prize when opening door #3. In this case contestants win the main prize when they change their minds.

As we see, contestants win the main prize in two of three cases when they change their minds. These probabilities do not change if contestants first choose door #2 or door #3. This is the mirror of the approach described above. The relationship can be represented using the probability tree in Fig. 6.11.

Marilyn vos Savant explains why switching makes sense with an extended example: "The first door has a 1/3 chance of winning, but the second door has a 2/3 chance. Here's a good way to visualize what happened. Suppose there are a million doors, and you pick door #1. Then the host, who knows what's behind the doors and will always avoid the one with the prize, opens them all except door #777,777. You'd switch to that door pretty fast, wouldn't you?"[6]

Bayes' theorem can also be used to solve the Monty Hall problem. First, we must define the events $G_i = \{\text{"main prize is located behind door } i\text{"}\}$ and $H_j = \{\text{"Host opens door } j\text{"}\}$. The a priori probability of a door containing the main prize

[6] See https://people.math.harvard.edu/~knill/oldtexas/Teaching/MA464/PUZZLES/marilyn.gameshow.html.

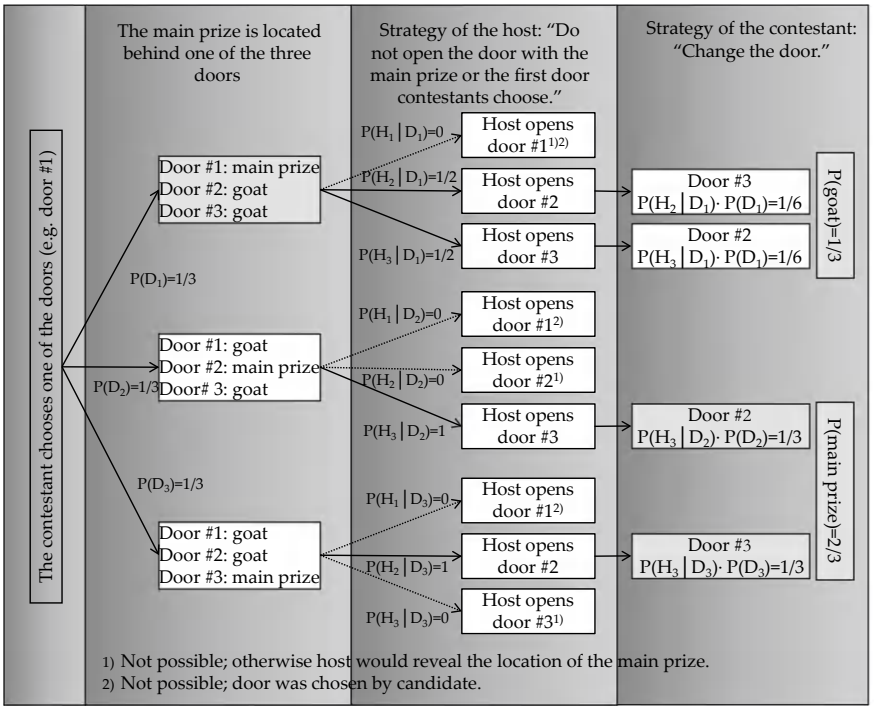


Fig. 6.11 Probability tree of the Monty Hall problem. This probability tree assumes that the host does not open the door with the main prize or the first door selected. It also assumes that contestants can choose any door. That is, even if contestants pick a door other than #1, the probability of winning stays the same. The winning scenarios are in grey

is $P(D_1) = P(D_2) = P(D_3) = 1/3$. In our example, the contestant decides for door #1, though the results are the same regardless which door the contestant chooses. Now the host opens either door #2 or door #3. Assuming he opens door #3, the contestant must gauge the probability of winning if he changes his mind ($P(D_2 \mid H_3)$) and the probability of winning if he doesn't change his mind ($P(D_1 \mid H_3)$). We also know the following:

- If the main prize is behind door #3, the host must open door #2, otherwise he would reveal the location of the prize. In this event, the probability of opening door #3 is zero: $P(H_3 \mid D_3) = 0$.
- If the main prize is behind door #2, the host must open door #3, otherwise he would reveal the location of the main prize with a probability of $P(H_3 \mid D_2) = 1$.
- If the main prize is behind door #1, the host can open door #2 or door #3, so that $P(H_3 \mid D_1) = 1/2$.

Using Bayes' theorem we get:

$$P(D_1|H_3) = \frac{P(H_3|D_1) \cdot P(D_1)}{P(H_3|D_1) \cdot P(D_1) + P(H_3|D_2) \cdot P(D_2) + P(H_3|D_3) \cdot P(D_3)}$$
$$= \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}} = \frac{1}{3} \tag{6.46}$$

$$P(D_2|H_3) = \frac{P(H_3|D_2) \cdot P(D_2)}{P(H_3|D_1) \cdot P(D_1) + P(H_3|D_2) \cdot P(D_2) + P(H_3|D_3) \cdot P(D_3)}$$
$$= \frac{1 \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}} = \frac{2}{3} \tag{6.47}$$

These formulas show the probability of winning is significantly greater when contestants change their minds than when they don't.

6.4 Chapter Exercises

Exercise 1
Inge is a sales representative for a shampoo brand. She plans to visit ten hairstylists who work at different salons as she makes her daily rounds.

(a) How many possible routes can Inge take?
(b) Of the ten hairstylists, five have blond hair, three have black hair, and two have red hair. How many possible routes exist if Inge can only differentiate hairstylists by the colour of their hair?
(c) Inge wants to invite six of the ten hairstylists to her 25th birthday party. How many possible combinations does she have to choose from?

Exercise 2
Twenty companies have advertised a job opening. There are twenty people seeking work through the job centre who qualify for the positions. Sixteen applications are from women and four from men. If the clerk at the job centre can distinguish applicants only by gender, how many different possibilities exist for assigning them to the job openings?

Exercise 3
A marketing executive decides to limit his company's ads to three German TV channels. Under consideration are the German networks ARD, ZDF, RTL, SAT1, and PRO 7.

(a) How many combinations of three channels exist for the executive to choose from? List all the possibilities.

(b) The marketing executive wants to run five different commercials five times a day on the three selected channels, with the option of repeating one multiple times. How many possibilities does he have if the order of the commercials does not matter?

Exercise 4

(a) Ten horses take part in a horse race. How many possibilities exist for picking the top three horses in the order they finish?
(b) A track and field club has six equally fast sprinters from which to make a $4 \cdot 100$ relay team. How many possibilities exist assuming distinctions are made between the relay positions?

Exercise 5
A cigarette dispenser has twelve slots and ten different brands. How many ways are there to stock the dispenser, if the number of packages of a brand is all that matters?

Exercise 6
Inge knows fifty hairstylists. How great is the likelihood that at least two of the 50 hairstylists have the same birthday? This is known as the birthday paradox.

Exercise 7
One-hundred students attend a statistics lecture. For the exam, students have two practice exams available. Fifty per cent of the students complete practice exam #1. Fifteen per cent complete practice exam #1, but not practice exam #2. Twenty per cent complete practice exam #2 but not practice exam #1. Student A is picked at random. Student A indicates that he/she completed practice exam 1. B indicates that he/she completed practice exam #2.

(a) Using the appropriate logical operators between A and B represent the following events and determine the probability that the student completed
    1. no practice exam,
    2. exactly one practice exam,
    3. both practice exams,
    4. at least one practice exam, and
    5. at least one practice exam.
(b) Are the events stochastically independent? Explain your answer.
(c) The student has completed practice exam #1. What is the probability that he also completed practice exam #2?

Exercise 8
Counterfeiters expand their activities! The probability that a euro bill is counterfeit is 0.5%. Retailers are vigilant and install machines at the cash registers to identify counterfeit money. The devices are able to detect a counterfeit euro bill in 95% of cases. In 1% of cases, they record a euro bill as counterfeit that is genuine.

(a) What is the probability of the machines detecting a counterfeit bill?
(b) What is the probability that the money is really counterfeit?

Exercise 9
From experience you know that 78% of students pass the statistics exam. Of those who pass the exam, 67% go on holiday after the results are posted. Those who fail usually want to study some more, but of those 25% holiday for a while anyway.

(a) Represent the individual probabilities using a probability tree.
(b) What is the probability that a student will go on holiday after the exam results are posted?
(c) You call up another student from the statistics class and hear on his answering machine, "I'm not home. I'm in Teneriffa!" What is the probability that he passed the statistics exam?

Exercise 10
A market research institute determines that 80% of all newly introduced products flop in a certain branch. But even when the decision to mass produce a new product is based on the results of a test market, mistakes happen. From past experience, we know that 70% of successful products performed well in the test market. In 20% of cases, they performed mediocrely, and in 10% of cases they performed poorly. With flops, it's exactly the opposite.

(a) Represent the individual probabilities using a probability tree.
(b) How great is the probability that test market success will mean general market success?
(c) A product achieves good test market results. What are the odds that it nevertheless flops?
(d) What is the probability that a product performs well in the test market or turns out a flop?
(e) What is the probability that a product performs poorly in the test market and is a flop?

Exercise 11
Three machines A, B, and C produce 50%, 30%, and 20% of the entire factory output, respectively. The reject rates for the machines are 3%, 4%, and 5%, respectively.

(a) Represent the individual probabilities using a probability tree.
(b) What is the probability that a randomly selected piece is defective?
(c) What is the probability that a randomly selected defective part is from machine A?
(d) What is the probability that a randomly selected part is from machine A or defective?
(e) What is the probability that a randomly selected part is from machine B and is not defective?

Exercise 12
A finish inspection at a car manufacturer paint shop assigns one of three scores per car body:

$A_1{}^*$—perfect with $P(A_1{}^*) = 0.85$
$A_2{}^*$—streaks present $P(A_2{}^*) = 0.12$
$A_3{}^*$—air bubbles present with $P(A_3{}^*) = 0.08$.

A_1 means perfect; A_2 only streaks are present; A_3 only air bubbles are present; A_4 streaks and air bubbles are present.

(a) Represent the events A_1, A_2, A_3, and A_4 using a Venn diagram.
(b) Calculate the probability for A_i, with $i = 1, 2, 3, 4$!
(c) Are A_2* and A_3* independent events?
(d) What are the sizes of the conditional probabilities $P(A_2{}^* \mid A_3{}^*)$ and $P(A_3{}^* \mid A_2{}^*)$?

6.5 Exercise Solutions

Solution 1

(a) This represents a permutation of N different elements:
Because $P_N^N = N!$ there are $P_{10}^{10} = 10! = 3{,}628{,}800$ possible routes.
(b) This represents a permutation of N elements with k groups:
$P_{n_1;\ldots;n_k}^N = \frac{N!}{n_1!n_2!\cdot\ldots\cdot n_k!} = 5$; hence, there are $P_{5;3;2}^{10} = \frac{10!}{5!3!2!} = 2520$ different routes.
(c) This represents a combination without repetition where order does not matter. From ten hairstylists, six are invited. The order the six persons are invited is irrelevant. Hence, there exist

$$C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!} \rightarrow C_6^{10} = \binom{10}{6} = \frac{10!}{6!(10-6)!} = \frac{10!}{4!6!}$$
$$= 210 \text{ possibilities.}$$

Solution 2
This represents a permutation of N elements with k different groups:
Hence, there exist $P_{n_1;\ldots;n_k}^N = \frac{N!}{n_1!n_2!\cdot\ldots\cdot n_k!} \rightarrow P_{16;4}^{20} = \frac{20!}{16!4!} = 4845$ different possibilities.

Solution 3

(a) This represents a combination without repetition where order does not matter. Hence, there exist:

$$C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!} \rightarrow C_3^5 = \binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{5!}{3!2!}$$
$$= 10 \text{ possibilities.}$$

Written out:

1	2	3	4	5	6	7	8	9	10
ARD	ARD	ARD	ARD	ARD	ARD	PRO7	PRO7	PRO7	RTL
PRO7	PRO7	PRO7	RTL	RTL	SAT1	RTL	RTL	SAT1	SAT1
RTL	SAT1	ZDF	SAT1	ZDF	ZDF	SAT1	ZDF	ZDF	ZDF

(b) This represents a combination with repetition where order does not matter. Hence, there exist:

$$\tilde{C}_n^N = \binom{N+n-1}{n} \rightarrow \tilde{C}_5^3 = \binom{3+5-1}{5} = \binom{7}{5} = 21 \text{ possibilities.}$$

Solution 4

(a) This represents a variation without repetition where order does matter. Hence, there exist

$$V_n^N = n!\binom{N}{n} = \frac{N!}{(N-n)!} \rightarrow V_3^{10} = \frac{10!}{(10-3)!} = \frac{10!}{7!} = 720 \text{ possibilities.}$$

(b) This represents a variation without repetition where order matters. If the order of the first four starters plays a role, there exist:

$$V_n^N = n!\binom{N}{n} = \frac{N!}{(N-n)!} \rightarrow V_4^6 = \frac{6!}{(6-4)!} = 360 \text{ possibilities.}$$

Solution 5
This represents a combination with repetition (twelve Marlboro packs are possible) where order plays no role (only the number packs). Hence, there exist:

$$\tilde{C}_n^N = \binom{N+n-1}{n} \rightarrow \tilde{C}_{12}^{10} = \binom{10+12-1}{12} = \binom{21}{12}$$
$$= 293{,}930 \text{ possibilities.}$$

Solution 6
The year has $N = 365$ days.

With $V_n^{365} = \frac{365!}{(365-n)!}$ possibilities, all n persons have birthdays on different days (variation without repetition). For n persons there is a total of $\tilde{V}_n^{365} = 365^n$ possibilities of birthday configurations (variation with repetition).

$\overline{A}$: All n persons have different birthdays.
A: At least two people have the same birthday.

$$P(A) = 1 - P(\overline{A}) = 1 - \frac{\frac{365!}{(365-n)!}}{365^n} = 1 - \frac{365 \cdot 364 \cdot \ldots \cdot (365-n+1)}{365^n}$$
$$= 1 - 0.0296 = 0.97037$$

n	2	10	20	25	30	40	60	80
$P(A)$	0.27%	9.46%	41.14%	56.86%	70.63%	89.12%	99.41%	99.99%

Solution 7

(a) A: Took practice exam #1; B: Took practice exam #2;
$\overline{A}$: Did not take practice exam #1; $\overline{B}$: Did not take practice exam #2;

	A	$\overline{A}$	
B	0.35	0.2	0.55
$\overline{B}$	0.15	0.3	0.45
	0.5	0.5	1

1. $P(\overline{A} \cap \overline{B}) = 0.3$
2. $P(A \cup B) - P(A \cap B) = P(A) - P(A \cap B) + P(B) - P(A \cap B) = 0.55 - 0.35 + 0.5 - 0.35 = 0.35$
   Alternatively: $P(A \cap \overline{B}) + P(\overline{A} \cap B) = 0.15 + 0.2 = 0.35$
3. $P(A \cap B) = 0.35$
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.7$
   Alternatively: $P(A \cap B) + P(\overline{A} \cap B) + P(A \cap \overline{B}) = 0.35 + 0.2 + 0.15 = 0.7$

5. $1 - P(A \cap B) = 0.65$
   Alternatively: $P(\overline{A} \cap \overline{B}) + P(\overline{A} \cap B) + P(A \cap \overline{B}) = 0.3 + 0.2 + 0.15 = 0.65$

(b) The events are stochastically dependent. Only if $P(A) = P(A \mid B)$ are they independent.
   Proof: $P(A) = 0.5 \neq P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.35}{0.55} = 0.636$

(c) We can use the following formula:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.35}{0.5} = 0.7$$

or Bayes' theorem: $P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)} = \frac{0.636 \cdot 0.55}{0.5} = 0.7$

Solution 8

(a) Events:
   F: counterfeit money $\overline{F}$: no counterfeit money
   D: machine detects counterfeit money $\overline{D}$: machine does not detect counterfeit money
   This yields the following probabilities:

$$P(F) = 0.005;\ P\left(\overline{F}\right) = 0.995;\ P(D|F) = 0.95;\ P\left(D\middle|\overline{F}\right) = 0.01$$

According to the law of total probability, the probability that the machine detects counterfeit money is:

$$\begin{aligned} P(D) &= P(D|F) \cdot P(F) + P\left(D|\overline{F}\right) \cdot P\left(\overline{F}\right) \\ &= 0.95 \cdot 0.005 + 0.01 \cdot 0.995 = 0.0147 \end{aligned}$$

This means that the machine detects counterfeit money in 1.47% of cases.

(b) According to Bayes' theorem for conditional probability $P(F|D)$, the probability that a euro bill is really counterfeit when the machine's alarm goes off is

$$P(F|D) = \frac{P(D|F) \cdot P(F)}{P(D)} = \frac{0.95 \cdot 0.005}{0.0147} = 0.323, \text{ or } 32.3\%.$$

Solution 9

(a) Events:
   B: The student passes the statistics exam.
   $\overline{B}$: The student fails the statistics exam.
   U: The student goes on vacation.

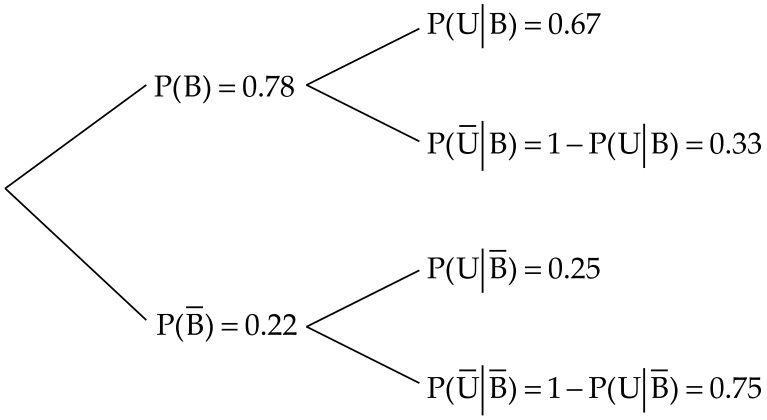


Fig. 6.12 Probability tree for statistics exam and holiday

$\overline{U}$: The student does not go on vacation.
The probability of passing the statistics exam is $P(B) = 0.78$. The probability of failing is:

$$P\left(\overline{B}\right) = 1 - P(B) = 1 - 0.78 = 0.22$$

Of the students who passed the exam, 67% go on holiday after the results are posted, so that $P(U \mid B) = 0.67$. Those who fail want to study more, but 25% go on holiday anyway after the results are posted, so that $P\left(U \mid \overline{B}\right) = 0.25$. This yields the probability tree in Fig. 6.12.

(b) The law of total probability gives us:

$$\begin{aligned} P(U) &= P(U|B) \cdot P(B) + P\left(U|\overline{B}\right) \cdot P\left(\overline{B}\right) \\ &= 0.67 \cdot 0.78 + 0.22 \cdot 0.25 = 0.5776 \end{aligned}$$

A student has a 57.76% probability of going on holiday after the results are posted.

(c) Bayes' theorem gives us:
$P(B|U) = \frac{P(U|B) \cdot P(B)}{P(U)} = \frac{0.67 \cdot 0.78}{0.5776} = 0.905$. This means that the student has a 90.5% probability of having passed the exam.

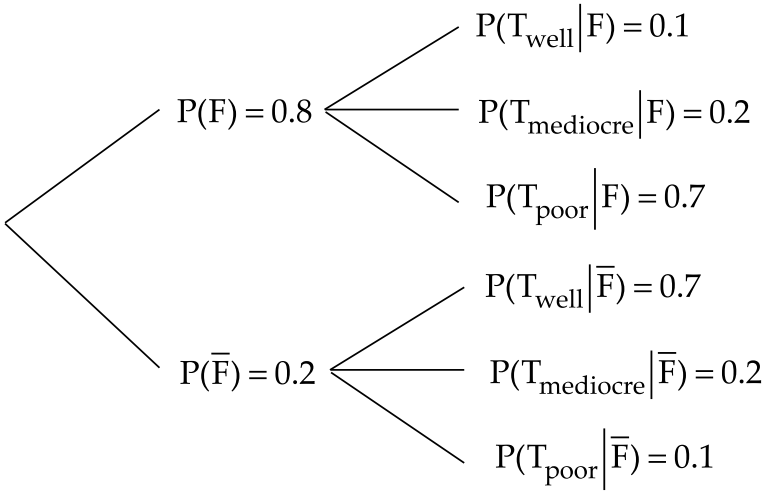


Fig. 6.13 Probability tree for test market

Solution 10

(a) Events: F: Flop; $\overline{F}$: no Flop
T_{well}: Product performs well in the test market
$T_{mediocre}$: Product performs mediocrely in the test market
T_{poor}: Product performs poorly in the test market
See the probability tree in Fig. 6.13.

(b) Total probability: $P(T_{well}) = 0.1 \cdot 0.8 + 0.7 \cdot 0.2 = 0.22$

(c) Bayes' theorem:

$$P(F|T_{well}) = \frac{P(F \cap T_{well})}{P(T_{well})} = \frac{P(T_{well}|F) \cdot P(F)}{P(T_{well})} = \frac{0.1 \cdot 0.8}{0.1 \cdot 0.8 + 0.7 \cdot 0.2} = 0.3636$$

(d) Inclusion–Exclusion Principle:

$$P(F \cup T_{well}) = P(F) + P(T_{well}) - P(F \cap T_{well}) = 0.8 + 0.22 - 0.1 \cdot 0.8 = 0.94$$

(e) Law of multiplication: $P\left(F \cap T_{poor}\right) = P\left(T_{poor}|F\right) \cdot P(F) = 0.7 \cdot 0.8 = 0.56$

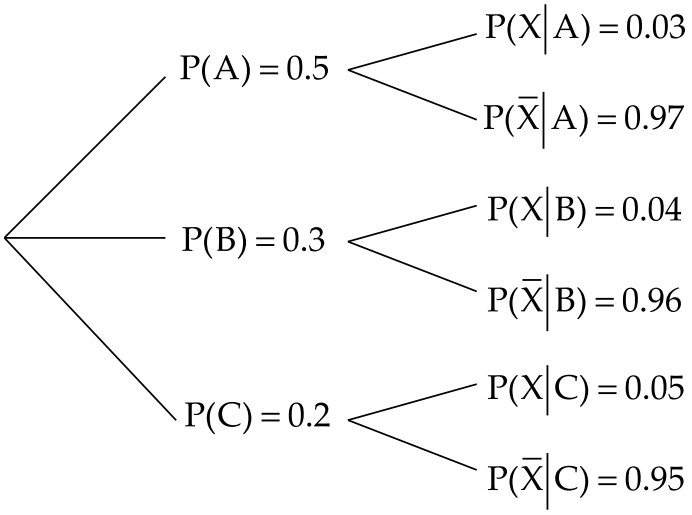


Fig. 6.14 Probability tree for defective products

Solution 11

(a) Events:
A: Machine A manufactures a part.
B: Machine B manufactures a part.
C: Machine C manufactures a part.
X: Part is defective.
$\overline{X}$: Part is not defective.
See the probability tree in Fig. 6.14.

(b) Total probability: $P(X) = 0.03 \cdot 0.5 + 0.04 \cdot 0.3 + 0.05 \cdot 0.2 = 0.037$

(c) Bayes' theorem:

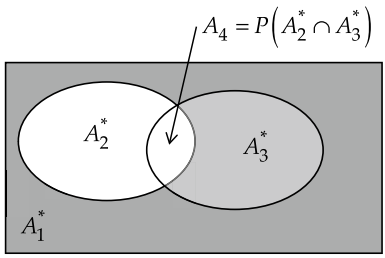
$$P(A|X) = \frac{P(A \cap X)}{P(X)} = \frac{P(X|A) \cdot P(A)}{P(X)} = \frac{0.03 \cdot 0.5}{0.037} = 0.4054$$

(d) Inclusion–Exclusion Principle:

$$P(A \cup X) = P(A) + P(X) - P(A \cap X) = 0.5 + 0.037 - 0.03 \cdot 0.5 = 0.522$$

(e) Law of multiplication: $P\left(B \cap \overline{X}\right) = P\left(\overline{X}|B\right) \cdot P(B) = 0.96 \cdot 0.3 = 0.288$

Fig. 6.15 Paint shop



Solution 12

(a) See Fig. 6.15.

(b) $P(A_1) = P\left(A_1^*\right) = 0.85$

$P(A_4) = P\left(A_2^* \cap A_3^*\right) = 0.85 + 0.12 + 0.08 - 1 = 0.05$ (all perfect: (1) perfect; (2) with streaks; (3) with air bubbles)

$P(A_2) = P\left(A_2^*\right) - P\left(A_2^* \cap A_3^*\right) = 0.12 - 0.05 = 0.07$ (all with streaks without (streaks with air bubbles))

$P(A_3) = P\left(A_3^*\right) - P\left(A_2^* \cap A_3^*\right) = 0.08 - 0.05 = 0.03$

(c) $P\left(A_2^* \cap A_3^*\right) = 0.05 \neq P\left(A_2^*\right) \cdot P\left(A_3^*\right) = 0.12 \cdot 0.08 = 0.096 \rightarrow$ dependence

(d) The probabilities are:

$$P\left(A_2^* \middle| A_3^*\right) = \frac{P\left(A_2^* \cap A_3^*\right)}{P\left(A_3^*\right)} = \frac{0.05}{0.08} = 0.625 \text{ and}$$

$$P\left(A_3^* \middle| A_2^*\right) = \frac{P\left(A_2^* \cap A_3^*\right)}{P\left(A_2^*\right)} = \frac{0.05}{0.12} = 0.41\overline{6}$$

References

Bourier, G. (2018). *Wahrscheinlichkeitsrechnung und Schließende Statistik. Praxisorientierte Einführung mit Aufgaben und Lösungen* (9th ed.). SpringerGabler.

de Finetti, B. (2008). *Philosophical lectures on probability, collected, edited, and annotated by Alberto Mura.* Springer.

Drogenbeauftragte der Bundesregierung—Bundesministerium für Gesundheit. (2009). *Drogen- und Suchtbericht Mai 2009.*

Hartung, J. (2009). *Statistik. Lehr- und Handbuch der angewandten Statistik* (15th ed.). Oldenbourg.

Krämer, W. (2015). *So lügt man mit Statistik* (17th ed.). Campus.

Savage, L. J. (1954). *Foundation of statistics.* Wiley.
Schira, J. (2016). *Statistische Methoden der VWL und BWL, Theorie und Praxis* (5th ed.). Pearson.
Schwarze, J. (2013). *Grundlagen der Statistik 2: Wahrscheinlichkeitsrechnung und induktive Statistik* (10th ed.). NWB Verlag.
Statistisches Bundesamt. (2009). *Deutschland—Land und Leute*. Destatis.
Velleman, P. (2002). *ActivStats computer software.*
Wewel, M. C., & Blatter, A. (2019). *Statistik im Bachelor-Studium der BWL und VWL. Methoden, Anwendung, Interpretation* (4th ed.). Pearson.



Random Variables and Probability Distributions

7

In the previous chapter we learned about various principles and calculation techniques dealing with probability. Building on these lessons, we will now examine some theoretical probability distributions that allow us to make inferences about populations from sample data. These probability distributions are based on the idea of random variables. A random variable is a variable whose numerical values represent the outcomes of random experiments. Random variables are symbolized with capital letters such as "X". The individual values of random variables are represented either with a capital Roman letter followed by a subscript "i" (e.g. "X_i") or with the lower case of the random variable letter (e.g. "x"). Generally, there are two types of random variables:

1. Discrete random variables result from random experiments that can produce a finite or countably infinite number of values. Frequently, though not necessarily, their values are integers. The sum of the value probabilities from random variables always equals 1. An example of a discrete random variable X is the roll outcome of a die, where the possible values are $\{1, 2, 3, 4, 5, 6\}$ for X_i.
2. Continuous random variables result from random experiments that can produce an infinite number of values. Their value range is $\mathbb{R}$. But the infinite number of possible values can nevertheless be restricted to a specific interval. The value of the integral below the probability function always equals 1. An example of a continuous random variable is the lifespan of electronic devices.

Both types of random variables form probability distributions indicating the likelihood that a specific value or value interval occurs in a random experiment. And both types of probability distributions can be represented by a probability function and a distribution function. Figure 7.1 illustrates the probability function and distribution function of a discrete random variable from a die experiment.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*,
Springer Texts in Business and Economics,
https://doi.org/10.1007/978-3-031-78070-7_7

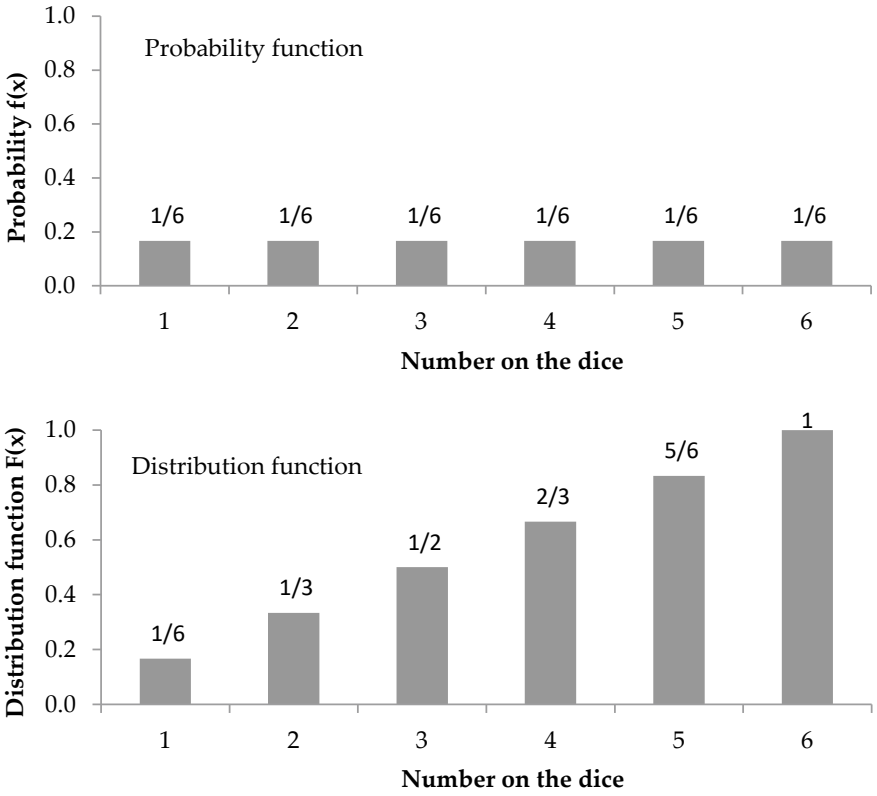


Fig. 7.1 Probability function and distribution function of a die roll experiment

As with the discrete frequency distribution, we can calculate mean and variance for the discrete probability distribution. For random variables, the mean is referred to as the *expected value*. It is calculated by adding the random variable values and weighting them with their probabilities. The die roll experiment illustrated in Fig. 7.1 produces the following expected value:

$$E(X) = \mu_X = \sum_{i=1}^{N} X_i \cdot P(X_i) = \sum_{i=1}^{6} X_i \cdot \frac{1}{6}$$
$$= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5 \qquad (7.1)$$

The variance of the expected value is:

$$\sigma_X^2 = \sum_{i=1}^{N} (X_i - \mu_X)^2 \cdot P(X_i)$$
$$= (1 - 3.5)^2 \frac{1}{6} + (2 - 3.5)^2 \cdot \frac{1}{6} + \cdots + (6 - 3.5)^2 \cdot \frac{1}{6} = 2.9166 \qquad (7.2)$$

Theoretically, numerous different empirical probability distributions result from real-life observations of relative frequency distributions. It is certain, for instance, that the probability distributions of random variable $X = \{\text{number of customers in a store}\}$ differ from those of random variable $Y = \{\text{shelf life of yoghurt}\}$.

However, many of these empirical distributions can be approximated with just a few theoretical probability functions. Accordingly, some of these theoretical probability functions have special importance for scientific and empirical research.

The simplest form can be found in Fig. 7.1, which shows a discrete uniform probability function. Other theoretical distributions have become very popular because they are well-suited to describe typical discrete phenomena in business and economics. These are the binominal distribution, the hypergeometric distribution, and the Poisson distribution. We briefly discuss them below.

7.1 Discrete Distributions

7.1.1 Binomial Distribution

One of the most important discrete distributions is the binomial distribution. The term *binomial* derives from the Latin phrase *ex binis nominibus*, meaning "consisting of two expressions". Accordingly, a binomial distribution is made up of only two unique values (A and B), which taken together occupy the entire probability space, so that $A \cup B = \Omega$. Each outcome is independent of the other and the individual probabilities add up to $P(A) + P(B) = 1$. The restriction to two values may seem to limit practical application. Yet some very common variables have only two possible outcomes—for example, *client* or *no client*. Moreover, a variety of binomial situations consisting of an occurrence (A) or a non-occurrence (B) of a certain outcome are pertinent for business:

- "What is the probability of a purchase being made or not?"
- "What is the probability that a machine breaks down or not?"
- "What is the probability that an employee is sick or healthy?"
- "What is the probability that a loan is cancelled?"
- "What is the probability that a customer takes notice of a certain advertisement or not?"

All these questions are relevant for business and economics, and all are about the probability of a variable with two possible values.

A common way to describe binomial distribution is the so-called urn model. Say we have an urn with ten black balls and five white balls. There are only two possible outcomes. The probability of drawing a white ball is $p = P(\text{ball} = \text{white}) = 5/15 = 1/3$. The probability of drawing a black ball is given by the difference between one and the probability of drawing a white ball: $q = P(\text{ball} = \text{black}) = 1 - 1/3 = 2/3$. We can see right away that we only need to know one of the probability values—p—since the second probability necessarily is given by the

complementary probability ($q = (1 - p)$). Since the drawings must be independent, the probabilities do not change for the second drawing: 1/3 for drawing a white ball and 2/3 for drawing a black ball. But this is only the case when we replace each ball after drawing it, so that there are always five white balls and ten black balls to draw from the urn.

The binomial distribution is frequently linked to "drawing with replacement (of that drawn)". This is the case with the urn model but it need not always apply. It is more accurate to think of binomial distribution as "drawing with constant probabilities".

What is the probability of picking two white balls and three black balls over five drawings, assuming that the probabilities remain constant ("drawing with replacement")? Let us first look at a certain sequence of outcomes: white, white, black, black, black. The probability of this sequence is:

$$P(\text{white}) \cdot P(\text{white}) \cdot P(\text{black}) \cdot P(\text{black}) \cdot P(\text{black}) = P(\text{white})^2 \cdot P(\text{black})^3 \tag{7.3}$$

This can be abbreviated as:

$$p^2 \cdot (1-p)^3 = \left(\frac{1}{3}\right)^2 \cdot \left(1 - \frac{1}{3}\right)^3 = 3.29\% \tag{7.4}$$

Two white balls and three black balls can be picked in other orders, such as black and white balls in alternation. Here, too, the probability remains the same:

$$P(\text{black}) \cdot P(\text{white}) \cdot P(\text{black}) \cdot P(\text{white}) \cdot P(\text{black}) = \left(\frac{1}{3}\right)^2 \cdot \left(1 - \frac{1}{3}\right)^3 = 3.29\% \tag{7.5}$$

Using combinatorial analysis and binomial coefficients we can calculate the number of possible combinations of two white balls and three black balls:

$$\binom{5}{2} = \frac{5!}{2!(5-2)!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{(1 \cdot 2)(1 \cdot 2 \cdot 3)} = 10 \tag{7.6}$$

$$B\left(n,k,p\right)=\binom{n}{k}\cdot p^{k}\cdot\left(1-p\right)^{n-k}=\frac{n!}{k!(n-k)!}\cdot p^{k}\cdot\left(1-p\right)^{n-k}$$

The probability of success (e.g. white ball) on an individual trial

Number of successes in the sample with a certain property (e.g. number of white balls)

Sample size

Fig. 7.2 Binomial distribution

This binomial coefficient—pronounced five choose two—yields a total of ten possibilities for the combination of drawing two white balls and three black balls, assuming each ball is replaced after drawing. This enables us to determine the total probability for a combination of two white balls and three black balls independent of their order:

$$\begin{aligned}\binom{5}{2}\cdot\left(\frac{1}{3}\right)^{2}\cdot\left(\frac{2}{3}\right)^{3}&=\frac{5!}{2!(5-2)!}\cdot\left(\frac{1}{3}\right)^{2}\cdot\left(\frac{2}{3}\right)^{3}\\&=\frac{1\cdot2\cdot3\cdot4\cdot5}{(1\cdot2)(1\cdot2\cdot3)}\cdot\left(\frac{1}{3}\right)^{2}\cdot\left(\frac{2}{3}\right)^{3}=10\cdot3.29\%=32.9\%\end{aligned}\tag{7.7}$$

From this we can formulate the general formula for binomial distribution as shown in the Fig. 7.2.

Let us approach binomial distribution using another example: a company runs three different TV ads on three different days in a certain week. The ad company performed a media analysis and discovered that the probability of an individual watching one of the ads is 10%. What is the probability that a customer watches two of the three ads if the order doesn't matter?

The probability that a customer sees an ad is $p = 0.1$, while the probability that a customer does not see the ad results from the complementary probability: $q = (1 - p) = (1 - 0.1) = 0.9$. These probabilities remain constant over time and hence follow a binomial distribution. But this example also shows why the binomial distribution cannot always be thought of in terms of "drawing with replacement". For instance, if a person misses the ad that runs on Monday, he won't have a chance to see it again. Once the Monday ad airs, it is not replaced. But since the probability does not change for seeing the other ads, it still follows a binomial distribution. The probability for seeing two ads is:

$$\begin{aligned}B(3,2,0.1)&=\binom{3}{2}p^{x}\cdot(1-p)^{1-x}=\frac{n!}{x!(n-x)!}p^{x}\cdot(1-p)^{1-x}\\&=\frac{3!}{2!\cdot1!}\cdot0.1^{2}\cdot0.9^{1}=2.7\%.\end{aligned}\tag{7.8}$$

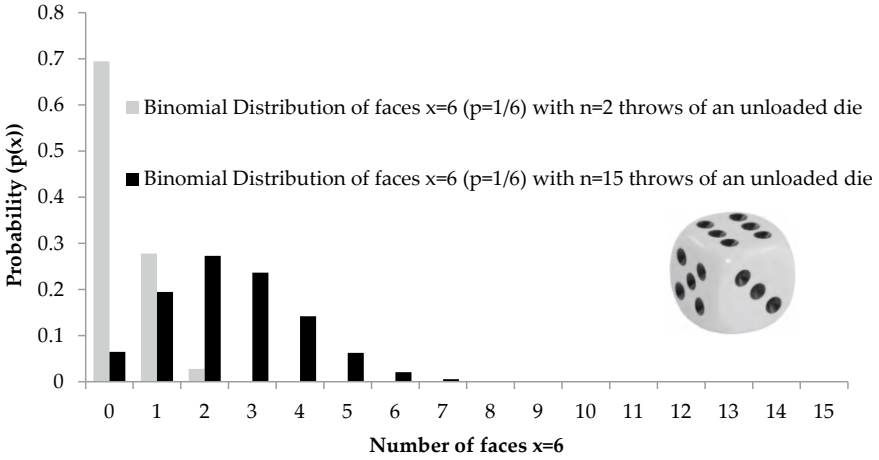


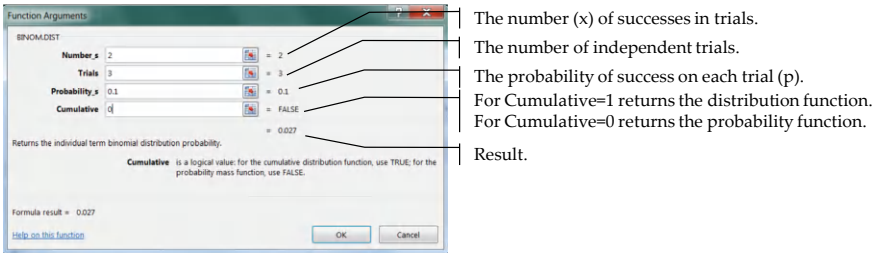
Fig. 7.3 Binomial distribution of faces $x = 6$ with n throws of an unloaded die

In brief, the characteristics of a binomial distribution are:

1. The shape of a binomial distribution can be symmetric or skewed. The distribution is symmetric whenever $p = 0.5$, regardless the size of the sample (n). However, when $p < 0.5$ ($p > 0.5$), the binomial distribution tends to be right-skewed (left-skewed).The larger the sample size n, the less skewed the binomial distribution will be (see Fig. 7.3).
2. The expected value of a random variable in a binomial distribution is $E(X) = n \cdot p$. For the TV ad example, this yields: $E(X) = 3 \cdot 0.1 = 0.3$. This means that for three draws we can expect 0.3 ad views.
3. The variance of a random variable in a binomial distribution is $\text{Var}(X) = n \cdot p \cdot (1 - p)$. For the TV ad example, this yields $\text{Var}(X) = 3 \cdot 0.1 \cdot 0.9 = 0.27$.
4. Random variables in a binomial distribution are reproductive, which means that the merging of two (or more) random variables in a binomial distribution leads to another random variable in a binomial distribution.
5. The probability of random variables in a binomial distribution can be calculated using other distributions under certain circumstances. Random variables in binomial distributions approximately follow a
   (a) … normal distribution ($N(n \cdot p; \sqrt{n \cdot p(1-p)})$), if $n \cdot p \cdot (1 - p) > 9$.
   (b) … Poisson distribution ($\text{Po}(n \cdot p)$), if $n \cdot p \leq 10$ and $n \geq 1500 \cdot p$.

7.1.1.1 Calculating Binomial Distributions Using Excel

Using Excel, we can calculate the probabilities of a binomial distribution. Under *Formulas* select *Insert Function.* Then select the function *BINOM.DIST* from under the category *Statistical*. As the Fig. 7.4 shows, next indicate the parameters "The



Syntax: BINOM.DIST (number_s; trials; probability_s; cumulative); BINOM.DIST (2;3;0.1;0)

Fig. 7.4 Calculating binomial distributions with Excel

number of successes in trials" (Number_s = x), "The number of independent trials" (Trials = n), "The probability of success on each trial" Probability_s = p), and "A logical value that determines the form of the function". For Cumulative = 1 *BINOM.DIST* returns the distribution function; for Cumulative = 0 *BINOM.DIST* returns the probability function.

7.1.1.2 Calculating Binomial Distributions Using Stata

In Stata, enter *display binomialp(n,k,p)* to calculate the probability of k successes in n trials for the individual probability $p(B(n,k,p))$. In a similar fashion, you can calculate the value of the binomial distribution function for a maximum of k successes by selecting *display binomial(n,k,p)* and for a minimum of k successes by selecting *display binomialtail(n,k,p)*.[1] Figure 7.5 provides an overview of Stata commands for binomial distributions.

7.1.2 Hypergeometric Distribution

Like the binomial distribution, the hypergeometric distribution consists of only two values (A and B) that together constitute the entire probability space, so that $A \cup B = \Omega$. Unlike the binomial distribution, the probability of an outcome in a hypergeometric distribution depends on the number of experiments that have been performed. Hence, the probability that a value occurs changes with each experiment.

To explain hypergeometric distribution, let us once again use a two-coloured urn model. Say we have an urn with five red balls and three green balls. The probability of drawing a red ball first is $P(\text{ball} = \text{red}) = 5/8$. The probability of drawing a green ball can be calculated from the complementary probability, in this case: $P(\text{ball} = \text{green}) = 1 - 5/8 = 3/8$. The ball we draw is not replaced, leaving only seven balls in the urn. The probability of drawing a red or a green ball

[1] A very good explanation of how to calculate binomial distributions using Stata can be found on the *Stata Learner* YouTube channel (see: https://www.youtube.com/watch?v=4O641mAAf_I).

Case	Syntax
Returns the probability of observing k successes in n trials when the probability of a success on one trial is p: B(n,k=x,p)	display binomialp(n,k,p)
Ex: Returns the probability of observing two successes in three trials when the probability of a success on one trial is p=0.1: B(3,k=2,0.1)	display binomialp(3,2,0.1) .027
Returns the probability of observing k or fewer successes in n trials when the probability of a success on one trial is p: B(n,k≤x,p)	display binomial(*n,k,p*)
Ex: Returns the probability of observing two or fewer successes in three trials when the probability of a success on one trial is p=0.1: B(3,k≤2,0.1)	display binomial(3,2,0.1) .999
Returns the probability of observing k or more successes in n trials when the probability of a success on one trial is p: B(n,k≥x,p)	display binomialtail(*n,k,p*)
Ex: Returns the probability of observing two or more successes in three trials when the probability of a success on one trial is p=0.1: B(3,k≥2,0.1)	display binomial-tail(3,2,0.1) .028

Fig. 7.5 Calculating binomial distributions using Stata

now depends on which colour was drawn first. If we draw the red ball first, then four of the remaining balls are red, so that the probability of drawing a red ball is $P_2(\text{ball} = \text{red}) = 4/7$ and the probability of drawing a green ball is $P_2(\text{ball} = \text{green}) = 3/7$. By contrast, if we draw a green ball first, then five of the remaining balls are red, so that the probability of drawing a red ball is $P_2(\text{ball} = \text{red}) = 5/7$ and the probability of drawing a green ball is $P_2(\text{ball} = \text{green}) = 2/7$. The probability of drawing a red or green ball is obviously dependent on previous drawings. With this mind, what is the probability of drawing a red ball and a green ball in two consecutive drawings without replacement? To answer this question, we must answer three questions:

- What is the total number of unique possible outcomes when drawing a sample of two out of eight balls? The binomial distribution yields a total of 28 possible (unique) events:

$$\binom{N}{n} = \binom{8}{2} = \frac{8!}{2!(8-2)!} = 28 \tag{7.9}$$

- What is the total number of all unique possibilities of drawing one (of five) red balls, regardless whether it is drawn in the first or second turn? The binomial distribution yields a total of five possibilities:

$$\binom{M}{x} = \binom{5}{1} = \frac{5!}{1!(5-1)!} = 5. \tag{7.10}$$

The number of unique possibilities of drawing one (of three) green balls results from the complementary probability:

$$\binom{N-M}{n-x} = \binom{8-5}{2-1} = \binom{3}{1} = \frac{3!}{1!(3-1)!} = 3 \text{ possibilities} \tag{7.11}$$

The probability is given by the quotient of:

$$H(8,5,2,1) = \frac{\begin{pmatrix}\text{Number of events with}\\ 1 \text{ (of 5) red ball}\end{pmatrix} \cdot \begin{pmatrix}\text{Number of events with}\\ 1 \text{ (of 3) green ball}\end{pmatrix}}{\begin{pmatrix}\text{Number of unique outcomes when drawing}\\ \text{a sample of two out of eight balls}\end{pmatrix}} \tag{7.12}$$

$$H(8,5,2,1) = \frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}} = \frac{\binom{5}{1} \cdot \binom{8-5}{2-1}}{\binom{8}{2}}$$

$$= \frac{\frac{5!}{1!(5-1)!} \cdot \frac{3!}{1!(3-1)!}}{\frac{8!}{2!(8-2)!}} = \frac{5 \cdot 3}{28} = 0.54 = 54\% \tag{7.13}$$

From this we can formulate the formal association for hypergeometric distributions as shown in Fig. 7.6.

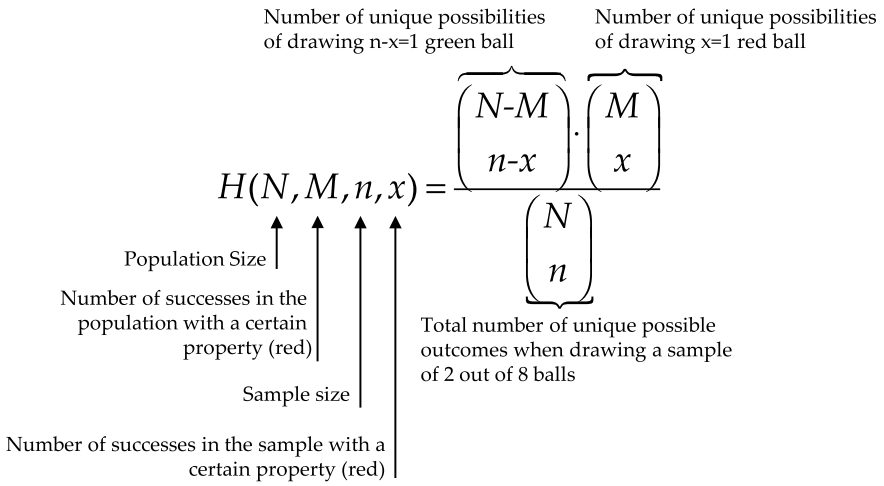


Fig. 7.6 Hypergeometric distribution

A typical example of hypergeometric distributions is the lottery drawing. For instance, what is the probability of holding three right numbers or six right numbers in Lotto 6/49? Since lottery balls are drawn without being replaced, each drawing necessarily changes the probability of another ball being drawn. Here are the probabilities for drawing three right numbers and six right numbers:

$$H(49,6,6,3) = \frac{\binom{49-6}{6-3}\binom{6}{3}}{\binom{49}{6}} = \frac{\frac{43!}{3!(43-3)!} \cdot \frac{6!}{3!(6-3)!}}{\frac{49!}{6!(49-6)!}}$$
$$= \frac{12{,}341 \cdot 20}{13{,}983{,}816} = 0.0177 = 1.77\% \tag{7.14}$$

$$H(49,6,6,6) = \frac{\binom{49-6}{6-6}\binom{6}{6}}{\binom{49}{6}} = \frac{\frac{43!}{0!(43-0)!} \cdot \frac{6!}{6!(6-6)!}}{\frac{49!}{6!(49-6)!}}$$
$$= \frac{1 \cdot 1}{13{,}983{,}816} = 0.00000715\% \tag{7.15}$$

The characteristics of the hypergeometric distribution are:

First, the expected value of a random variable in a hypergeometric distribution is defined as $E(X) = n\frac{M}{N}$. Accordingly, the expected value for drawing two red balls is

$$E(X) = 2 \cdot \frac{5}{8} = 1.25. \tag{7.16}$$

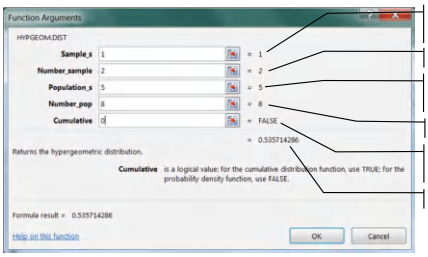
Second, the variance of a random variable in a hypergeometric distribution is defined by

$$\mathrm{Var}(X) = \frac{n(N-n)}{N-1} \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right). \tag{7.17}$$

This gives us the following variance for the above example:

$$\mathrm{Var}(X) = \frac{n(N-n)}{N-1} \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right) = \frac{2 \cdot (8-2)}{8-1} \cdot \frac{5}{8} \cdot \left(1 - \frac{5}{8}\right) = 0.402 \tag{7.18}$$

Third, the probability of random variables in a hypergeometric distribution can be calculated in certain circumstances by other distributions. This is especially the case with large total populations, as here the probabilities of drawing a certain item change little in subsequent drawings despite there being no replacement. Imagine we were picking people at random from the German population of approximately



Syntax: HYPGEOM.DIST(sample_s;number_sample_s;population_s;number_pop;cumulative); HYPGEOM.DIST(1;2;5;8)

Fig. 7.7 Calculating hypergeometric distributions with Excel

80,000,000 inhabitants. With the first drawing, the possibility of selecting a person is 1/80,000,000. The probability of selecting a person from the remaining population is marginally less: 1/79,999,999. Random variables in a hypergeometric distribution ($H(N;M;n)$) approximately follow

- … a normal distribution $\left(N\left(\underbrace{n\frac{M}{N}}_{E(x)}; \underbrace{\sqrt{\frac{n(N-n)}{N-1}\cdot\frac{M}{N}\cdot\left(1-\frac{M}{N}\right)}}_{\sigma} \right) \right)$, if $n > 30$ and $0.1 < \frac{M}{N} < 0.9$,
- … a Poisson distribution $\left(\text{Po}\left(\underbrace{n\frac{M}{N}}_{E(x)} \right) \right)$, if $0.1 \geq \frac{M}{N}$ or $\frac{M}{N} \geq 0.9$ and $n > 30$ and $\frac{n}{N} < 0.05$,
- … a binomial distribution $\left(B\left(n; \frac{M}{N}\right)\right)$, if $n > 10$ and $0.1 < \frac{M}{N} < 0.9$ and $\frac{n}{N} < 0.05$.

7.1.2.1 Calculating Hypergeometric Distributions Using Excel

Using Excel, we can calculate the probabilities of a hypergeometric distribution. Under *Formulas* select *Insert Function.* Then select the function *HYPGEOM.DIST* from under the selected category *Statistical*. As the Fig. 7.7 shows, next indicate the parameters for x (the number of successes in the sample: *Sample_s*), for n (the size of the sample: *Number_sample*), for M (the number of successes in the population with a certain property: *Population_s*), and the population size N (*Number_pop*).

7.1.2.2 Calculating the Hypergeometric Distribution Using Stata

Assume you want to calculate the hypergeometric distribution for $x = k$ successes of a specific attribute in a sample with a size n, and the attribute occurs M times in a population N. To do so, enter *display hypergeometricp(N,M,n,x)* along with the values for N, M, n and $x = k$ in the command line. Then calculate the value of

Case	Syntax
Returns the hypergeometric probability of x successes (where success is obtaining an element with the attribute of interest) out of a sample of size n, from a population of size N containing M elements that have the attribute of interest: H(N,M,n,x = k)	display hypergeometricp(N,M,n,x)
Ex: Returns the hypergeometric probability of x=1 successes out of a sample of size n=2, from a population of size N=8 containing M=5 elements that have the attribute of interest: H(8,5,2,x=1)	display hypergeometricp(8,5,2,1) .536
Returns the cumulative hypergeometric probability of x or less successes (where success is obtaining an element with the attribute of interest) out of a sample of size n, from a population of size N containing M elements that have the attribute of interest: H(N,M,n,x ≤ k)	display hypergeometric(N,M,n,x)
Ex: Returns the cumulative hypergeometric probability of one or less successes (x≤1) out of a sample of size n=2, from a population of size N=8 containing M=5 elements that have the attribute of interest: H(8,5,2,x ≤ 1)	display hypergeometric(8,5,2,1) .643
Returns the hypergeometric probability of more than x successes (where success is obtaining an element with the attribute of interest) out of a sample of size n, from a population of size N containing M elements that have the attribute of interest: H(N,M,n,x > k)	display 1-hypergeometric(N,M,n,x)
Ex: Returns the hypergeometric probability of more than one successes (x>1) out of a sample of size n=2, from a population of size N=8 containing M=5 elements that have the attribute of interest: H(8,5,2,x > 1)	display 1-hypergeometric(8,5,2,1) .357

Fig. 7.8 Calculating hypergeometric distributions using Stata

a hypergeometric distribution function $H(N,M,n,x)$ for a maximum of k successes ($x \leq k$) by entering *display hypergeometric(N,M,n,k)* and for more than k ($x > k$) successes by entering *display 1-hypergeometric(N,M,n,k)*. Figure 7.8 provides an overview of Stata commands for hypergeometric distributions.

7.1.3 The Poisson Distribution

The final discrete probability distribution we will describe here involves a Bernoulli trial, i.e. whether a certain event occurs or not, and hence also derives from the binomial distribution. The Poisson distribution, named after the French physicist and mathematician Siméon Denis Poisson (1781–1840), is frequently used to estimate whether a certain event occurs within a certain period of time,

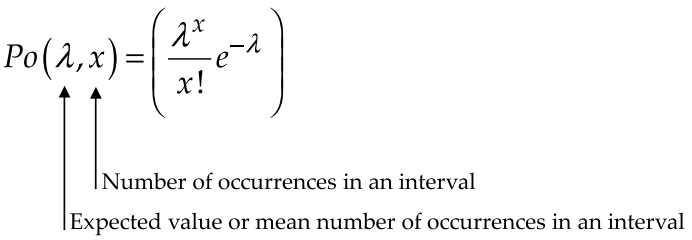


Fig. 7.9 Poisson distribution

such as the number of machine failures or defective products in a day or the number of vehicles that are sold in an hour. The Poisson distribution assumes that events are independent of each other. For instance, the probability of a machine failure in one time span does not change the probability of a machine failure in the second.

Usually, a variable is considered to follow a Poisson distribution if the probability that a certain event occurs is relatively low. In this case, the parameter p of the binomial distribution is relatively small, while parameter n—the size of the sample—tends to be large. Here we can show that a random variable with a binomial distribution obeys the following formula:

$$B(n, k = x, p) = \lim_{n \to \infty} \binom{n}{x} p^x (1-p)^{n-x} = \frac{\mu^x}{x!} e^{-\mu} \tag{7.19}$$

As illustrated in Fig. 7.9, this is also the formula for calculating Poisson random variables.

A special property of the Poisson distribution is that the distribution can be described with a single parameter λ, which equals both the expected value μ and the variance of the Poisson distribution:

$$E(X) = \mu_X = \sigma_X^2 = \lambda \tag{7.20}$$

The reproductive property is another feature of the Poisson distribution. The sum of n stochastic independent and Poisson distributed random variables with the parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$ results in a Poisson distribution with $\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_n$.

Here's an example to illustrate these two features. Four streets meet at an intersection. The numbers of vehicles that cross the intersection are stochastically independent and follow a Poisson distribution. From experience, we know that the average number of vehicles that pass the intersection per day on each street is 2.0, 1.6, 3.4, and 3.0, respectively. What is the probability that on a given day exactly nine vehicles cross the intersection, and what is the probability on a given day that no more than nine vehicles cross the intersection?

The expected values of the Poisson distributions correspond to the parameters for $\mu_1 = \lambda_1 = 2, \ldots, \mu_4 = \lambda_4 = 3.4$. Due to the reproductive property, the sum

of individual Poisson distributions also results in a Poisson distribution. Hence, we can expect the average number vehicles to be:

$$\mu = \mu_1 + \mu_2 + \mu_3 + \mu_4 = 2 + 1.6 + 3.4 + 3 = 10 = \lambda \tag{7.21}$$

The probability that exactly nine vehicles drive past the intersection is:

$$\text{Po}(\lambda, x = 9) = \frac{\mu^x}{x!}\text{e}^{-\mu} = \frac{10^9}{9!}\text{e}^{-10} = 12.511\% \tag{7.22}$$

To calculate the probability of no more than nine vehicles passing the intersection, the individual probabilities of zero cars to nine cars must be calculated and added. A computer can help with the calculation.

As mentioned, the Poisson distribution with $\lambda = n \cdot p$ is derived from the binomial distribution, which itself can be approximated using the continuous normal distribution. When $\lambda \geq 10$, therefore, the Poisson distribution can also be approximated using the continuous normal distribution.[2]

7.1.3.1 Calculating the Poisson Distribution Using Excel

In Excel you can calculate the probabilities for the above example by going to *Formulas → Insert Functions* and selecting *POISSON.DIST*. Next enter the parameters for the number of cases ($x = 9$) and for the mean ($\lambda = 10$). The function *cumulative = 1*: *POISSON.DIST* calculates the value for the distribution function, while *cumulative = 0*: *POISSON.DIST* calculates the probability function, as can be seen in Fig. 7.10. Again, the probability of exactly nine vehicles passing the intersection is 12.51%. Accordingly, the probability of no more than nine vehicles passing the intersection each day is 45.79%.

7.1.3.2 Calculating the Poisson Distribution Using Stata

Figure 7.11 summarizes the calculations of the probability distribution and the distribution function using Stata. Enter *display poissonp(λ,k)* in the command line to calculate the probability for k successful outcomes with an expected value of λ. Calculate the value of the Poisson distribution function for no more than k successful outcomes by entering *display poisson(λ,k)* and for at least k successful outcomes by entering *display poissontail(λ,k)*.[3]

[2] When approximating a discrete distribution with a continuous distribution, the probability of a certain value x_i from a random variable is positive or zero in the discrete case, while for the continuous distribution it always equals zero. To determine the value for $P(X = 10)$ using approximation, we calculate the difference of the normal distribution function, so that $P(X = 10) = P(X \leq 10.5) - P(X \leq 9.5)$.

[3] A very good explanation of how to calculate Poisson distributions using Stata can be found on the *Stata Learner* YouTube channel (see: https://www.youtube.com/watch?v=R9a61ViJBwc).

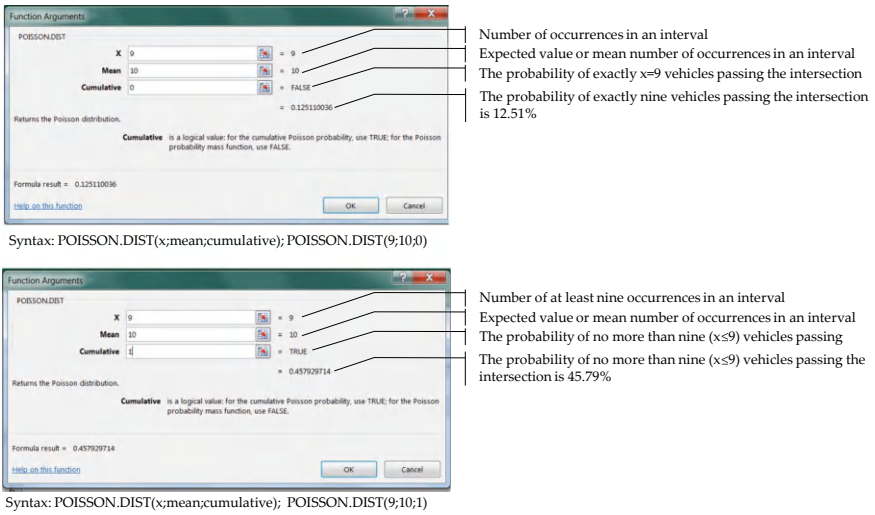


Fig. 7.10 Calculating the Poisson distribution with Excel

Case	Syntax
Returns the probability of observing x=k outcomes that are distributed as Poisson with a mean μ=λ: Po(λ,x=k)	display poissonp(λ,k)
Ex: Returns the probability of observing x=2 outcomes that are distributed as Poisson with a mean μ=10: Po(λ=10, x=2)	display poissonp(10,9) .125
Returns the probability of observing x=k or fewer outcomes that are distributed as Poisson with a mean μ=λ: Po(λ,x≤k)	display poisson(λ,k)
Ex: Returns the probability of observing x=2 or fewer outcomes that are distributed as Poisson with a mean μ=10: Po(λ=10, x≤2)	display poisson(10,9) .458
Returns the probability of observing x=k or more outcomes that are distributed as Poisson with a mean μ=λ: Po(λ,x≥k)	display poissontail(λ,k)
Ex: Returns the probability of observing x=2 or more outcomes that are distributed as Poisson with a mean μ=10: Po(λ=10, x≥2)	display poissontail(10,9) .667

Fig. 7.11 Calculating the Poisson distribution using Stata

7.2 Continuous Distributions

In the previous section, we learned about the probability distributions of discrete random variables. Discrete random experiments result in a finite number or a countably infinite number of values. Random experiments that can result in an

infinite number of values produce continuous random variables. Frequently, continuous random variables measure time, weight, or length such as when determining the probability of a certain shelf life of a product.

While probabilities for specific values require discrete probability functions, continuous probabilities rely on so-called density functions $f(x)$. The latter do not supply probabilities directly. Instead, an integral calculation must be performed to determine the probability of a specific interval. Figure 7.12 shows typical density functions.

The total area beneath each density function $f(x)$ equals

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1 = 100\% \tag{7.23}$$

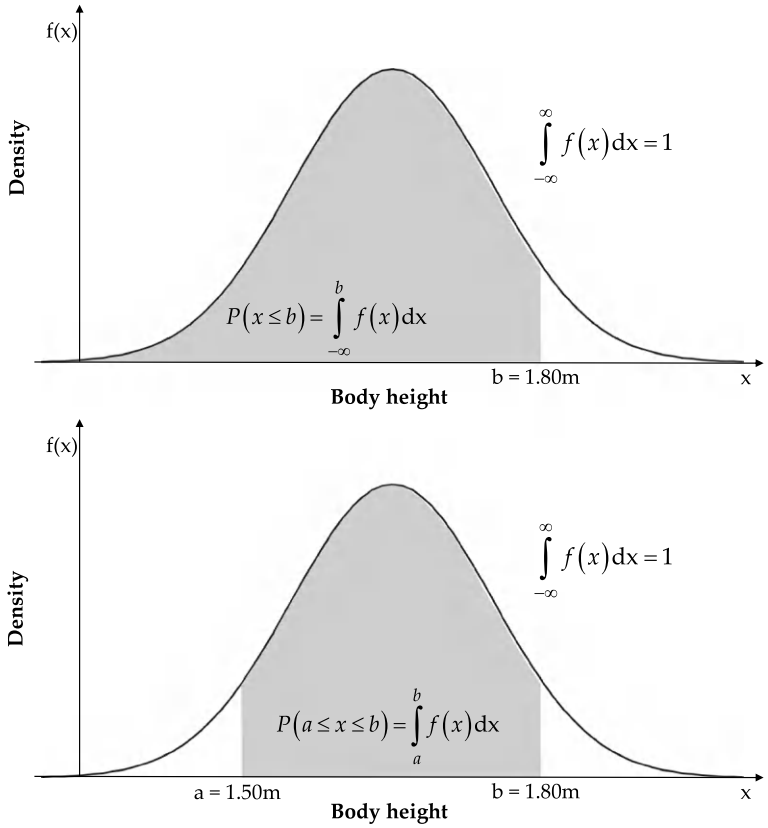


Fig. 7.12 Density functions

In other words, the integral from $-\infty$ to $+\infty$ always equals one in density functions.[4]

This is actually very intuitive. Say you are investing the body height of students. It is fairly obvious that 100% of the observations you record will lie between $-\infty$ to $+\infty$. Now say you want to find the probability that none of the students is taller than 1.80 m. The probability that a value of no more than $b = 1.80$ m occurs is represented by the gray area in the upper graph in Fig. 7.12 and is described by the formula

$$P(x \leq b = 1.80) = \int_{-\infty}^{b=1.80} f(x)\mathrm{d}x \tag{7.24}$$

Perhaps you also want to calculate the probability that a student will be between 1.50 m and 1.80 m tall. The integral of all values that lie between $a = 1.50$ m and $b = 1.80$ m are represented by the gray area in the lower graph in Fig. 7.12 and is described by the formula

$$P(a = 1.50 \leq x \leq b = 1.80) = \int_{a=1.50}^{b=1.80} f(x)\mathrm{d}x \tag{7.25}$$

The expected values and variance of discrete random variables are calculated through addition. The expected values and variance of continuous random variables are calculated by determining an integral. For the latter case, we use density function values, not probabilities, so that

$$E(X) = \mu_X = \int_{-\infty}^{\infty} x \cdot f(x)\mathrm{d}x \tag{7.26}$$

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 \cdot f(x)\mathrm{d}x \tag{7.27}$$

In the following sections we will take a closer look at the most important continuous distribution functions.

[4] In addition to the integral from $-\infty$ to $+\infty$ equaling 1, density functions must be real functions that are non-negative and integrable.

7.2.1 The Continuous Uniform Distribution

One of the simplest continuous distributions is the uniform distribution, sometimes also known as the rectangular distribution. Here the density functions have the constant value $\frac{1}{b-a}$ within the interval between some minimum a and some maximum b. The probability of occurrence is the same for all events:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a < X \leq b \\ 0 & \text{otherwise} \end{cases} \tag{7.28}$$

The integral of the density function between the interval limits a and b equals 1, the required value for density functions:

$$P(a \leq x \leq b) = \int_a^b f(x)\mathrm{d}x = 1 \tag{7.29}$$

Let's illustrate this with a brief example. Say it takes at least 120 min and at most 140 min to manufacture a product. The formula for its uniform distribution is

$$f(x) = \begin{cases} \frac{1}{140-120} = \frac{1}{20} & \text{for } 120 \leq X \leq 140 \\ 0 & \text{otherwise} \end{cases} \tag{7.30}$$

Figure 7.13 provides a graph of the distribution.

The expected value—the average time it takes to manufacture a product—is

$$E(X) = \frac{a+b}{2} == \frac{120+140}{2} = 130 \text{ min} \tag{7.31}$$

The variance is

$$\mathrm{Var}(X) = \sigma^2 = \frac{(b-a)^2}{12} = \frac{(140-120)^2}{12} = \frac{400}{12} = 33.\overline{3} \text{ min}^2 \tag{7.32}$$

What is the probability that the manufacturing process takes.

(a) no more than 125 min?
(b) no less than 130 min?
(c) between 125 and 130 min?

The graph on the left in Fig. 7.14 shows the answer to part (a). The black rectangle represents the possible values between 120 and 125 min. Accordingly, height · width yields

$$P(a \leq x \leq c) = \int_a^c f(x)\mathrm{d}x = \frac{1}{(b-a)} \cdot (c-a) \tag{7.33}$$

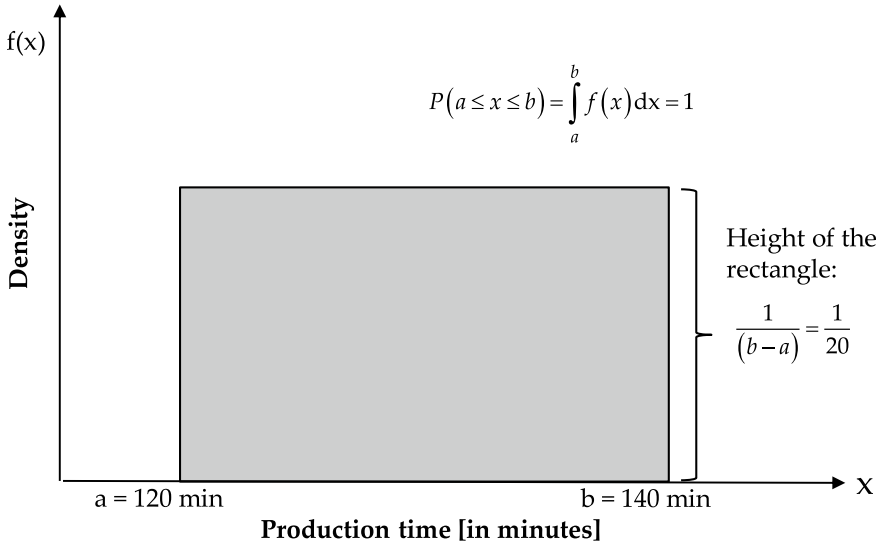


Fig. 7.13 Uniform distribution

so that the probability of the manufacturing process taking no more than 120 min is:

$$P(120 \le X \le 125) = \frac{1}{(140-120)} \cdot (125-120) = 25\% \tag{7.34}$$

We can arrive at the answer to part (b) in analogous fashion, as shown in the middle graph in Fig. 7.14:

$$P(d \le X \le b) = \int_d^b f(x)dx = \frac{1}{(b-a)} \cdot (b-d) = \frac{1}{(140-120)} \cdot (140-130) = 50\% \tag{7.35}$$

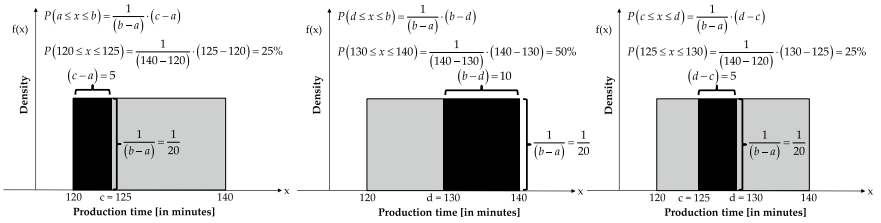


Fig. 7.14 Production times

The right graph from Fig. 7.14 shows the general formula for calculating probability from uniformly distribution density functions. Once again, the probability that the production time lies between 125 and 130 min results from multiplying the height and width of the black rectangle:

$$P(c \leq X \leq d) = \frac{1}{(b-a)} \cdot (d-c) \tag{7.36}$$

$$P(125 \leq X \leq 130) = \frac{1}{(140-120)} \cdot (130-125) = 25\% \tag{7.37}$$

7.2.2 The Normal Distribution

The most important theoretical distribution is certainly the normal distribution. The Huguenot Abraham de Moivre (1738), who earned his living giving advice to gamblers, described the formation of what would later be called a "normal distribution" in the second edition of his 1738 work *Doctrine of Chances*. The mathematician Pierre-Simon Laplace (1749–1827) and Carl Friedrich Gauss (1777–1855) used de Moivre's insights to develop the idea of an error probability curve for minimizing measurement error. In the eighteenth and nineteenth centuries, natural scientists often arrived at varying results in scientific experiments due to the inaccuracy of their measurement instruments. Measuring the distance of fixed stars, say, could produce quickly different individual results. But Laplace and Gauss observed that as more measurements were taken, the better the aggregated results became. Mapping all the results on a graph produces a symmetric, bell-shaped curve, whose mean represented the most frequent value, as shown in Fig. 7.15. This mean could then serve as the error-corrected value of, say, a star's actual distance.

Adolph-Lambert Quételet (1796–1874) was the first person to use the term "normal distribution". A fastidious and somewhat idiosyncratic thinker, Quételet noticed that the frequency distributions of vastly different measurements resembled the Laplace-Gauß curve. Whatever he measured—be it human body height or the chest measurements of Scottish soldiers—the data created the same distribution. Quételet thus surmised that the distribution was natural, hence the term "normal". Of course, there are many properties in nature that do not follow a normal distribution, and many result from the interaction of different measurements with normal distributions (Swoboda, 1971, p. 76).

Today, the normal distribution counts as one of the most important foundations of statistics. What does a normal distribution look like and what is so special about it? Figure 7.15 shows an ideal normal distribution. For more advanced readers, the density formula of the normal distribution is provided here as well:

$$f_x(x) = \frac{1}{\sigma\sqrt{2 \cdot \Pi}} \cdot \mathrm{e}^{-\left(\frac{(x-\mu)^2}{2 \cdot \sigma^2}\right)} \tag{7.38}$$

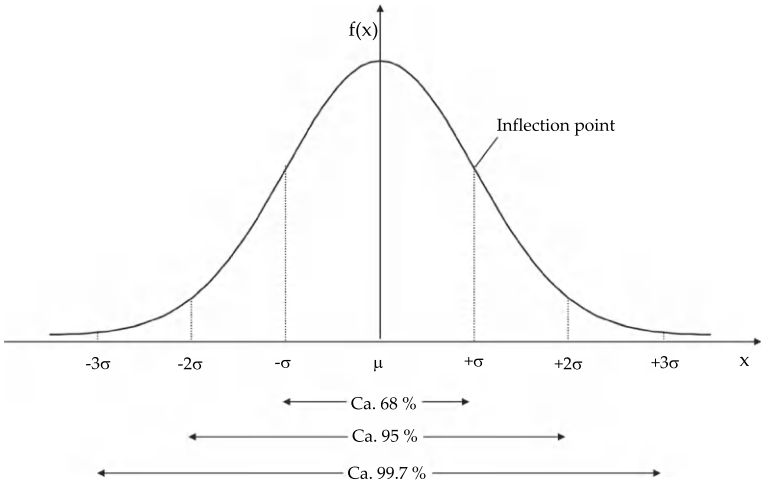


Fig. 7.15 Ideal density of a normal distribution

The normal distribution is shaped like a bell, with a single, symmetric peak. It can be fully described with only two parameters—the expected value μ and the standard deviation σ—regardless of the x-axis measurement values. This can be seen in the density formula of the normal distribution above. A random variable with a normal distribution can thus be given as:

$$X \sim N(\mu; \sigma) \tag{7.39}$$

The expected value represents the axis of symmetry and the most common value of the distribution and is identical with the median and the mean of the distribution. The expected value describes the position of a normal distribution on the x-axis. As shown in Fig. 7.16, the larger the expected value, the more the normal distribution shifts along the x-axis to the right.

The density of the curve falls off to the right and left, approaching the abscissa asymptotically. The size of the standard deviation of the normal distribution influences the shape of the normal distribution. As Fig. 7.17 shows, the larger the value of the standard deviation, "the flatter the bell".

The inflection point of the normal distribution is located at a distance of one standard deviation (σ) from μ. A tangent drawn through the inflection points will intersect the normal distribution at a distance of two standard deviations ($2 \cdot \sigma$). Approximately 68% of observations lie within the interval between $-\sigma$ and $+\sigma$; approximately 95% lie between -2σ and $+2\sigma$; and approximately 99.7% lie between -3σ and $+3\sigma$. This is illustrated in Fig. 7.15.

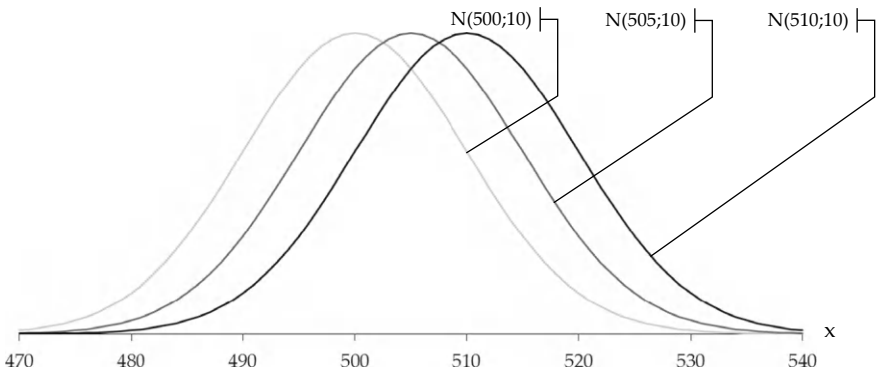


Fig. 7.16 Positions of normal distributions

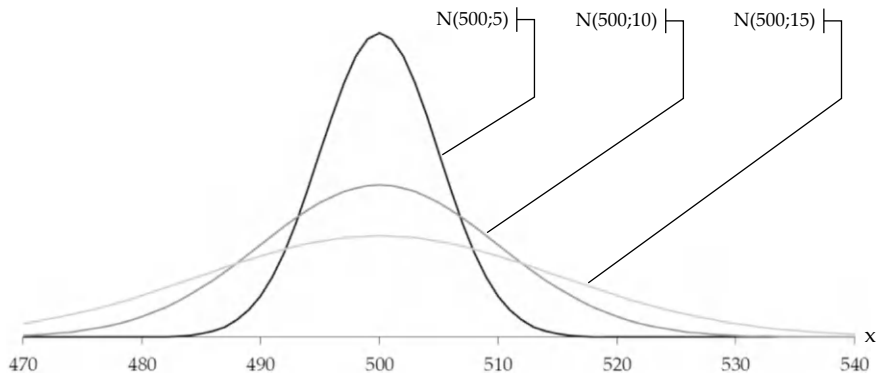


Fig. 7.17 Different spreads of normal distributions

The reproductive property is another particularity of the normal distribution: If two random variables X_1 and X_2 with the same or different expected values and spreads are normally distributed with $N(\mu_1; \sigma_1)$ and $N(\mu_2; \sigma_2)$, then the random variable $X = X_1 + X_2$ is also normally distributed with:

$$N\left(\mu_1 + \mu_2; \sqrt{\sigma_1^2 + \sigma_2^2}\right) \tag{7.40}$$

Another property of the normal distribution is that the area between the x-axis and the normal distribution always equals one (= 100%), regardless whether it is narrow, broad, or flat. The distribution is, therefore, especially well-suited for determining probabilities.

In empirical reality, the ideal and theoretical normal distribution is never 100% satisfied. But it is often justified as an approximation of actual values. As the physicist Gabriel Lippmann is said to have observed: Everyone believes in the normal

law, the experimenters because they imagine that it is a mathematical theorem and the mathematicians because they think it is an experimental fact (Swoboda, 1971, p. 80).

Of all the types of normal distributions, the one with an expected value of $\mu = 0$ and a standard deviation of $\sigma = 1$ has acquired special importance. It is known as the *standard normal distribution*. Before computers, statisticians relied on this distribution's "standard" values (like those tabulated in Appendix A) for making the complicated calculations associated with normal distributions easier. They discovered that any normal distribution could be transformed into a standard normal distribution by subtracting the expected value from every value of a random variable X and dividing it by the standard deviation. This procedure is known as the z-transformation. Its formula looks like this:

$$Z = \frac{X - \mu}{\sigma} \tag{7.41}$$

where $X \sim N(\mu; \sigma)$ yields

$$Z = \frac{X - \mu}{\sigma} \sim N(0; 1). \tag{7.42}$$

Each normal random variable X thus corresponds with a standard normal distribution random variable Z to help calculate probabilities.

We want to illustrate the usefulness of the z-transformation with an example. Say the shelf life of yoghurt is distributed with $N(30;10)$. That is to say, the average shelf life is 30 days with a standard deviation of ten days. What is the probability that the actual shelf life is no more than 50 days? What we need is the value for $P(X \leq 50)$, which corresponds to the grey area in the upper graphic in Fig. 7.18.

Unfortunately, the value for 50 cannot be found in the table in Appendix A. Hence the normal distribution $N(30;10)$ must undergo a z-transformation. The upper limit of the grey area of the interval in the upper part of Fig. 7.18 must be transferred to a random variable distributed with $N(0;1)$ so that the surface remains identical. Accordingly, a value of 50 for a random variable distributed with $N(30;10)$ yields a value of two for a standard normal distributed random variable (see the lower graph in Fig. 7.18). The calculation is as follows:

$$\begin{aligned} P(X \leq 50) &\Rightarrow P(X - \mu \leq 50 - \mu) \Rightarrow P\left(\frac{X - \mu}{\sigma} \leq \frac{50 - \mu}{\sigma}\right) \\ &\Rightarrow P\left(Z \leq \frac{50 - 30}{10}\right) \Rightarrow P(Z \leq 2) \end{aligned} \tag{7.43}$$

The values for $P(Z \leq 2)$ can be taken from the table of standard normal distribution values in Appendix A. The value of $P(Z \leq 2) = 0.9772$ corresponds to that of row 2.0 and column 0.00.

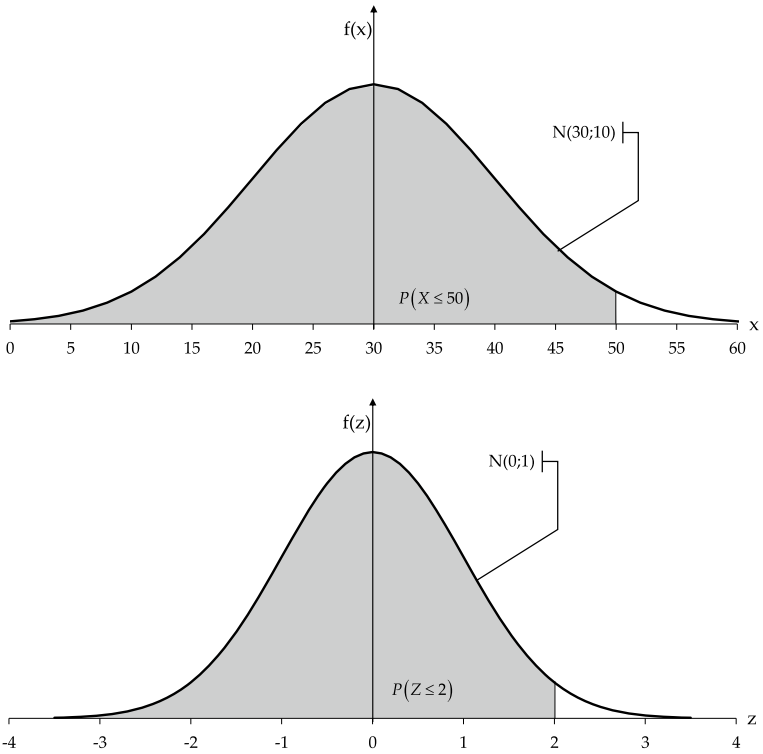


Fig. 7.18 Shelf life of yoghurt (1)

We can also determine the probabilities of an interval with fixed upper and lower limits:

$$P\left(X_{\text{lower}} \leq X \leq X_{\text{upper}}\right) \Leftrightarrow P\left(X_{\text{lower}} - \mu \leq X - \mu \leq X_{\text{upper}} - \mu\right) \tag{7.44}$$

$$\Leftrightarrow P\left(\frac{X_{\text{lower}} - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{X_{\text{upper}} - \mu}{\sigma}\right) \tag{7.45}$$

$$\Leftrightarrow P\left(\frac{X_{\text{lower}} - \mu}{\sigma} \leq Z \leq \frac{X_{\text{upper}} - \mu}{\sigma}\right) \tag{7.46}$$

Let us assume again that the shelf life of yoghurt is distributed with $N(30;10)$. What is the probability that the actual shelf life of the product is between 30 and 40 days? We are looking for the value of $P(30 \leq X \leq 40)$, which corresponds to the grey area in the upper graph in Fig. 7.19.

The z-transformation converts the defined limits into a random variable distributed with $N(0;1)$ over an identical area. This is shown in the lower graph in

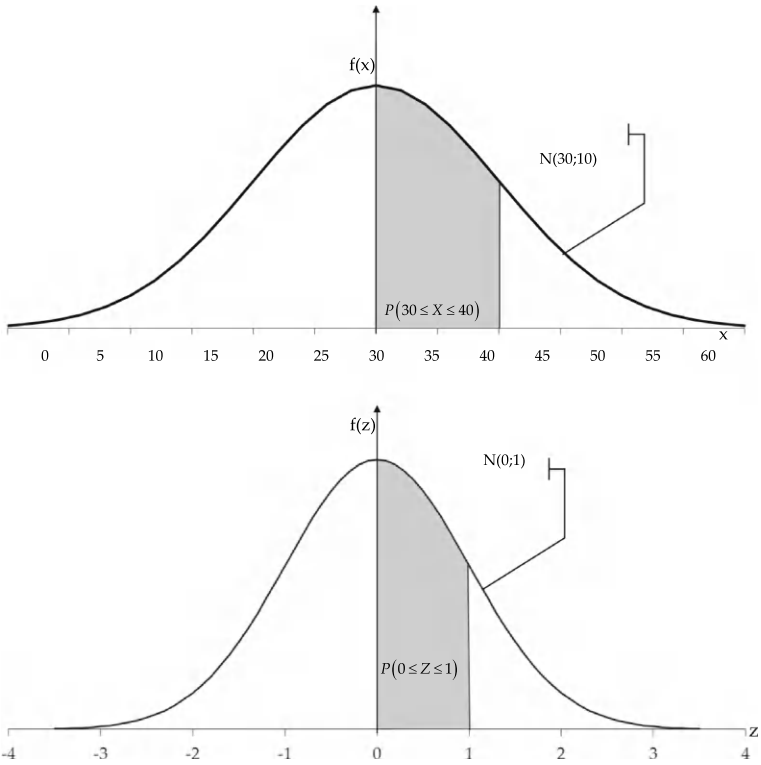


Fig. 7.19 Shelf life of yoghurt (2)

Fig. 7.19. The calculation is as follows:

$$P\left(\frac{X_{\text{lower}}-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{X_{\text{upper}}-\mu}{\sigma}\right) \Rightarrow P\left(\frac{30-30}{10} \leq Z \leq \frac{40-30}{10}\right)$$
$$\Rightarrow P(0 \leq Z \leq 1) \tag{7.47}$$

The probability we are looking for corresponds to the entire area of the standard normal distribution of $-\infty$ to the upper interval limit $P(Z < 1)$ (see the upper graph in Fig. 7.20) minus the area of $-\infty$ to the lower interval limit $P(Z < 0)$ (see the middle graph in Fig. 7.20). The probability is thus:

$$P(0 \leq Z \leq 1) = P(Z \leq 1) - P(Z < 0) \tag{7.48}$$

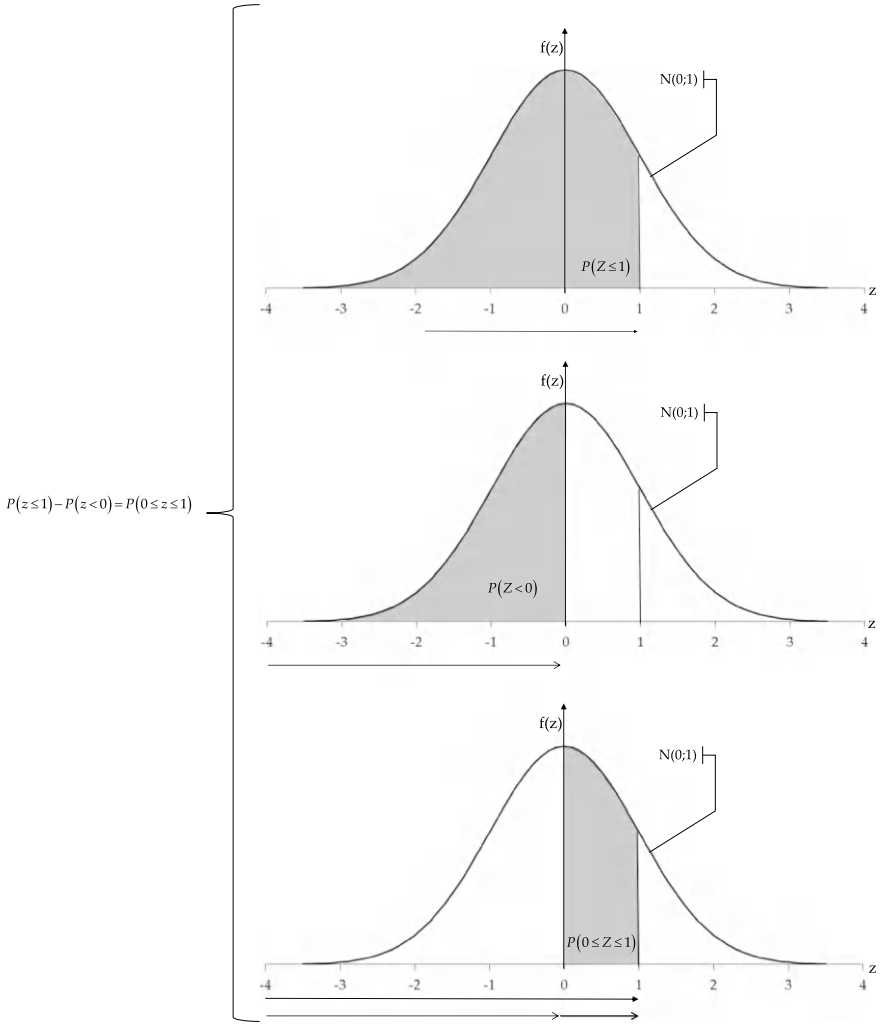


Fig. 7.20 Calculating the probability of a z-transformed random variable

The values for $P(Z \leq 1)$ and $P(Z < 0)$ can be again taken from the table of standard normal distribution in Appendix A. The value for $P(Z < 0) = 0.5000$ corresponds to that of row 0.0 (for the value zero and the first decimal place zero) and the column 0.00 (for the second decimal place zero). The value for $P(Z \leq 1) = 0.8413$ corresponds to that of row 1.0 and the column 0.00. This produces a probability of 34.13%:

$$P(0 \leq Z \leq 1) = P(Z \leq 1) - P(Z < 0) = 0.8413 - 0.5000 = 34.13\% \quad (7.49)$$

Figure 7.21 shows more examples of probability calculations using the table of standard normal distribution values.

The last example in Fig. 7.21 shows that due to the symmetry of the standard normal distribution the area to the right of a positive x-value (e.g. to the right of 2.5) is just as large as the area to the left of same x-value with a negative sign (e.g. to the left of $-$ 2.5). Thus:

- $P(Z \geq 1.65) = P(Z \leq (-1.65)) \approx 0.05$
- $P(Z \geq 1.96) = P(Z \leq (-1.96)) \approx 0.025$
- $P(Z \geq 2.58) = P(Z \leq (-2.58)) \approx 0.005$
- $P(Z \geq 3.00) = P(Z \leq (-3.00)) \approx 0.001$.

The lifetime of a lightbulb is normally distributed with a mean of 900 hours and a standard deviation of 100 hours. The lifetime is thus N(900;100) distributed. Determine the probabilities of lifetimes…

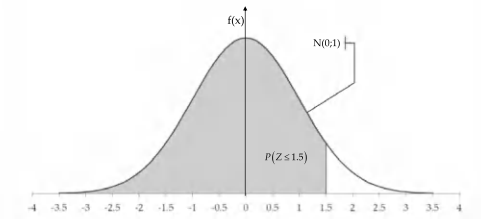
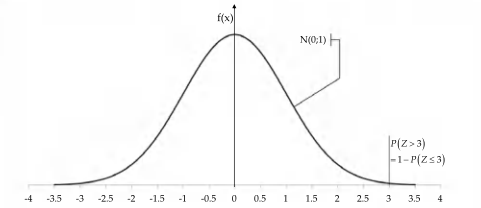
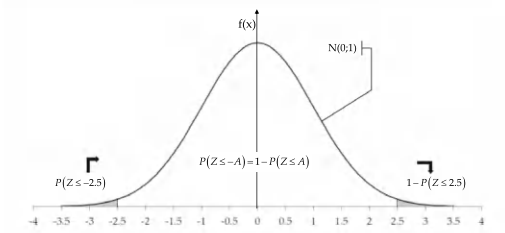
…of no more than 1050 hours	$P\left(Z \leq \frac{X_{upper} - \mu}{\sigma}\right)$ $\Rightarrow P\left(z \leq \frac{1050 - 900}{100}\right)$ $\Rightarrow P(Z \leq 1.5) = 93.32\%$	
…of more than 1200 hours	$P\left(Z > \frac{X_{upper} - \mu}{\sigma}\right)$ $\Rightarrow 1 - P\left(Z \leq \frac{X_{upper} - \mu}{\sigma}\right)$ $\Rightarrow 1 - P\left(z \leq \frac{1200 - 900}{100}\right)$ $\Rightarrow 1 - P(Z \leq 3) = 0.13\%$	
…of less than 650 hours	$P\left(Z \leq \frac{X_{upper} - \mu}{\sigma}\right)$ $\Rightarrow P\left(Z \leq \frac{650 - 900}{100}\right)$ $\Rightarrow P(Z \leq -2.5)$ $\Rightarrow 1 - P(Z \leq 2.5) = 0.62\%$	

Fig. 7.21 Calculating probabilities using the standard normal distribution

You can find very good explanations of normal distributions on the on the *Stata Learner* YouTube channel at:

- https://www.youtube.com/watch?v=9la6r8EuuCw
- https://www.youtube.com/watch?v=y7sISN6A3EM.

7.2.2.1 Calculating the Normal Distribution Using Excel

To calculate the probability values in Excel (Fig. 7.22) go to *Formulas* → *Insert Functions* and select *NORM.DIST*. Inserting the x-value, the expected value, the standard deviation, and the value 1 for the parameter *cumulative* yields the cumulative probability for p. For the given value of $x = 2.33$, an expected value of $\mu = 0$, and a standard deviation of $\sigma = 1$, the cumulative probability is $p \approx 0.99$. Use the function *NORM.DIST(2.33;0;1)* to make the calculation.

Conversely, the function *NORM.INV* calculates the x-value of a cumulative normal distribution with a given expected value and a given standard deviation. For example, a probability of $p = 0.9$, $\mu = 0$, and $\sigma = 1$ produces the x-value of $x \approx 1.28$ (= *NORM.INV(0.9;0;1)*).

A two-sided closed interval with the area for α equally divided on both sides of the normal distribution generates an x-value of $x_{1-\frac{\alpha}{2}}$ for the upper limit and an x-value of $x_{\frac{\alpha}{2}}$ for the lower limit. Using *NORM.INV(0.95;0;1),* the above example of a probability of $p = 0.9$, $\mu = 0$, and $\sigma = 1$ produces an x-value of $x_{95\%} \approx 1.65$; using *NORM.INV(0.05;0;1)*, it produces an x-value of $x_{5\%} \approx -1.65$.

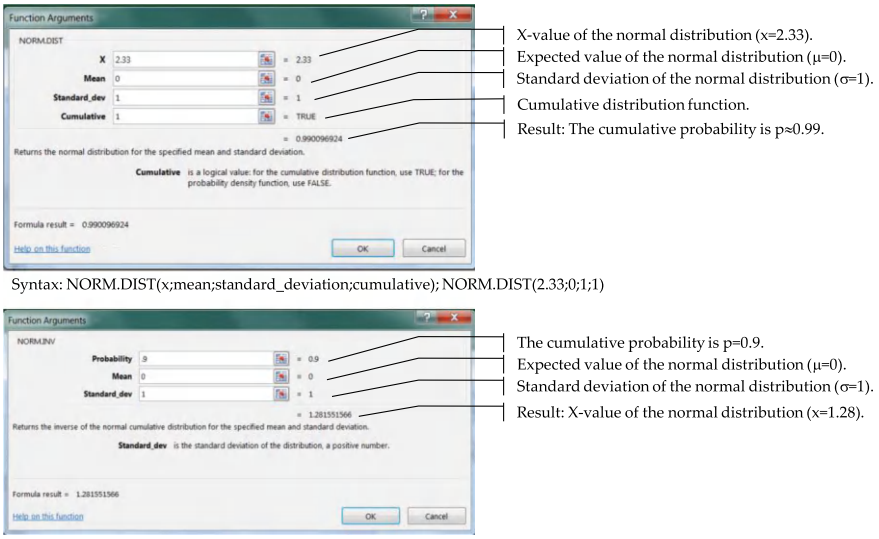


Fig. 7.22 Calculating the normal distribution using Excel

Case	Syntax
Display normal(z) returns the probability p for a given z-value of a standard normal distribution.	display normal(z)
Ex.: *display normal(1.96)* returns the cumulative standard normal distribution for z=1.96: p=0.975.	display normal(1.96) . 9750021
Ex.: Given a normal distribution N(3;4), *display normal((7-3)/4)* returns the cumulative standard normal distribution for x=7: p≈0.841.	display normal((7-3)/4) .84134475
Display invnormal(p) returns the z-value by assuming the cumulative probability p of a standard normal distribution.	display invnormal(p)
Ex.: Given a standard normal distribution (N(0;1)), *display invnormal(0.9)* returns a z-value of z=1.28.	display invnormal(0.9) 1.2815516
Ex.: Given a normal distribution (N(3;4)), *display 3+invnormal(0.9)*4* returns an x-value of x≈8.13.	display 3+invnormal(0.9)*4 8.1262063.

Fig. 7.23 Calculating the normal distribution using Stata

7.2.2.2 Calculating the Normal Distribution Using Stata

The command *display normal(z)* calculates the probability p with a given z-value of a standard normal distribution. For example, *display normal(1.96)* produces the value $p = 0.975$. If the normal distribution is not a standard normal distribution, standardize the x-value and select *display normal*$((x-\mu)/\sigma)$. The command *display normal((7–3)/4)* produces a probability of $p \approx 0.8413$ for $N(3;4)$ and the value $x = 7$.

The command *display invnormal(p)* calculates the z-value of a standard normal distribution by assuming the cumulative probability p. *Display invnormal(0.9)* produces a z-value of $z \approx 1.28$, as shown in Fig. 7.23.[5] For normal distributions with any μ and σ, use $\mu + z \cdot \sigma$ to transform the z-value into an x-value. For instance, the command *display 7 + invnormal(0.9)*4* calculates the value $x \approx 8.13$ of a random variable with a $N(3;4)$ distribution for the probability $p = 0.9$.

7.3 Important Distributions for Testing

Besides the normal distribution, three other theoretical distributions have achieved a certain renown. These are the t-distribution, the F-distribution, and the chi-squared distribution (or χ^2-distribution). Their popularity has less to do with their

[5] A very good explanation of how to calculate normal distributions using Stata can be found on the *Stata Learner* YouTube channel (see: https://www.youtube.com/watch?v=Os4kEJdwIyU).

connection to normal distributions and more to do with their usefulness in statistical hypothesis testing, which is why they are sometimes called test distributions. Chapter 9 will look at statistical testing more closely. For now, however, let's see what makes these three distributions so special.

7.3.1 The Chi-Squared Distribution

The chi-squared distribution—also known as the χ^2-distribution—forms the basis of many important statistical tests. This distribution was first described in 1876 by the German mathematician Friedrich Robert Helmert, but it was Karl Pearson who showed its usefulness for statistical test methods (Swoboda, 1971, p. 326). As illustrated Fig. 7.24, the density functions of the chi-squared distribution assume different shapes depending on the number of degrees of freedom.

The chi-squared distribution derives from a squared random variable following a standard normal distribution. The square of a random variable following an $N(0;1)$ distribution produces a chi-squared distribution with one degree of freedom:

$$Z^2 \sim \chi_1^2 \tag{7.50}$$

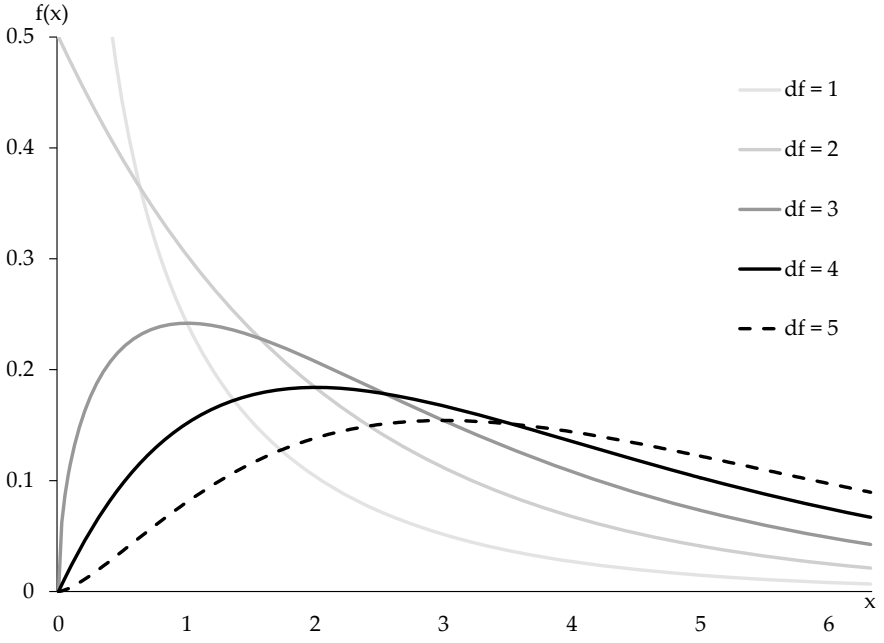


Fig. 7.24 Density function of a chi-squared distribution with different degrees of freedom (df)

The sum of the squares of two independent random variables Z_1 and Z_2 with standard normal distributions produces a chi-squared distribution with two degrees of freedom:

$$Z_1^2 + Z_2^2 \sim \chi_2^2 \tag{7.51}$$

The general formula looks like this:

$$Z_1^2 + Z_2^2 + \cdots + Z_n^2 \sim \chi_n^2 \tag{7.52}$$

where the sum of the squares of n independent random variables following standard normal distributions produces a chi-squared distribution with n degrees of freedom.

The entire area below the density function of a chi-squared distribution equals 1. The area between a given interval describes the probability that a random chi-squared value lies in this interval (Bortz & Schuster, 2010, p. 75). As with a normal distribution, the calculation of the probability without computer or without or a chi-square table takes some time. Appendix B shows a chi-square table for a selection of important quantiles and degrees of freedom. The probability of $(1 - \alpha) = 90\%$ and 20 degrees of freedom produces the quantile of $\chi^2_{90\%;20} = 28.412$.

Two special features of the chi-squared distribution should be mentioned here:

- As the number of degrees of freedom increases, the chi-squared distribution begins to resemble a normal distribution.
- Like the normal distribution and the Poisson distribution, the reproductive property of the chi-squared distribution is given. That is to say, the sum of two independent and chi-squared random variables produces a chi-squared distribution whose number of degrees of freedom equals the sum of the degrees of freedom of the original distributions.

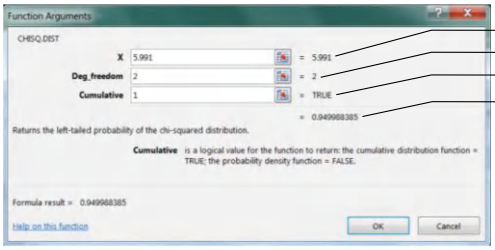
7.3.1.1 Calculating the Chi-Squared Distribution Using Excel

To calculate the probability values in Excel (Fig. 7.25) go to *Formulas* → *Insert Functions* and select *CHISQ.DIST*. Inserting the χ^2-value, the degrees of freedom, and the value 1 for the parameter *cumulative* yields the cumulative probability for p. For the given χ^2-value of $\chi^2 = 5.991$ and DF = 2 degrees of freedom, the cumulative probability is $p = 0.95$. Use the function *CHISQ.DIST*(5.991;2;1) to make the calculation.

Conversely, the function *CHISQ.INV* returns the χ^2-value for a given cumulative probability and a given number of degrees of freedom. For example, a probability of $p = 0.9$ and DF = 20 degrees of freedom produces the χ^2-value $\chi^2_{90\%;20} = 28.412$(*CHISQ.INV(0.9;20)*).

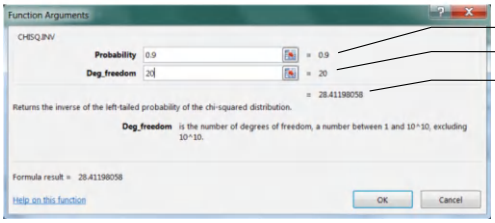
7.3.1.2 Calculating the Chi-Squared Distribution Using Stata

Figure 7.26 summarizes the calculation of a χ^2-distribution using Stata. The command *display invchi2tail(n,α)* calculates the χ^2-value when assuming the



χ^2-value of the χ^2-distribution (χ^2=5.991).
Number of Degrees of Freedom (DF=2)
Cumulative distribution function.
Result: The cumulative probability is p≈0.95.

Syntax: CHISQ.DIST(x;degrees_of_freedom;cumulative); CHISQ.DIST(5.991;2;1)



The cumulative probability is p=0.9.
Number of Degrees of Freedom (DF=20).
Result: χ^2-value of the χ^2-distribution (χ^2=28.41).

Syntax: CHISQ.INV(probability; degrees_of_freedom); CHISQ.INV(0.9;20)

Fig. 7.25 Calculating the chi-squared distribution with Excel

cumulative probability $P = (1 - \alpha)$ and n degrees of freedom. For the Stata command, use the value for α and not, as in Excel, the value for $P = (1 - \alpha)$. To find the χ^2-value of a cumulative χ^2-distribution for, say, $P = (1 - \alpha) = 0.9$, enter $\alpha = 0.1$. The command *display invchi2tail(20,0.1)* for $p = 0.9$ and $n = 20$ degrees of freedom yields the χ^2-value $\chi^2_{90\%;20} \approx 28.412$.

The command *display chi2tail(n,χ^2)* returns α for the reverse cumulative (upper tail) χ^2-distribution with n degrees of freedom. Using *display chi2tail(2,5.991)* results in $\alpha = (1 - p) = 0.05$.

Case	Syntax
Display chi2tail(n,x) returns α for the reverse cumulative (upper tail) χ^2-distribution with n degrees of freedom.	display chi2tail(n,x)
Ex.: *display chi2tail(2,5.991)* results in α=0.05 for n=2 degrees of freedom and a χ2-value of χ2=5.991.	display chi2tail(2,5.991) .05001162
Display invchi2tail(n,α) returns the χ^2-value when assuming the cumulative probability p=(1-α) and n degrees of freedom. Ex.: *display invchi2tail(2,0.1)* yields the χ2-value of χ2=28.41 for p=(1-α)=0.9 and n=2 degrees of freedom.	display invchi2tail(n,α) display invchi2tail(20,0.1) 28.411981

Fig. 7.26 Calculating the chi-squared distribution with Stata

7.3.2 The *t*-Distribution

The *t*-distribution was developed by the British chemist and mathematician William Sealy Gosset (1876–1937) in the early twentieth century. A chemist at the Guinness Brewery, Gosset used the *t*-distribution—derived from normal and chi-squared distributions—to determine the number of yeast cells in small batches of beer. Because he was forbidden by his employer from circulating the results under his own name, Gosset published his research using the pen name "Student". Accordingly, his discovery is often referred to as the Student's *t*-distribution. Sir Ronald A. Fisher (1890–1962) recognized the importance of Gosset's work and expanded its application in statistical testing.

Intuitively, it would seem, small sample sizes do not make for particularly representative results, but in many cases there's no other choice. A prime example is destructive testing. Destructive testing destroys the product whose quality is being tested, and when those products are expensive, researchers want to keep samples as small as possible. The same goes for determining the side-effects of new drugs. Nevertheless, for many years, dogmatic probability theorists disparaged such statistical inferencing, rejecting the idea that accurate insights about a population could be gleaned from just a few values. In 1928, the probability theorist Richard von Mises (1883–1953) described small sample theory as "the erroneous practice of drawing statistical conclusions from short sequences of observations" (von Mises, 1957).

As we know today, Richard von Mises was greatly mistaken. Many test procedures based on the *t*-distribution have since become well-established tools of statistical assessment. They include what would later be known as the *t*-test, which is based on the *t*-distribution, and regression analysis, which uses the *t*-values of regression coefficients.

What does the *t*-distribution look like? Figure 7.27 shows different *t*-distributions. Their single-peaked curves and symmetrical shapes, all centred around $E(X) = 0$ and

$$\mathrm{Var}(X) = \sqrt{\frac{n}{n-2}}, \tag{7.53}$$

recall a normal distribution. And, in fact, the properties of the *t*-distribution do indeed very much resemble those of the normal distribution. And the area between the *x*-axis and the *t*-distribution always equals one (= 100%) regardless of the shape of the curve.

What are the specific applications of *t*-distributions? We know from Sect. 7.2.2 that the following formula can be used to transform normally distributed values into a standard normal distribution:

$$z = \frac{\overline{x} - \mu}{\sigma_{\overline{x}}} = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0; 1) \tag{7.54}$$

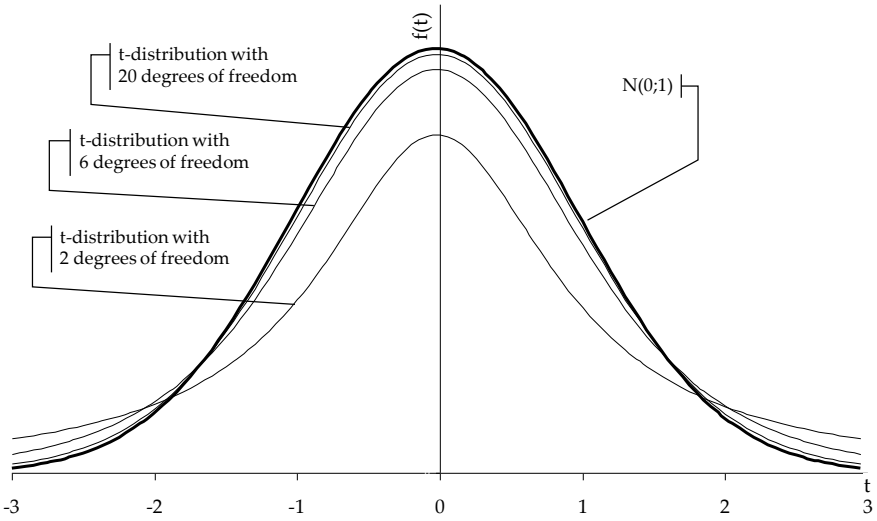


Fig. 7.27 *t*-distribution with varying degrees of freedom

When selecting a sample, we rarely know the mean or standard deviation of the population. If the sample n is greater than 30, μ or S can be used to estimate them. This is because the mean of a sample approximates a normal distribution even if the individual values of the population do not follow a normal distribution. We will return to this subject in more detail below. If the sample is smaller than $n = 30$, and the values of the populations are approximately normally distributed, the values can be standardized as follows:

$$t = \frac{\overline{x} - \mu}{S_{\overline{x}}} = \frac{\overline{x} - \mu}{\frac{S}{\sqrt{n}}} = \frac{\overline{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \tag{7.55}$$

The estimated values for S vary depending on the sample size. Accordingly, the shape of a *t*-distribution also depends on the sample size. Now, the concept of sample size is rarely mentioned in the context of *t*-distributions. What's crucial in the context of *t*-distributions is the number of degrees of freedom. But the number of degrees of freedom is calculated from the sample size. And when the sample size is small, its uncertainty is reflected in the degrees of freedom as well.[6]

Figure 7.27 shows different shapes of *t*-distributions. In contrast to the normal distribution, these distributions can vary with regard to their degrees of freedom.

[6] Ultimately, the number of degrees of freedom is linked to the derivation of the *t*-distribution. A *t*-distribution with n degrees of freedom results from the quotient of a standard normal distribution random variable ($N(0;1)$) and an independent chi-squared random variable with n degrees of freedom: $t^n = \frac{z}{\sqrt{\chi_n^2 / n}}$. For more, see Bortz and Schuster (2010, p. 75).

The smaller the sample, the narrower the peaks of the *t*-distribution and the more observation values on the left and right tails. The larger the sample, the more "centered" the estimated values are in most cases. Accordingly, the observation values tend to group less under the left and right tails and more at the axis of symmetry.

The larger the sample is, the more it approximates a normal distribution. Theoretically we can show that the *t*-distribution converges to a normal distribution as the sample size tends towards infinity ($n \to \infty$). At a sample size of over 30 observations, the difference between the values of a normal distribution and those of a *t*-distribution is marginal. This is why the normal distribution is used in most cases when sample sizes have more than 30 observations.

The probability values of a *t*-distribution can be determined using a *t*-table. This table is given in Appendix C for select probability values (in the columns) and degrees of freedom (in the rows). For instance, a probability value of $p = 0.9$ and 30 degrees of freedom yields a *t*-value of $t_{90\%}^{30} = 1.3104$.

7.3.2.1 Calculating the *t*-Distribution Using Excel

To calculate the probability values in Excel (see Fig. 7.28) go to *Formulas* → *Insert Functions* and select *T.DIST*. Inserting the *t*-value, the degrees of freedom, and the value 1 for the parameter *cumulative* yields the cumulative probability for *p*. For the given *t*-value of $t = 3.078$ and a degree of freedom of $n = 1$, the cumulative probability is $p = 0.9$. Use the function *T.DIST*(3.078;1;1) to make the calculation.

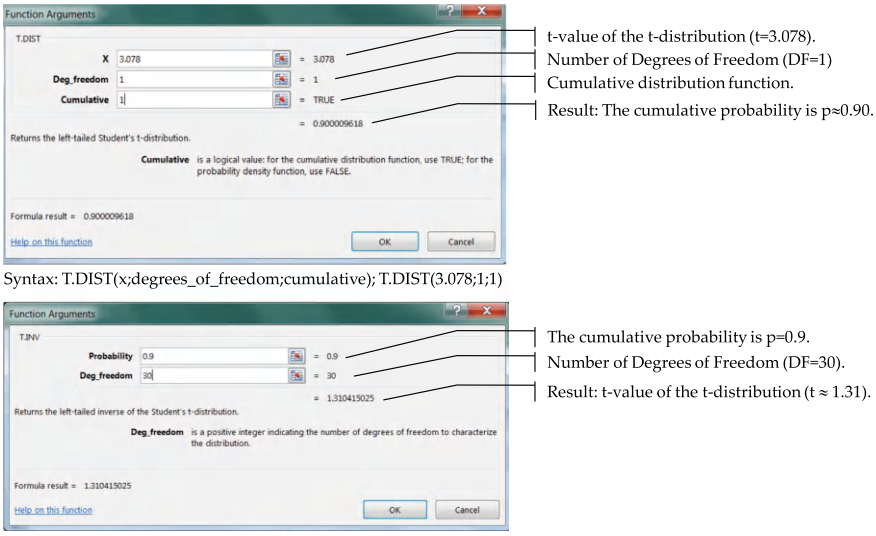


Fig. 7.28 Calculating the *t*-distribution using Excel

Case	Syntax
*Display ttail(n,t) r*eturns the probability α for the reverse cumulative (upper tail) t-distribution with n degrees of freedom. Ex.: *display ttail(1,3.078)* results in α=0.1 for n=1 degree of freedom and a t-value of t=3.078.	display ttail(n,t) display ttail(1,3.078) .09999028
*Display invttail(n,α) r*eturns the t-value when assuming the cumulative probability p=(1-α) and n degrees of freedom. Ex.: *display invttail(30,0.1)* yields the t-value of t=1.31 for p=(1-α)=0.9 and n=30 degrees of freedom.	display invttail(n,α) display invttail(30,0.1) 1.3104

Fig. 7.29 Calculating the *t*-distribution using Stata

Conversely, the function *T.INV* calculates the *t*-value for a specific cumulative probability and degrees of freedom. For example, a probability of $p = 0.9$ and $n = 10$ degrees of freedom produces the *t*-value $t^{10}_{90\%} = 1.372$.

A two-sided interval with the area for α equally divided on both sides of the *t*-distribution generates a *t*-value of $t^n_{1-\frac{\alpha}{2}}$ for the upper limit and a *t*-value of $t^n_{\frac{\alpha}{2}}$ for the lower limit. Using *T.INV(0.95;10)*, the above example of a probability of $p = 0.9$ and $n = 10$ degrees of freedom produces a *t*-value of $t^{10}_{95\%} = 1.812$; using *T.INV(0.05;10)*, it produces a *t*-value of $t^{10}_{5\%} = -1.812$.

7.3.2.2 Calculating the *t*-Distribution Using Stata

Figure 7.29 summarizes the calculation of a *t*-distribution using Stata. The command *display invttail(n,α)* calculates the *t*-value when assuming the cumulative probability $p = (1 - \alpha)$ and n degrees of freedom. For the Stata command, use the value for α and not, as in Excel, the value for $p = (1 - \alpha)$. To find the *t*-value of a cumulative Student's *t*-distribution for, say, $p = (1 - \alpha) = 0.9$, enter $\alpha = 0.1$. The function *display invttail(30,0.1)* for $p = 0.9$ and $n = 30$ degrees of freedom yields the *t*-value $t^{30}_{90\%} = 1.3104$.

The command *display ttail(n,t)* calculates α for the reverse cumulative (upper tail) *t*-distribution with n degrees of freedom. Using *display ttail(1,3.078)* results in $\alpha = 0.1$.[7]

[7] For some good explanations of the Stata approach, see https://www.youtube.com/watch?v=Es-Yyq_MgBY and https://www.youtube.com/watch?v=YNMGbm4CBvA on the *Stata Learner* YouTube channel.

7.3.3 The *F*-Distribution

The *F*-distribution was named after the British biologist and statistician Ronald Aylmer Fisher (1890–1962). It is closely related to the *t*-distribution and the chi-squared distribution and is particularly useful for checking the variance uniformity of two samples and the difference between the means of multiple samples. We will learn more about latter technique—known as the analysis of variance—in Sect. 9.5.1.

Theoretically, an *F*-distribution arises when two independent random variables—let's call them U and V—both follow a chi-squared distribution. From Sect. 7.3.1 we know that a chi-squared distribution varies depending on degrees of freedom. Say that the random variable U follows chi-squared distribution with m degrees of freedom and the random variable V follows a chi-squared distribution with n degrees of freedom, so that

$$U \sim \chi_m^2 \text{ and } V \sim \chi_n^2 \tag{7.56}$$

The quotient

$$F = \frac{U/m}{V/n} \sim F_{m;n} \tag{7.57}$$

is *F*-distributed with m degrees of freedom in the numerator and n degrees of freedom in the denominator. Like the other distributions used for testing, there is not one *F*-distribution but a whole array of possible *F*-distributions. Their shapes depend on the combination of the degrees of freedom in the nominator and the denominator. Like the examples in Fig. 7.30, *F*-distributions always have positive values, are right skewed, and are asymptotic to the x-axis.

7.3.3.1 Calculating the *F*-Distribution Using Excel

To calculate the probability values in Excel (Fig. 7.31) go to *Formulas* → *Insert Functions* and select *F.DIST*. Inserting the *F*-value, the two degrees of freedom m and n, and the value 1 for the parameter *cumulative* yields the cumulative probability for p. For the given *F*-value of $F = 2.364$ and for $m = 13$ and $n = 9$ degrees of freedom, the cumulative probability is $p = 0.9$. Use the function *F.DIST*(2.364;13;9;1) to make the calculation.

Conversely, the function *F.INV* calculates the *F*-value for a specific cumulative probability with m and n degrees of freedom. For example, a cumulative probability of $p = 0.9$ with $m = 13$ and $n = 9$ degrees of freedom produces the *F*-value $F_{13;9;90\%} = 2.3640(F.INV(0, 9; 13; 9))$.

7.3.3.2 Calculating the *F*-Distribution Using Stata

Figure 7.32 summarizes the calculation of an *F*-distribution using Stata. The command *display invF(m,n,p)* calculates the *F*-value when assuming the cumulative probability p with m and n degrees of freedom. To find the *F*-value for, say, $p =$

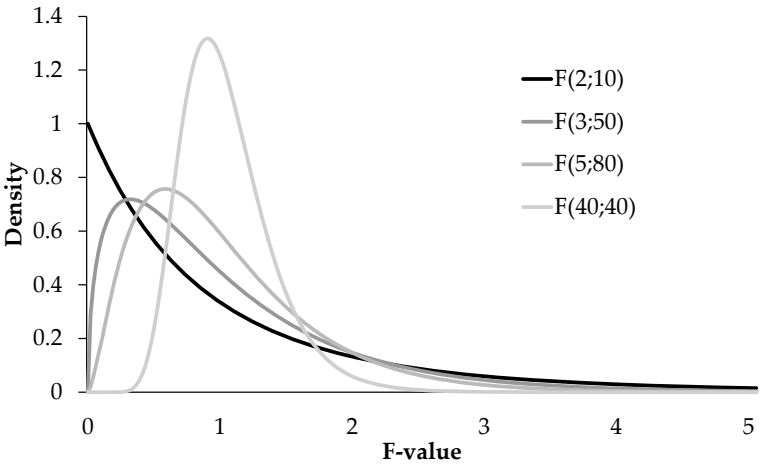


Fig. 7.30 *F*-distributions

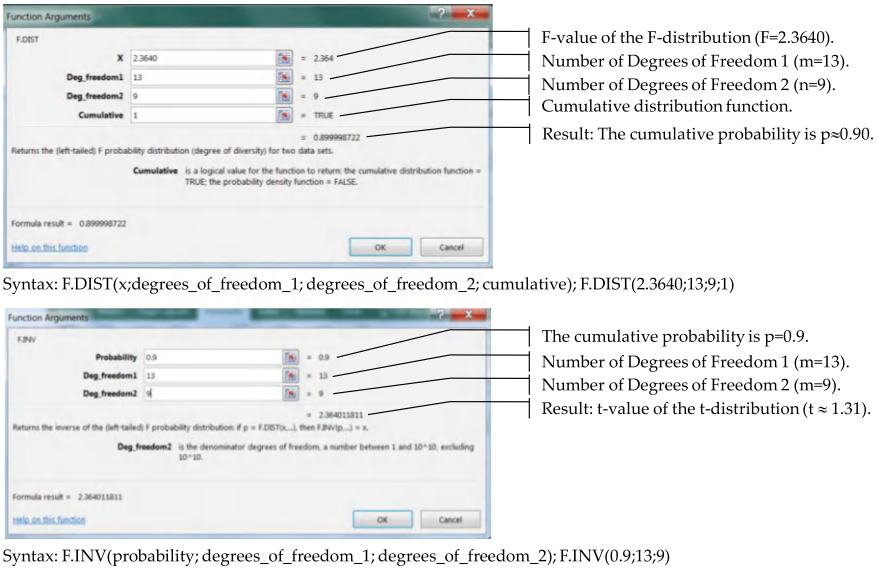


Fig. 7.31 Calculating the *F*-distribution using Excel

0.9, $m = 13$ and $n = 9$ degrees of freedom use the command *display invF(13,9,0.9)*. It results the F-value $F_{13;9;90\%} = 2.3640$.

The command *display F(m,n,f)* calculates the cumulative probability p for a given F-value and for given degrees of freedom. Using *display F(13,9,2.364)* results in $p = 0.9$.

Case	Syntax
The command *display F(m,n,f)* calculates the cumulative probability p for a given F-value and for given degrees of freedom. Ex.: Using *display F(13,9,2.364)* results in p=0.9.	display F(m,n,f) display F(13,9,2.364) .89999872
The command *display invF(m,n,p)* calculates the F-value when assuming the cumulative probability p with m and n degrees of freedom. Ex.: Using the command *display invF(13,9,0.9)* results in $F_{13;9;90\%}$ =2.3640.	display invF(m,n,p) display invF(13,9,0.9)) 2.3640118

Fig. 7.32 Calculating the F-distribution using Stata

7.4 Chapter Exercises

Exercise 33
The local train leaves the station once every 15 min. What is the probability that a randomly arriving passenger will have to wait more than ten minutes?

Exercise 34
True to the motto "just in time"—the principle of flexible and quick response—a car parts supplier keeps only a small number of injection pumps in stock. Within an hour of receiving an order by telephone, the supplier delivers the injection pump to the car repair shop. The delivery time x fluctuates according to the density function:

$$f(x) = \begin{cases} \frac{1}{5} - \frac{x}{50} & 0 \leq x \leq a \\ 0 & \text{else} \end{cases} \tag{7.58}$$

(a) How large is a?
(b) Draw the density function.
(c) Determine $P(2 < X < 4)$, $P(1 < X < 5)$, $P(X \geq 6)$, $P(X = 2)$ and $P(X \leq 4)$ using the density function.
(d) What is the associated distribution function?
(e) Draw the distribution function.
(f) Calculate the probability from (c) using the distribution function.
(g) Show the relationship between density and distribution function using your drawings.
(h) Calculate and interpret the percentiles $x_{0.25}$; $x_{0.5}$ and $x_{0.75}$ as well as the quartile range.
(i) How can these values be calculated graphically?
(j) What are the expected value, variance, and standard deviation of the delivery time?

Exercise 35
The student Peter wants to start his car. He knows that the battery has a 98% probability of being in working order and that each of the four spark plugs has a 95% probability of firing. He also knows that his car will start if the battery and at least three of the spark plugs work. All events are independent when viewed in pairs. What is the probability that his car will start?

Exercise 36
Of all the business students at Pforzheim University, 45% have been to Spain (including the Balearic Islands and the Canary Islands) and 23% have been to Portugal. 11% have been to both Portugal and Spain.

1. What is the probability that a randomly selected business student has been to the Iberian Peninsula (which includes Spain and Portugal)?
2. What is the probability that of five randomly selected students two have been to Spain or Portugal?

Exercise 37
A delivery of N televisions contains M B-grade devices. To estimate the share of B-grade televisions in the batch, a sample n is taken.

(a) What is the distribution of B-grade televisions if each unit is immediately returned after being selected?
(b) What is the distribution of B-grade television if the units are not returned after being selected?
(c) What is the probability when selecting a sample without the units being replaced that $N = 15$, $M = 5$ and $n = 3$ includes more than one B-grade television?

Exercise 38
Of the 25 employees at a mid-sized company, five are fine with longer opening hours. A journalist investigating attitudes towards the new opening hours selects six employees from the company at random.

(a) What is the probability that none of the interviewees speaks out against the longer hours?
(b) What is the probability that none of the interviewees speaks out when each is "returned" to the sample after selection?

Exercise 39
In the production of a certain component, the share of rejected items is $\theta = 0.01$. A researcher selects a random sample of $n = 100$ from the assembly line. Calculate the probability that there is no more than one defective component in the sample.

Exercise 40
The calls per minute to a telephone switchboard follow a Poisson distribution of $\lambda = 2$. Calculate the probabilities that in a given minute…

1. …no calls occur;
2. …no more than two calls occur.

Exercise 41
The company *Compsolutions* has 40 employees. Of these, 20% are members of the union and would support a strike in the coming labour dispute. Mr. Miller's marketing department consists of six employees. What is the chance that Mr. Miller's entire department will continue to work in the event of a strike?

Exercise 42
In the accounting department of the same company, an accountant completes 500 accounting entries per day. The probability of making a mistake is 0.004. What is the probability that the accountant will make more than one error in a day?

Exercise 43
Researchers count vehicles at two points of an intersection. The number of vehicles that pass by per minute has a Poisson distribution of $\lambda_1 = 1.2$ and $\lambda_2 = 0.8$. What is the probability that within five minutes no more than six vehicles will pass an observation point?

Exercise 44
The company Hedonic has made coffee dispensers available to its workers. The coffee consumption of its 100 employees has a normal distribution with an expected value of 130 l/week and a standard deviation of five l/week.

(a) What is the probability that 125 L of coffee are enough for a week?
(b) Management wants the coffee supply to have a 95% probability of meeting demand. How many litres of coffee must they have available each week?

Exercise 45
The average weight of chicks the same age and variety is 105 g with a standard deviation of 2.5 g and a normal distribution.

1. What is the probability that the weight of a chick is less than 100 g?
2. What is the probability that the weight of a chick is more than 101 g and less than 108 g?
3. What weight do 40% of chicks fall below?

Exercise 46
A machine fills bags of coffee in such a way that the weight of a filled bag has a normal distribution of $\mu = 510$ g and $\sigma = 20$ g. The packages are labelled "Net weight: 500 g".

1. What per cent of bags are underweight?
2. What per cent of bags weigh more than 550 g?
3. How large must μ be with the same standard deviation so that only 1% of bags are underweight?

Exercise 47
The company *Autotrans* specializes in the transport of vehicles between manufacturers and retailers. Every once and a while a car is damaged during transport—a median of 1.21 vehicles per month. The standard deviation of the damaged vehicles is 1.1 vehicles.

(a) What is the probability that three cars are damaged in a month? Which distribution applies? Explain your answer.
(b) What is the probability that 16 vehicles are damaged in twelve months?
(c) What is the probability that between ten and twenty vehicles are damaged in twelve months? Explain why another distribution applies.

Exercise 48
A storeowner wants to get rid of 25 fireworks from last year in the coming New Year's season. The manufacturer guarantees that 60% of the fireworks—in this case, 15 units—will ignite after a year. Let us assume that firework functionality is independent and thus follows a hypergeometric distribution. Given this information, let us assume that a customer buys five of the 25 fireworks.

(a) What is the probability that only one firework will ignite?
(b) Assume that the storeowner has 1000 fireworks from the previous year and a customer buys 50 of them. Once again, firework functionality is guaranteed to be 60%. Which distribution would you assume to determine the probability

that a certain number of fireworks ignite if you didn't have a calculator to determine the actual distribution? Explain your answer. You don't need to perform the calculation here.

Exercise 49
Each day a wholesaler makes deliveries to 50 retailers. The probability that a retailer (independently of the others) will return one of the products is 0.03.

(a) What is the probability that exactly one retailer returns a product?
(b) What would be the approximate probability of ten products being returned if the wholesaler made deliveries to 300 retailers?

Exercise 50
Imagine a factory that produces a certain type of machine.

(a) The operators know from experience that the probability of a day passing without an error occurring is 0.9048. Which distribution is suited to describe the number of errors (random variable X) that occur during a day? Calculate the value of the distribution's parameter.
(b) What is the probability that four machines (independently of each other) make exactly one error in a single day?

7.5 Exercise Solutions

Solution 33
The waiting time follows a uniform distribution:

$$f(x) = \frac{1}{15} \quad 0 \leq x \leq 15 \tag{7.59}$$

The probability that a train arrives in the next ten minutes is given by:

$$F(x = 10) = \int_{a}^{x=10} \frac{1}{b-a} \, \mathrm{d}x = \frac{x-a}{b-a} = \frac{10-0}{15-0} = \frac{2}{3} \tag{7.60}$$

The probability that a train does not arrive in the next ten minutes is given by:

$$1 - F(x = 10) = 1 - \frac{2}{3} = \frac{1}{3} \tag{7.61}$$

Solution 34

(a)

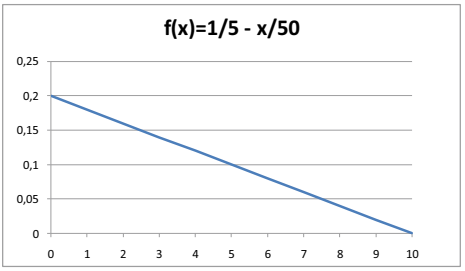
$$F(X) = \int_0^a \left(\frac{1}{5} - \frac{x}{50}\right) \mathrm{d}x = 1 \Rightarrow F(X) = \left(\frac{x}{5} - \frac{x^2}{100}\right)\Big|_0^a \tag{7.62}$$

$$\Rightarrow \frac{a}{5} - \frac{a^2}{100} = 1 \Rightarrow a^2 - 20a + 100 = 0 \Rightarrow a_{1|2} = 10 \pm \sqrt{100 - 100} \tag{7.63}$$

$$\Rightarrow a = 10 - \sqrt{100 - 100} = 10 \tag{7.64}$$

(b)

x	f(x)	F(x)
0	0.2	0
1	0.18	0.19
2	0.16	0.36
3	0.14	0.51
4	0.12	0.64
5	0.1	0.75
6	0.08	0.84
7	0.06	0.91
8	0.04	0.96
9	0.02	0.99
10	0	1



(c)

$$P(2 < X < 4) = \int_2^4 \left(\frac{1}{5} - \frac{x}{50}\right) \mathrm{d}x = \left(\frac{x}{5} - \frac{x^2}{100}\right)\Big|_2^4 \tag{7.65}$$
$$= \frac{4}{5} - \frac{16}{100} - \left(\frac{2}{5} - \frac{4}{100}\right) = 0.28$$

$$P(1 < X < 5) = \int_1^5 \left(\frac{1}{5} - \frac{x}{50}\right) \mathrm{d}x = \left(\frac{x}{5} - \frac{x^2}{100}\right)\Big|_1^5$$
$$= 1 - \frac{25}{100} - \left(\frac{1}{5} - \frac{1}{100}\right) = 0.56 \tag{7.66}$$

$$P(X \geq 6) = \int_6^{10} \left(\frac{1}{5} - \frac{x}{50}\right) \mathrm{d}x = \left(\frac{x}{5} - \frac{x^2}{100}\right)\Big|_6^{10}$$
$$= 2 - 1 - \left(\frac{6}{5} - \frac{36}{100}\right) = 0.16 \tag{7.67}$$

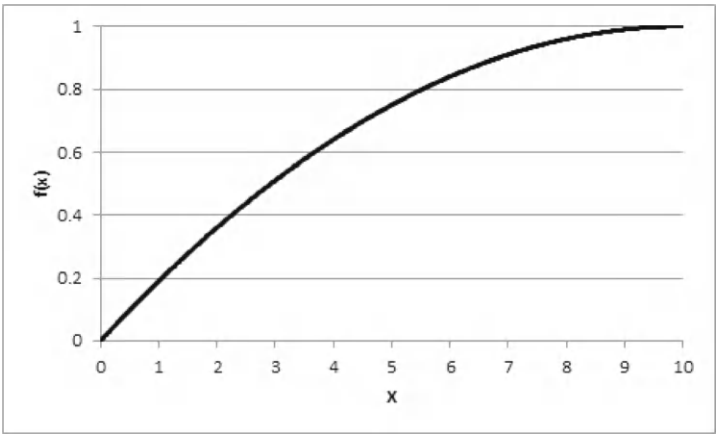
$$P(X = 2) = \int_2^2 \left(\frac{1}{5} - \frac{x}{50}\right) \mathrm{d}x = 0 \tag{7.68}$$

$$P(X \leq 4) = \int_0^4 \left(\frac{1}{5} - \frac{x}{50}\right) \mathrm{d}x = \left(\frac{x}{5} - \frac{x^2}{100}\right)\Big|_0^5 = \frac{4}{5} - \frac{16}{100} = 0.64 \quad (7.69)$$

(d)

$$F(x) = \begin{cases} 0 & x < 0 \\ \int_0^x \left(\frac{1}{5} - \frac{u}{50}\right) \mathrm{d}u = \left(\frac{u}{5} - \frac{u^2}{100}\right)\Big|_0^x = \frac{x}{5} - \frac{x^2}{100} & 0 \leq x < 10 \\ 1 & x \geq 10 \end{cases} \quad (7.70)$$

(e) $F(X) = x/5 - x^2/100$



(f) $P(2 < X < 4) = F(4) - F(2) = 0.64 - 0.36 = 0.28$; $P(1 < X < 5) = F(5) - F(1) = 0.75 - 0.19 = 0.56$; $P(X \geq 6) = 1 - F(6) = 1 - 0.84 = 0.16$; $P(X = 2) = F(2) - F(2) = 0$; $P(X \leq 4) = F(4) = 0.64$.

(g) The distribution function indicates the area below the density function up to point x. The density function indicates the slope of the distribution function at point x. This means that integrating the density function yields the distribution function and extrapolating the distribution function yields the density function.

(h) $$P\big(X \leq x_p\big) = F(x) = p \Rightarrow \frac{x_p}{5} - \frac{x_p^2}{100} = p \Rightarrow x_p^2 - 20x_p + 100p = 0 \quad (7.71)$$

$$\Rightarrow x_{1/2} = 10 \pm \sqrt{100 - 100p} \Rightarrow x_p = 10 - \sqrt{100 - 100p} \quad (7.72)$$

$$x_{0.25} = 10 - \sqrt{100 - 100 \cdot 0.25} = 1.34 \quad (7.73)$$

$$x_{0.5} = 10 - \sqrt{100 - 100 \cdot 0.5} = 2.93 \quad (7.74)$$

$$x_{0.75} = 10 - \sqrt{100 - 100 \cdot 0.75} = 5 \tag{7.75}$$

$$\text{IQR} = x_{0.75} - x_{0.25} = 5 - 1.34 = 3.66 \tag{7.76}$$

(i) Using the inverse functions of $F(x)$.

(j) $$E(X) = \int_0^{10} x \left(\frac{1}{5} - \frac{x}{50}\right) \mathrm{d}x = \left(\frac{x^2}{10} - \frac{x^3}{150}\right)\Big|_0^{10} = 10 - \frac{1000}{150} = \frac{10}{3} \tag{7.77}$$

Calculation of variance using the law of displacement:

$$\text{Var}(X) = E\left[(X - E(X))^2\right] = E\left(X^2\right) - E(X)^2 \tag{7.78}$$

$$\begin{aligned} E\left(X^2\right) &= \int_0^{10} x^2 \left(\frac{1}{5} - \frac{x}{50}\right) \mathrm{d}x = \left(\frac{x^3}{15} - \frac{x^4}{200}\right)\Big|_0^{10} \\ &= \frac{1000}{15} - \frac{10000}{200} = \frac{200}{3} - 50 = \frac{50}{3} \end{aligned} \tag{7.79}$$

$$\begin{aligned} \text{VAR}(X) &= \frac{50}{3} - \left(\frac{10}{3}\right)^2 = \frac{150 - 100}{9} = \frac{50}{9} \\ &= 5.55 \Rightarrow \sigma_X = \sqrt{\text{Var}(X)} = 2.36 \end{aligned} \tag{7.80}$$

Solution 35

Let the following variables be defined thus:

B battery works
Z at least three spark plugs work
A car starts.

Due to the independence of the events, the probability that the car starts is determined by the formula:

$$P(A) = P(Z \cap B) = P(B)P(Z) = 0.98 \cdot P(Z) \tag{7.81}$$

To calculate $P(Z)$ the random variable X can be defined as the number of working spark plugs. This means that $P(Z) = P(X = 3) + P(X = 4)$. Due to the independence of the events, the random variable X has a binomial distribution, with $n = 4$ and $p = 0.95$.

$$P(X = 3) = \binom{4}{3} 0.95^3 \cdot 0.05^1 = \frac{4!}{3! \cdot 1!} 0.95^3 \cdot 0.05^1$$

$$= 4 \cdot 0.857 \cdot 0.05 = 0.171 \tag{7.82}$$

$$P(X = 4) = \binom{4}{4} 0.95^4 \cdot 0.05^0 = 0.815 \tag{7.83}$$

Hence, $P(Z)$ yields: $P(Z) = P(X = 3) + P(X = 4) = 0.171 + 0.815 = 0.986$. The probability of the car starting can be calculated as follows:

$$P(A) = P(Z \cap B) = P(B)P(Z) = 0.98 \cdot P(Z) = 0.98 \cdot 0.986 = 0.96628. \tag{7.84}$$

Solution 36

1. Let A be the event "Student has been to Spain", B the event "Student has been to Portugal". Hence, $P(A \cup B)$ means that the student has already been to Spain or Portugal. According to the Inclusion–Exclusion Principle,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.45 + 0.23 - 0.11 = 0.57$$

2. X is the number of persons who were already in Spain or in Portugal. This yields a binomial distribution $X \sim B(5; 0.57)$, whereby:

$$\begin{aligned} P(X \geq 2) &= P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) \\ &= 1 - P(X = 0) - P(X = 1) \end{aligned}$$

$$\begin{aligned} P(X \geq 2) &= 1 - \binom{5}{0} 0.57^0 \cdot 0.43^5 - \binom{5}{1} 0.57^1 \cdot 0.43^4 \\ &= 1 - 0.0147 - 0.0974 = 0.8879 \end{aligned}$$

Solution 37

(a) Binomial distribution;
(b) Hypergeometric distribution;
(c) $P(X > 1) = 1 - P(X \leq 1) = 1 - F_x(1) = 1 - \left(\frac{120}{455} + \frac{5 \cdot 45}{455}\right) = 0.24175.$

Solution 38
Let X be the number of employees in the sample who are fine with the longer opening hours. Furthermore, let

N be the total number of all objects = 25.
M be the number of objects with a specific property = 5.

n be the number of selections = 6.
x be the number of objects with a specific property in the selections = 0.

(a) Hypergeometric distribution:

$$H(25, 5, 6, 0) = \frac{\binom{N-M}{n-x}\binom{M}{x}}{\binom{N}{n}} = \frac{\binom{20}{6}\binom{5}{0}}{\binom{25}{6}} = \frac{38{,}760}{177{,}100} = 0.219 \tag{7.85}$$

(b) Binomial distribution: $P(X = 0) = \binom{6}{0}\left(\frac{5}{25}\right)^0\left(1 - \frac{5}{25}\right)^6 = 0.8^6 = 0.262$ (7.86)

Solution 39
The answer can be solved either with this formula:

$$P(X \leq 1) = P(X = 0) + P(X = 1) \tag{7.87}$$

$$\begin{aligned} P(X \leq 1) &= \binom{100}{0} \cdot 0.01^0 \cdot 0.99^{100} + \binom{100}{1} 0.01^1 \cdot 0.99^{99} \\ &= 0.3660 + 0.3697 = 0.7357 \end{aligned} \tag{7.88}$$

or by using an approximation with a Poisson distribution

$$\lambda = n \cdot \theta = 100 \cdot 0.01 = 1 \tag{7.89}$$

$$P(X \leq 1) = \frac{1^0 e^{-1}}{0!} + \frac{1^1 e^{-1}}{1!} = 0.3679 + 0.3679 = 0.7358 \tag{7.90}$$

Solution 40

(1)

$$P(X = 0) = \frac{2^0 e^{-2}}{0!} = e^{-2} = 0.1353 \tag{7.91}$$

$$
\begin{aligned}
&P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) \\
(2)\quad &P(X \leq 2) = \frac{2^0\, e^{-2}}{0!} + \frac{2^1\, e^{-2}}{1!} + \frac{2^2\, e^{-2}}{2!} \\
&\qquad = 0.1353 + 0.2707 + 0.2707 = 0.6767
\end{aligned}
\tag{7.92}
$$

Solution 41
Let the random variable X be the number of employees willing to strike. Furthermore, let

N be the total number of objects $= 40$.
M be the number of objectives with a specific property $= 8$.
n be the number of selections $= 6$.
x be the number of objects with a specific property in the selections $= 0$.

Hypergeometric distribution:

$$
H(40, 8, 6, 0) = \frac{\binom{N-M}{n-x}\binom{M}{x}}{\binom{N}{n}} = \frac{\binom{32}{6}\binom{8}{0}}{\binom{40}{6}} = 0.236. \tag{7.93}
$$

Solution 42
The probability of error is very low ($P = 0.004$), so assume a Poisson distribution. Let the random variable X be the number of erroneous accounting entries per day.

$$
\text{Poisson distribution: } P(X = x) = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda} \tag{7.94}
$$

$$
\lambda = E(X) = n \cdot p = 500 \cdot 0.004 = 2
$$

$$
P(X > 1) = 1 - P(X = 0) - P(X = 1) = 1 - \frac{2^0}{0!}\mathrm{e}^{-2} - \frac{2^1}{1!}\mathrm{e}^{-2} = 59.4\% \tag{7.95}
$$

Solution 43
Reproductive property $\lambda = \lambda_1 + \lambda_2 = 1.2 + 0.8 = 2$ vehicles per minute and repeated application of the reproductive property for each minute: $5 \cdot 2 = 10$ vehicles every five minutes. What is the probability that within five minutes no more than six vehicles pass the point of observation? With the Poisson distribution

this corresponds to the summarized Poisson probabilities for 0, 1, 2, 3, 4, 5, and 6 vehicles:

$$P(X \leq 6) = \sum_{i=0}^{6} \frac{\lambda^x}{x!} e^{-\lambda} = \frac{10^6}{6!} \mathrm{e}^{-10} + \cdots + \cdots = 0.13 \tag{7.96}$$

Using normal distribution if $\lambda \geq 10 : N(\mu; \mu^{0.5}) = N\left(10; 10^{0.5}\right)$, we get:

$$\begin{aligned} P(x \leq 6.5) &= Z\left(\frac{x - \lambda}{\sqrt{\lambda}}\right) \\ &= Z\left(\frac{6.5 - 10}{\sqrt{10}}\right) = Z(-1.11) = 1 - Z(1.11) = 1 - 0.8665 = 0.13 \end{aligned} \tag{7.97}$$

Here the upper value is selected to be 6.5, as in this case the *discrete* Poisson distribution is approximated by a *continuous* normal distribution. This means that all values of the continuous distribution between 6 and 6.5 is regarded to be a discrete value of 6. This is called a continuity correction.

Solution 44
Let the random variable X = coffee consumption in l/week; $X \sim N(130, 5)$.

(a)
$$\begin{aligned} P(X \leq 135) &= P\left(\frac{X - 130}{5} \leq \frac{135 - 130}{5}\right) \Rightarrow P\left(Z \leq \frac{135 - 130}{5}\right) \\ &\Rightarrow P(Z \leq 1) = \Phi(1) = 0.8413 \end{aligned} \tag{7.98}$$

(b)

$$P(X \leq x) = 0.95 \Rightarrow P(Z \leq 1.65) = 0.95 \Rightarrow \frac{x - 130}{5} = 1.65 \Rightarrow x = 138.25 \tag{7.99}$$

Solution 45
We have to calculate $P(X \leq 100$ g). First the random variable X must be transformed into a random variable Z with a standard normal distribution, so that:

1.
$$\begin{aligned} P(X \leq 100) &= P\left(\frac{X - 105}{2.5} \leq \frac{100 - 105}{2.5}\right) \\ P(X \leq 100) &= P(Z \leq -2) = 1 - P(Z \leq 2) \\ &= 1 - 0.9772 = 0.0228 = 2.28\% \end{aligned} \tag{7.100}$$

which is to say, 2.28% of chicks weigh less than 100 g.

2.
$$P(101 \leq X \leq 108) = P\left(\frac{X-101}{2.5} \leq Z \leq \frac{100-108}{2.5}\right)$$
$$P(101 \leq X \leq 108) = P(Z \leq 1.2) - P(Z \leq -1.6)$$
$$= P(Z \leq 1.2) - (1 - P(Z \leq 1.6)) = 0.8301 \tag{7.101}$$

3. $P(X \leq c) = 0.4.$ (7.102)

For the random variable Z with a standard normal distribution, we get

$$P\left(Z \leq \frac{c-105}{2.5}\right) = 0.4 \tag{7.103}$$

Since only values larger than or equal to 0.5 are tabulated,

$$P\left(Z \leq -\frac{c-105}{2.5}\right) = 0.6. \tag{7.104}$$

For the 60% quantile, the result is approximately 0.255, so that:

$$-\frac{c-105}{2.5} = 0.255 \text{ g} \tag{7.105}$$

This can be converted to: $c = 105 - 2.5 \cdot 0.255 = 104.3625$ g.

Solution 46

1. We need to calculate $P(X \leq 500 \text{ g})$. First, we need to transform the random variable X into a random variable Z with a standard normal distribution (z-transformation):

$$z = \frac{X - 150 \text{ g}}{20 \text{ g}} \sim N(0, 1). \tag{7.106}$$

This yields:

$$P(X \leq 500 \text{ g}) = P\left(\frac{X - 510 \text{ g}}{20 \text{ g}} \leq \frac{500 - 510 \text{ g}}{20 \text{ g}}\right) = F(-0.5)$$
$$= 1 - F(0.5) = 1 - 0.6915 = 0.3085. \tag{7.107}$$

This means that 30.85% of the bags are underweight.

2.
$$P(X \geq 550 \text{ g}) = P\left(\frac{X - 550 \text{ g}}{20 \text{ g}} \leq \frac{550 - 510 \text{ g}}{20 \text{ g}}\right)$$
$$= 1 - F(2) = 1 - 0.9772 = 0.0228 \tag{7.108}$$

3. The following applies:

$$P(X \leq 550\text{ g}) = P\left(\frac{X-\mu}{20\text{ g}} \leq \frac{500\text{ g}-\mu}{20\text{ g}}\right) = F\left(\frac{500\text{ g}-\mu}{20\text{ g}}\right) = 0.01 \tag{7.109}$$

Since only values larger than or equal to 0.5 are tabulated, we get:

$$1 - F\left(\frac{500\text{ g}-\mu}{20\text{ g}}\right) = 1 - 0.01 \Rightarrow F\left(-\frac{500\text{ g}-\mu}{20\text{ g}}\right) = 0.99 \tag{7.110}$$

This yields: $-\frac{500g-\mu}{20\text{ g}} = 2.326$, since 2.326 represents the 99% quantile of the standard normal distribution and -2.326 represents the 1% quantile of the standard normal distribution. Converting this we get: $\mu = 546.52$ g.

Solution 47

(a) Let the random variable X = damaged vehicles per month. Since the probability of a damaged car is very low, assume that the random variable X has a Poisson distribution:

$$P(X = x) = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda} \text{ with } \lambda = E(X) = 1.21$$
$$P(X = 3) = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda} = \frac{1.21^3}{3!}\mathrm{e}^{-1.21} = 8.80\%$$

(b) Let the random variable Y = damaged vehicles per year. As a result of its reproductive property: $\lambda = E(Y) = 12 \cdot 1.21 = 14.52$. The probability of 16 damaged vehicles in twelve months is: $P(Y = 16) = \frac{\lambda^y}{y!}\mathrm{e}^{-\lambda} = \frac{(14.52)^{16}}{16!}\mathrm{e}^{-(14.52)} = 9.22\%$.

(c) 11 partial probabilities must be calculated. Approximation by normal distribution is simpler and since $E(Y) > 10$ it is possible.

$$E(Y) = 14.52;\ \mathrm{Var}(Y) = 14.52;\ \sigma = 3.811;\ y_u = 9.5;\ y_o = 20.5 \tag{7.111}$$

$$\begin{aligned} P(9.5 \leq Y \leq 20.5) &= P(Y \leq 20.5) - P(Y \leq 9.5) \\ &= P\left(Z \leq \frac{20.5 - 14.52}{3.811}\right) - P\left(Z \leq \frac{9.5 - 14.52}{3.811}\right) \end{aligned} \tag{7.112}$$

$$\begin{aligned} P(9.5 \leq Y \leq 20.5) &= P(Z \leq 1.569) - P(Z \leq -1.317) \\ &= \Phi(1.569) - [1 - \Phi(1.317)] \end{aligned} \tag{7.113}$$

$$P(9.5 \leq Y \leq 20.5) = 0.9418 - [1 - 0.9066] = 0.8484 \tag{7.114}$$

Here the values 9.5 and 20.5 are selected as the lower and upper limits, since in this case the *discrete* Poisson distribution is approximated by a *continuous* normal distribution. This means that all values of the continuous distribution between 9.5 and 20.5 are regarded to be discrete values between ten and twenty. This is known as a continuity correction.

Solution 48

Let the random variable X be the number of firework that ignite. Furthermore, let

N be the total number of objects = 25;
M be the number of objects with a specific property = 15;
n be the number of selections = 5;
x be the number of objects with a specific property in the selections = 1.

(a) Hypergeometric distribution:

$$H(25, 15, 5, 1) = \frac{\binom{N-M}{n-x}\binom{M}{x}}{\binom{N}{n}} = \frac{\binom{10}{4}\binom{15}{1}}{\binom{25}{5}} = 5.93\%. \tag{7.115}$$

(b) The approximation of the normal distribution as M/N is between 0.1 und 0.9 and $n > 30$:

$$N\left(n\frac{M}{N}; \sqrt{n\frac{M}{N}\left(1 - \frac{M}{N}\right)\left(\frac{N-n}{N-1}\right)}\right) = N\left(30; \sqrt{11.4}\right) \tag{7.116}$$

Solution 49

(a) Let the random variable X be the number of returned products:

$$P(X = 1) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{50}{1} 0.03^1 \cdot (1 - 0.03)^{50-1} = 33.72\%$$

The result can also be calculated using a Poisson distribution, as $n \cdot p = 50 \cdot 0.03 = 1.5 < 10$ and $n > 1500p = 1500 \cdot 0.03 = 45$:

$$P(X = 1) = \frac{\lambda^x}{x!} e^{-\lambda} = \frac{np^x}{x!} e^{-np} = \frac{1.5^1}{1!} e^{-1.5} = 33.47\% \tag{7.117}$$

(b) $\binom{300}{10} p^{10} \cdot (1-p)^{300-10} = 12.04\%$. Normal calculators usually cannot perform this operation. Because $n \cdot p = 300 \cdot 0.03 = 9 < 10$ and $1500 \cdot p = 1500 \cdot 0.03 = 45 < n = 300$ is given: Approximation via a Poisson distribution of $\lambda = E(X) = n \cdot p = 9$:

$$P(X = 10) = \frac{\lambda^x}{x!} e^{-\lambda} = \frac{9^{10}}{10!} e^{-9} = 11.85\% \tag{7.118}$$

Solution 50

(a) Let the random variable X be the number of:

$$P(X = 0) = \frac{\lambda^x}{x!} e^{-\lambda} = \frac{\lambda^0}{0!} e^{-\lambda} = 0.9048 \Rightarrow \lambda = 0.1 \tag{7.119}$$

(b) Let the random variable Z be the number of errors that occur in four machines per day:

$$P(Z = 1) = \frac{(4 \cdot 0.1)^z}{z!} e^{-(4 \cdot 0.1)} = \frac{(0.4)^1}{1!} e^{-(0.4)} = 0.268 \text{ (reproductive property)} \tag{7.120}$$

References

Bortz, J., & Schuster, C. (2010). *Statistik für Sozialwissenschaftler* (7th ed.). Springer.
de Moivre, A. (1738). *Doctrine of chance* (2nd ed.). Woodfall.
Mises, R. von (1957). *Probability, Statistics and Truth* (2nd ed.). New York: Dover Publications.
Swoboda, H. (1971). *Exakte Geheimnisse: Knaurs Buch der modernen Statistik*. Knaur.



Parameter Estimation

8

Now that we have laid the theoretical foundations of probability calculus in the past chapters, let us recall what all the effort was about. The main purpose of inductive statistics is to develop methods for making generalizations about a population from sample data. This chapter will present these methods, known as statistical estimation of parameters. Statistical estimation is a procedure for estimating the value of an unknown population parameter—for example, average age. There are two types of estimation procedures: point estimation and interval estimation.

Point estimations are used to estimate mean values ($E(x)$ or μ) or variances ($\text{Var}(x)$ or σ) of a population using sample mean values $\bar{x}$ or sample variance S^2. A typical example of point estimation would be a statement like this: "The average price of a competitor's product is €2.38". Of course, as with everything in life, the more we tie ourselves to a single fixed value, the more likely we are to be off the mark. What if the average price is €2.39 or €2.40? For this reason, it is usually a good idea to estimate an interval. An example of interval estimation is this statement: "It is 99% certain that the average price of a competitor's product is between €2.36 and €2.40". An interval estimation is nothing more than a point estimation with a certain margin of error. The size of the margin of error depends on the amount of certainty required. The larger the interval, the more certain the estimation. In the following we consider point estimation and then on its basis discuss interval estimation.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78070-7_8.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*,
Springer Texts in Business and Economics,
https://doi.org/10.1007/978-3-031-78070-7_8

8.1 Point Estimation

Say we draw a simple random sample n and identify the mean value $\bar{x}$ for a certain variable such as the price of a new type of yoghurt. Can we assume that the mean value we identify is the mean value of the population μ? To be on the safe side, we could repeat the process and draw a second sample. If the mean values of the two samples $\bar{x}_1$ and $\bar{x}_2$ are close, we can generally assume that the mean values of the samples are good estimates for the actual mean of the population. Yet rarely are $\bar{x}_1$ and $\bar{x}_2$ identical and rarely do they equal the exact value of μ. To be very certain, we could, theoretically, draw an infinite number of samples (with replacement). These samples will usually differ with regard to the mean value $\bar{x}$. Taken together, all values for $\bar{x}$ produce a *sampling distribution*. This distribution tells us how well an individual mean value of a sample estimates the mean value μ of a population. The narrower the distribution is, the more exact the estimation provided by the sample mean value $\bar{x}$. In other words, if we know the sampling distribution, we can draw conclusions about the accuracy of an estimation. Let us take a closer look at the specific features of sample distributions.

One intuition might be to assume that the sampling distribution of a mean corresponds to the distribution of an empirical population's variables being investigated. That is to say, taking samples from a normally distributed population and calculating their mean values should eventually result in a normally distributed sampling distribution of the mean.

We can verify this assumption with an example. Say a population has a normal age distribution, a mean value of $\mu = 35$ years and a standard deviation of $\sigma = 10$ years, as shown in part 1 of Fig. 8.1. The first sample #1 we draw is small with a size of $n = 5$ and contains the values 31.46; 28.16; 44.62; 42.76; 34.41, for a mean value of $\bar{x}_1 = 36.28$ years. It would be pure luck if the mean from a sample of $n = 5$ corresponded exactly to the mean value of the population. As it happens, the mean value of the sample is not far from $\mu = 35$, the actual value of the population. If we perform multiple drawings of $n = 5$ and calculate the mean value of the sample each time, we should see that the mean values are distributed around the actual value of the population.

For simplicity's sake, let us use Excel to perform this task for us. First, use the add-ins manager to make sure that the *Analysis ToolPak* and the *Analysis ToolPak VBA* are activated.[1] Then under the Data tab click the *Data Analysis* command button and select the *Random Number Generation* entry from the list. We want to draw 1000 samples of size $n = 5$ from the population specified above with $N(\mu = 35; \sigma = 10)$. The results are partly illustrated in Fig. 8.2. Sample #1 above is taken from cells B2 to B6 in the Excel table. The 1000 samples yield many mean values (row #7) in the vicinity of the actual mean value of the population $\mu = 35$ (see column B to column ALM in the Excel table of Fig. 8.2). In a few rare cases, the sampled mean values fall far above or below the population mean.

[1] In Excel this can be reached by clicking *File* → *Options* → *Add-ins* → *Go*.

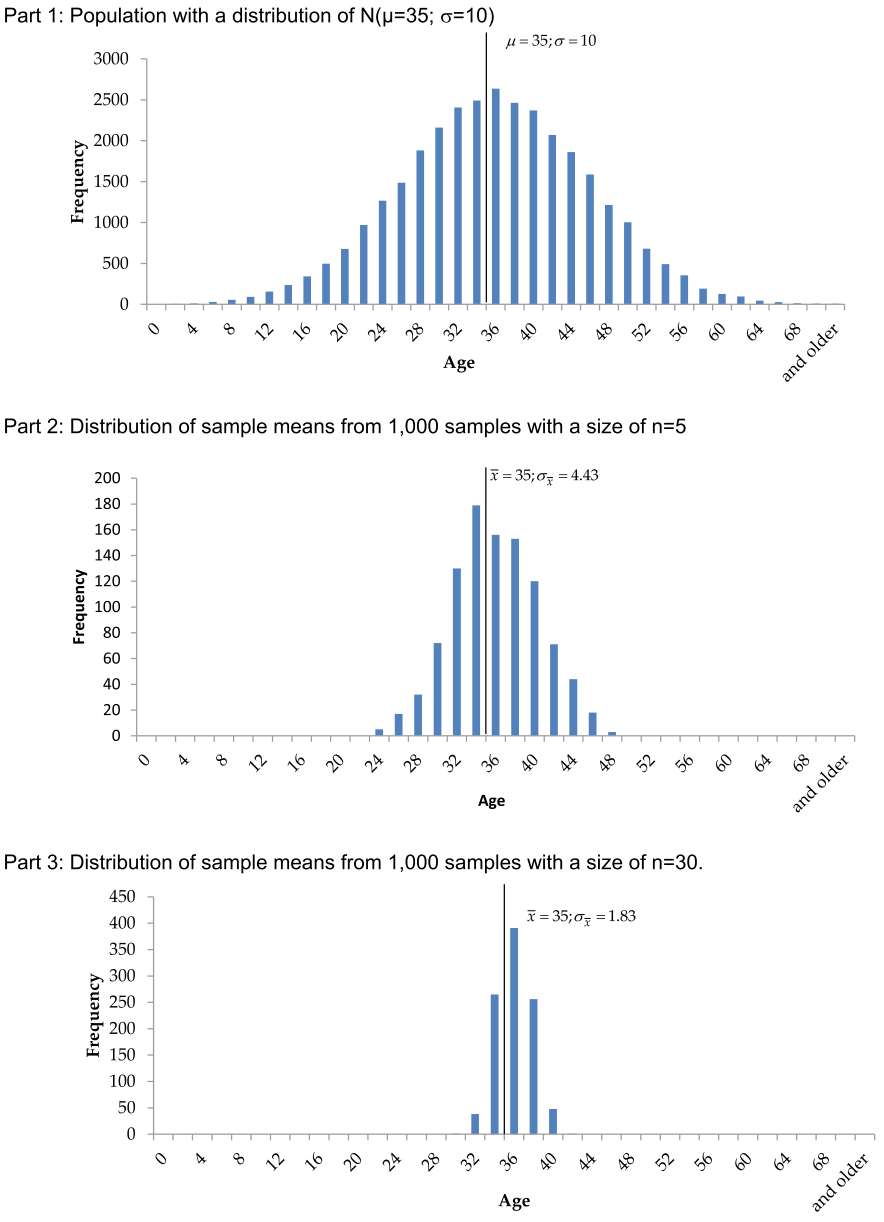
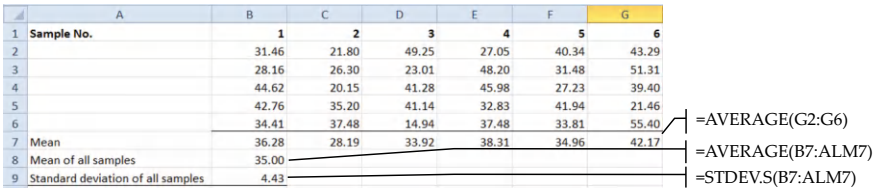


Fig. 8.1 Distribution of sample means in a normally distributed population



	A	B	C	D	E	F	G	
1	**Sample No.**	**1**	**2**	**3**	**4**	**5**	**6**	
2		31.46	21.80	49.25	27.05	40.34	43.29	
3		28.16	26.30	23.01	48.20	31.48	51.31	
4		44.62	20.15	41.28	45.98	27.23	39.40	
5		42.76	35.20	41.14	32.83	41.94	21.46	
6		34.41	37.48	14.94	37.48	33.81	55.40	=AVERAGE(G2:G6)
7	Mean	36.28	28.19	33.92	38.31	34.96	42.17	=AVERAGE(B7:ALM7)
8	Mean of all samples	35.00						=STDEV.S(B7:ALM7)
9	Standard deviation of all samples	4.43						

Fig. 8.2 Generating samples using Excel: 1000 samples with a size of $n = 5$ from a population with a distribution of $N(\mu = 35; \sigma = 10)$

If we create a histogram (see part 2 of Fig. 8.1), we will notice that the mean value $\bar{x}$ of the sample is normally distributed around the actual mean value of the population $\mu = 35$. What is striking, however, is that the standard deviation of the distribution of the sample mean values is smaller than that of the population. On average, the standard deviation of the sample mean values is $\sigma_{\bar{x}} = 4.43$. Comparing this value with the standard deviation of the population ($\sigma = 10$) we can identify the following relationship: the deviation of the sampling distribution of the mean equals the quotient of the standard deviation of the population and the square root of the sample size. The parameter $\sigma_{\bar{x}}$ is called the standard error of the sample. In our case:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{5}} = 4.47 \approx 4.43. \tag{8.1}$$

If samples with a size n are drawn out of a normally distributed population with $N(\mu; \sigma)$, the mean values of the samples will also have a normal distribution with

$$N\left(\mu; \frac{\sigma}{\sqrt{n}}\right) = (\mu; \sigma_{\bar{x}}). \tag{8.2}$$

This means that both $\bar{x}$ and $\sigma_{\bar{x}}$ allow us to infer the mean value and the standard deviation of the population.

As a rule, the larger the scope of the sample, the narrower the normal distribution of the mean values of the samples. This is shown in part 3 of Fig. 8.1 with a sample size of $n = 30$. The size yields a lower standard error than a sample with the size $n = 5$:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{30}} = 1.83. \tag{8.3}$$

The root term of sample size n in the denominator accounts for the narrow distribution, in accordance with the formal relationship described above. However, there is another, more intuitive explanation: the larger the sample, the more data points from the population go into calculating the sample mean. In large samples,

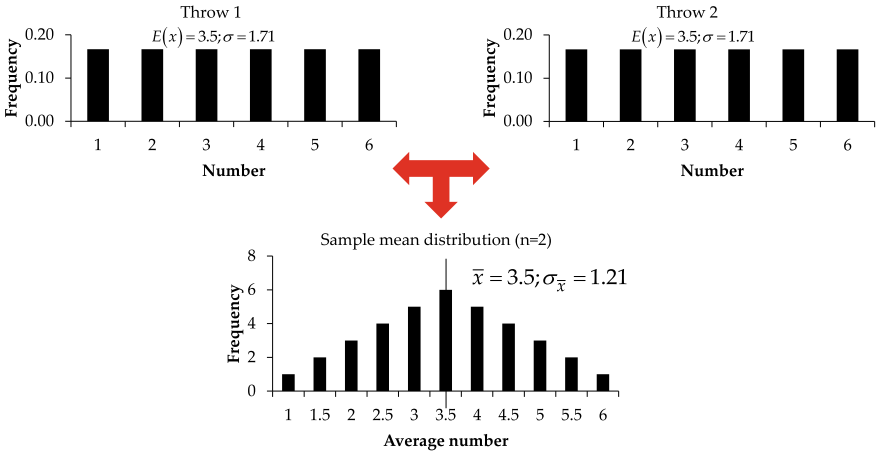


Fig. 8.3 Distribution of mean with $n = 2$ throws of an unloaded die

outliers like the large age deviations in our example are frequently counterbalanced by deviations in the other direction. In small samples, this compensatory effect tends to be smaller, leaving mean values more widely dispersed.

In our above example, we assumed a special case: a population with a normal distribution. However, we can only assume a normally distributed population if the population is known as such. Knowledge about the distribution of a population is rarely given. This necessarily leads us to ask about the relationship between a population without a normal distribution and the sampling distribution of the mean. To answer this question, let us start by considering an experiment involving a single unloaded die. The probability of throwing any number is 1/6 and follows a uniform (rather than a normal) distribution. The mean value of a single throw is $\mu = (1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$ with a standard deviation of $\sigma = 1.71$. In the experiment, the die is thrown twice and the mean value is determined for the sample ($n = 2$). The distribution of all possibilities is depicted in Fig. 8.3.[2] Although the results are uniformly distributed, the distribution of the mean value for $n = 2$ throws already shows the beginnings of a normal distribution, if still discrete and triangular.

In Fig. 8.4, the same experiment has been carried out with four throws. In this case, the average score is $\overline{x} = 3.5$ and $\sigma_{\overline{x}} = 0.85$, and the result very much resembles a normal distribution. Here too the mean of the sample distribution $\overline{x} =$

[2] The possible scores for the mean value are: (1; 1) → 1; (1; 2) → 1.5; (1; 3) → 2; (1; 4) → 2.5; (1; 5) → 3; (1; 6) → 3.5; (2; 1) → 1.5; (2; 2) → 2; (2; 3) → 2.5; (2; 4) → 3; (2; 5) → 3.5; (2; 6) → 4; (3; 1) → 2; (3; 2) → 2.5; (3; 3) → 3; (3; 4) → 3.5; (3; 5) → 4; (3; 6) → 4.5; (4; 1) → 2.5; (4; 2) → 3; (4; 3) → 3.5; (4; 4) → 4; (4; 5) → 4.5; (4; 6) → 5; (5; 1) → 3; (5; 2) → 3.5; (5; 3) → 4; (5; 4) → 4.5; (5; 5) → 5; (5; 6) → 5.5; (6; 1) → 3.5; (6; 2) → 4; (6; 3) → 4.5; (6; 4) → 5; (6; 5) → 5.5; (6; 6) → 6.

3.5 corresponds to the mean of a theoretical population (μ). The standard error of the sample distribution can be calculated using the formula:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.71}{\sqrt{4}} \approx 0.85. \tag{8.4}$$

Its likeness to a normal distribution is no coincidence. The *central limit theorem* tells us that the distribution of sample means gradually approaches a normal distribution as the sample size increases, regardless of the distribution of the population. To understand why, consider the distributions of the sample means of two populations without normal distributions illustrated in Fig. 8.5. Thirty thousand samples with a size of $n = 2$ were first drawn from a bimodal population (upper-left panel). Note that the mean values of these samples do not yet have a normal distribution (centre-left panel). However, starting with a sample size as small as $n = 5$ (lower-left panel), the sample mean distribution begins to approximate a normal distribution. The situation is similar with a left-skewed population, as shown on the right side of the same figure.

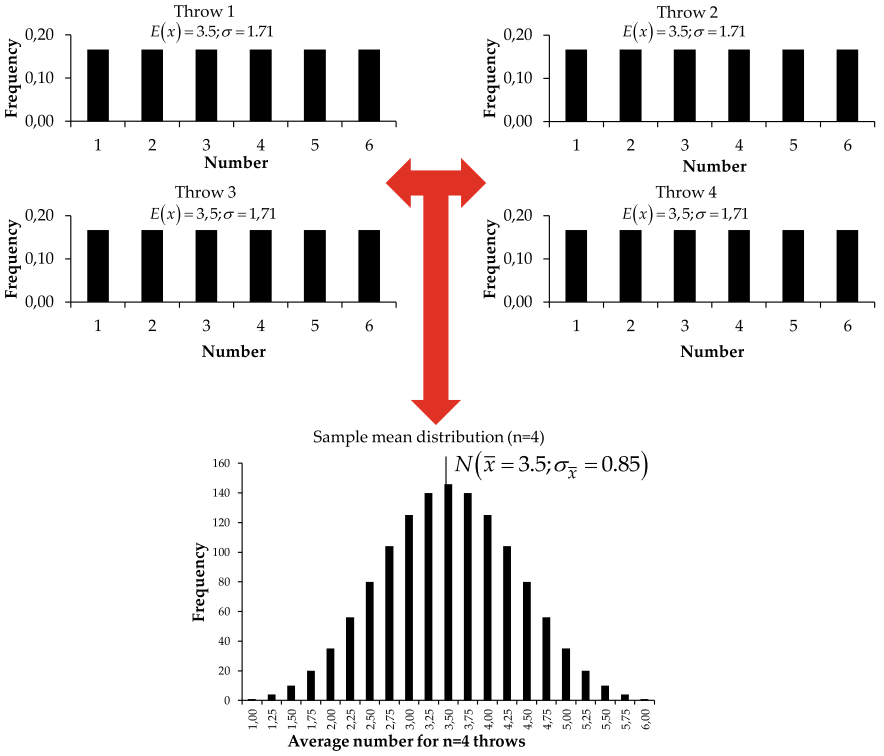


Fig. 8.4 Distribution of the mean with $n = 4$ throws of an unloaded die

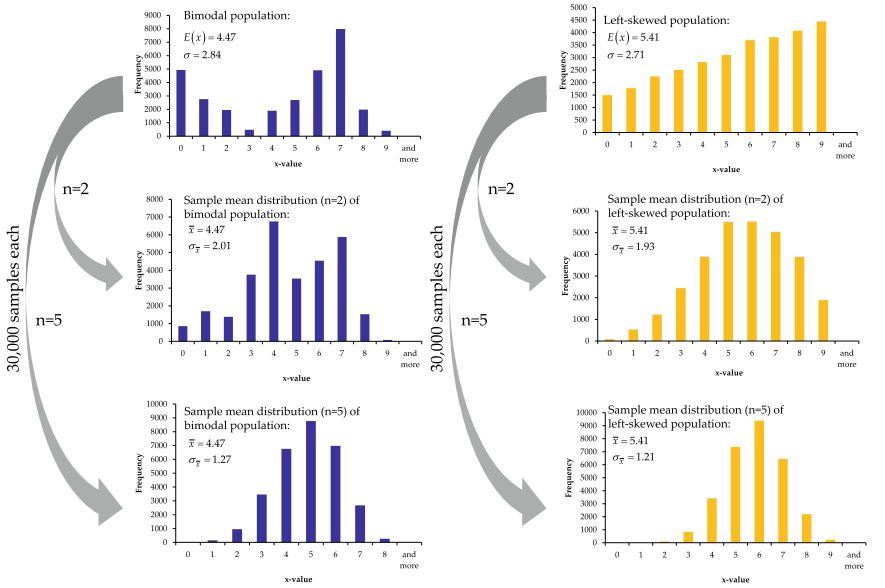


Fig. 8.5 Sample mean distribution of a bimodal and a left-skewed population for 30,000 samples of sizes $n = 2$ and $n = 5$

In sum, a sufficiently large sample produces sample means with a normal distribution whatever the original distribution of the population. Hence, we need not know the actual distribution of values in the population to assume a normal distribution of sample means in the form of

$$N\left(\mu; \frac{\sigma}{\sqrt{n}}\right). \tag{8.5}$$

Theoreticians and practitioners do not completely agree about what constitutes a sufficiently large sample. However, most textbooks assume a sample size of $n \geq 30$. The above examples make clear that even for not normally distributed random variables but a sufficiently large sample size, the mean $\bar{x}$ of the sampling distribution is equal to the mean of the population μ and the standard error of the sample tends to have a value of

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}. \tag{8.6}$$

Strictly speaking, to determine the standard error we have to know the standard deviation of the population σ, which is usually not the case. However, the standard deviation of the population can be accurately estimated by the standard deviation S (or $\hat{\sigma}$) of the individual values of a sample. Students often have difficulties distinguishing between the standard deviation of the sample S and the standard error

$\sigma_{\bar{x}}$. The standard error $\sigma_{\bar{x}}$ identifies the dispersion of the means from many samples drawn in succession, while the standard deviation of the sample S describes the dispersion of individual values in a given sample. It is possible to estimate the standard error without knowing the standard deviation of the population using the following formula[3]:

$$\hat{\sigma}_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{S_{\text{emp}}}{\sqrt{n-1}} \tag{8.7}$$

The sample parameters $\bar{x}$ and S have the following properties as estimated values for the population parameters μ and σ:

- The estimators are *consistent* because the estimated values approximate the value of the parameter to be estimated as the sample size increases.
- The estimators are *unbiased* because the estimated value of the sample distribution corresponds to the population parameter.

8.2 Interval Estimation

8.2.1 The Confidence Interval for the Mean of a Population (μ)

In the previous section, we first learned that a sample mean value $\bar{x}$ can be used for a point estimation of the mean of the population μ. We then learned that these point estimations differ from sample to sample and thus represent a random variable whose distribution we have to know to assess the accuracy of the estimation. The smaller the standard deviation of sample means, the more exact our estimate. Finally, we learned that though the distribution and its standard deviation are unknown at first, we can assume a normal distribution of the sample means if either:

(a) the population itself has a normal distribution, or
(b) the sample has a size of $n \geq 30$.

In those instances where we are correct to assume a normal distribution of sample mean, it is possible to determine a so-called confidence interval for the mean of the population μ. Confidence intervals do not estimate an exact value; they provide lower and upper limits for a given degree of certainty $(1 - \alpha)$. This certainty $(1 - \alpha)$ is called the *confidence level*. The interval estimation is nothing other than a point estimation with a built-in margin of error. In the price example described above—"it is 99% certain that the price of a competitor's product lies between

[3] In statistics, estimated parameters are indicated by adding a circumflex over them.

€2.36 and €2.40"—€2.36 represents the lower confidence limit and €2.40 represents the upper confidence limit. The confidence level $(1 - \alpha)$ amounts to 99%. The word "confidence" is used to express the likelihood that the mean of the population falls within the calculated interval. Accordingly, there is a risk of $\alpha = 1\%$ that the mean of the population does not fall within the calculated confidence interval.

The size of a confidence interval depends among other things on the amount of certainty researchers need for their study. The more certainty they need for their estimations, the larger the confidence level $(1 - \alpha)$, and the greater confidence interval they require.

How do we calculate confidence intervals? We know from the section on point estimation that sample means have a normal distribution in samples with a size of $n \geq 30$. The following thus applies:

$$\bar{x} \sim N(\mu; \sigma_{\bar{x}}) \tag{8.8}$$

Subjecting the normally distributed sample to a z-transformation and converting the normally distributed values into a standard normal distribution yields:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \sim N(0; 1) \tag{8.9}$$

We perform this step because tabulated values are available only for the standard normal distribution $N(0; 1)$. We no longer need the lower and upper interval limits for $\bar{x}$ but for its respective z-transformed value. The limits are determined with a confidence level of $(1 - \alpha)$ in the middle of the normal curve. In the case of so-called two-sided intervals, this will leave α equally allocated on both sides of the normal curve and the confidence interval—$\alpha/2$ above the upper limit and $\alpha/2$ below the lower interval limit.

This relationship is schematically illustrated for our price example in Fig. 8.6. With a confidence level of $(1 - \alpha) = 99\%$, the error of probability is $\alpha = 1\%$. Since the interval limits before the z-transform are €2.36 and €2.40, there is a $\alpha/2 = 0.5\%$ probability that the average price lies below €2.36 (the shaded area on the left) or above €2.40 (the shaded area on the right). Transformed to z-values, these errors are instances in which the z-values are smaller than the $z_{(\alpha/2)}$ percentile or larger than the $z_{(1 - \alpha/2)}$ percentile of the standard normal distribution. All z-values lie between the $z_{(0.5\%)}$ percentile and the $z_{(99.5\%)}$ percentile and thus satisfy the condition of the confidence level of $(1 - \alpha) = 99\%$.

The following formula summarizes this relationship:

$$P\left(z_{\frac{\alpha}{2}} \leq z \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha \Leftrightarrow P\left(z_{\frac{\alpha}{2}} \leq \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha \tag{8.10}$$

Since $z_{(\alpha/2)} = (-\, z_{(1-\alpha/2)})$, the following applies:

$$\Leftrightarrow P\left(-\, z_{1-\frac{\alpha}{2}} \cdot \sigma_{\bar{x}} \leq \bar{x} - \mu \leq z_{1-\frac{\alpha}{2}} \cdot \sigma_{\bar{x}}\right) = 1 - \alpha \tag{8.11}$$

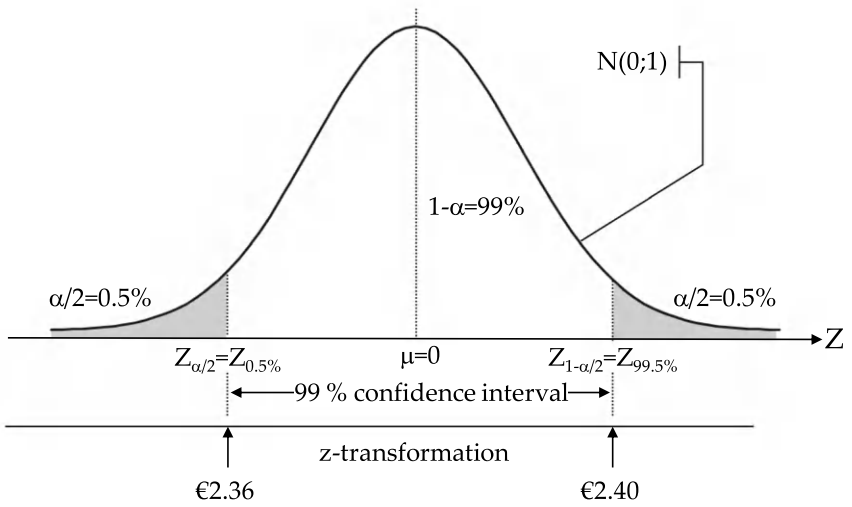


Fig. 8.6 Confidence interval in the price example

$$\Leftrightarrow P\left(-\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \sigma_{\bar{x}} \leq -\mu \leq -\bar{x} + z_{1-\frac{\alpha}{2}} \cdot \sigma_{\bar{x}}\right) = 1 - \alpha \tag{8.12}$$

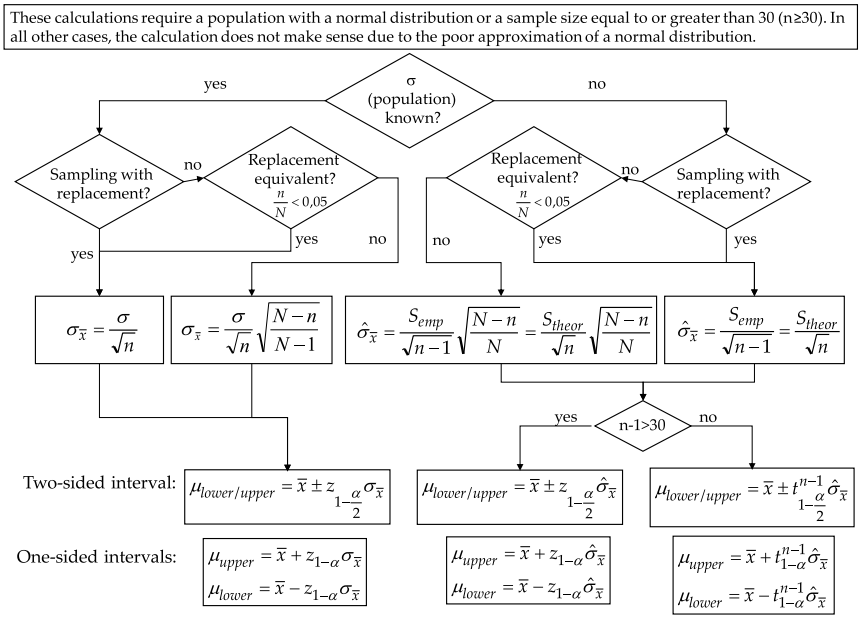
Multiplying by (-1) yields a formula for calculating confidence intervals:

$$P\left(\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \sigma_{\bar{x}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \sigma_{\bar{x}}\right) = 1 - \alpha. \tag{8.13}$$

The expected value of the population μ is covered by an interval with a confidence level of $(1 - \alpha)$. The interval limits are determined from the mean of the sample plus or minus the standard error weighted by $z_{(1-\alpha/2)}$. The size of the standard error $\sigma_{\bar{x}}$ rarely results directly from the distribution of population σ, since the latter is usually unknown. As with point estimation the estimation of the standard error occurs using the standard deviation of the sample:

$$\hat{\sigma}_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{S_{\text{emp}}}{\sqrt{n-1}}. \tag{8.14}$$

Figure 8.7 provides a flow chart summarizing the calculation of confidence intervals under various conditions. Again, one must carefully distinguish between cases where the standard deviation of the population is known from the beginning (left branch in Fig. 8.7) and those where it is not (right branch in Fig. 8.7) and between cases where there is sampling with replacement and where there is sampling without replacement (see Sect. 6.3). This flow chart provides us with two examples for determining the size of confidence intervals.



Note:
Two-sided intervals: In two-sided intervals, $\mu_{lower/upper}$ forms the lower and upper limits of the confidence interval. With a confidence level of (1-α), the mean of the population is contained in this interval. **One-side intervals**: To determine the **minimum** mean value (equal to or greater than a given condition) of the population with a given confidence level of (1-α), form the confidence interval using the lower limit μ_{lower} and the upper limit infinite. To determine the **maximum** mean value (equal to or less than a given condition) of the population with a confidence level of (1-α), form the confidence interval using the lower limit minus infinite and the upper limit μ_{upper}.

Fig. 8.7 Calculating confidence intervals for means

Example 1

Assume a standard deviation $\sigma = 0.6$ is known for a population with $N = 1000$ products. You draw a sample of $n = 100$ products and obtain a mean of six defective products. What confidence interval contains the population mean with 90% confidence, if

(a) the sample is drawn with replacement?
(b) the sample is drawn without replacement?

This example describes the rare case where the standard deviation and the size of a population is known (left branch in Fig. 8.7). The following information is available for part (a) of the exercise:

- sampling with replacement,
- a standard error of $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.6}{\sqrt{100}} = 0.06$,
- a sample mean of $\bar{x} = 6$, and a value for $\alpha = 10\%$.

For the desired confidence interval, this yields:

$$\mu_{\text{lower/upper}} = \overline{x} \pm z_{1-\frac{\alpha}{2}}\sigma_{\overline{x}} = 6 \pm z_{1-\left(\frac{0.1}{2}\right)} \cdot 0.1 = 6 \pm 1.65 \cdot 0.06$$
$$\Rightarrow P(5.901 \leq \mu \leq 6.099) = 0.9 \tag{8.15}$$

The lower and upper limits of the confidence interval are 5.901 and 6.099 products, respectively.

The calculation of part (b) proceeds similarly, though the sampling is without replacement. Here we have a finite population whose size decreases with each product picked. In all cases in which the population size N is known and the quotient $n/N \geq 0.05$, the factor (see Fig. 8.7)

$$\sqrt{\frac{N-n}{N-1}} \tag{8.16}$$

reduces the deviation of the distribution and, by extension, the confidence interval. The larger the sample is relative to the population, the more accurate the estimation. Based on the following conditions:

- sampling without replacement
- $n/N = 100/1000 = 0.1 \geq 0.05$
- $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \frac{0.6}{\sqrt{100}} \cdot \sqrt{\frac{1000-100}{1000-1}} = 0.06 \cdot 0.95 = 0.057;\ \overline{x} = 6;\ \alpha = 10\%.$

We obtain for the desired confidence interval:

$$\mu_{\text{lower/upper}} = \overline{x} \pm z_{1-\frac{\alpha}{2}}\sigma_{\overline{x}} = 6 \pm z_{1-\left(\frac{0.1}{2}\right)} \cdot 0.057 = 6 \pm 1.65 \cdot 0.057 \tag{8.17}$$

$$\mu_{\text{lower/upper}} = P(5.906 \leq \mu \leq 6.094) = 0.9 \tag{8.18}$$

The lower and upper limits of the confidence interval are now 5.906 and 6.094 products, respectively.

Example 2

Ten bags are selected from a large delivery of onions. The bags are determined to have the following weights in kilogrammes: 9.5; 10.5; 10.0; 10.0; 10.2; 10.0; 10.4; 9.6; 9.8; 10.0. The weight of the packaged onions approximates a normal distribution:

(a) Determine the 95% confidence interval for the average weight of the delivered bags.
(b) What is the lowest average weight of the bags? Identify the one-sided 95% confidence interval.
(c) What is the highest average weight of the bags? Identify the one-sided 95% confidence interval.

(d) What would be the result from (a) if the sample contained 41 bags, all other parameters being equal?

In this case, the dispersion of the population is unknown (right branch in Fig. 8.7):

- The size of the population N is assumed to be infinite so that $n/N < 0.05$.
- Because the sample size ($n - 1 = 10 - 1 = 9$) is smaller than 30, the confidence interval must be calculated using the t-distribution.
- $\overline{x} = 10;\ S_{\text{emp}} = 0.3;\ \alpha = 5\%$.
- $\hat{\sigma}_{\overline{x}} = \frac{S_{\text{emp}}}{\sqrt{n-1}} = \frac{0.3}{\sqrt{10-1}} = 0.1$.

For the two-sided confidence interval from part (a), we get:

$$\mu_{\text{lower/upper}} = \overline{x} \pm t_{1-\frac{\alpha}{2}}^{n-1} \cdot \hat{\sigma}_{\overline{x}} = 10 \pm t_{1-\frac{0.05}{2}}^{10-1} \cdot 0.1 = 10 \pm 2.262 \cdot 0.1 \tag{8.19}$$

$$\mu_{\text{lower/upper}} = P(9.774 \leq \mu \leq 10.226) = 0.95 \tag{8.20}$$

For the one-sided confidence interval from part (b), we get:

$$\begin{aligned} \mu_{\text{lower}} &= \overline{x} - t_{1-\alpha}^{n-1} \cdot \hat{\sigma}_{\overline{x}} = 10 - t_{1-0.05}^{10-1} \cdot 0.1 = 10 - 1.812 \cdot 0.1 \\ &\Rightarrow P(9.819 \leq \mu) = 0.95 \end{aligned} \tag{8.21}$$

For the one-sided confidence interval from part (c), we get:

$$\begin{aligned} \mu_{\text{upper}} &= \overline{x} + t_{1-\alpha}^{n-1} \cdot \hat{\sigma}_{\overline{x}} = 10 + t_{1-0.05}^{10-1} \cdot 0.1 = 10 + 1.812 \cdot 0.1 \\ &\Rightarrow P(\mu \leq 10.181) = 0.95 \end{aligned} \tag{8.22}$$

For part (d) we can forgo the t-distribution in favour of the normal distribution on account of the sample size: $n - 1 = 40 - 1 = 39 > 30$. This yields:

$$\mu_{\text{lower/upper}} = \overline{x} \pm z_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{\overline{x}} = \overline{x} \pm z_{1-\frac{\alpha}{2}} \frac{S_{\text{emp}}}{\sqrt{n-1}} = 10 \pm 1.96 \cdot \frac{0.3}{\sqrt{40}} \tag{8.23}$$

$$\mu_{\text{lower/upper}} = P(9.907 \leq \mu \leq 10.093) = 0.95 \tag{8.24}$$

The approximation of the t-distribution can be represented by the normal distribution. This is shown by the result that would have been calculated relative to a calculation using the t-distribution. In this case, the tabular value of the t-distribution with 40 degrees of freedom is 2.021. The attendant confidence interval would only be six grammes longer:

$$P(9.904 \leq \mu \leq 10.096) = 0.95 \tag{8.25}$$

8.2.2 Planning the Sample Size for Mean Estimation

Practitioners and scientists must choose not only the empirical sampling method (see Sect. 5.1) but also the size of the sample. The sample size is the number of elements to be included in a survey. Usually it is chosen in a way that statistical analyses based on the sample have a statistical significance of a particular power. The elements chosen for analysis are called sampling units. In many cases, these elements are people, but it also can involve larger groupings of individuals (e.g. households, institutions) or objects (e.g. products). The collection of all sampling units is called the gross sample. However, the gross sample size must be distinguished from the net sample size. It is not unusual that respondents refuse to participate in a survey (non-response) or that sampled respondents are not part of the target group. These individuals are omitted in the net sample. What ultimately matters for calculating statistical results is the size of the net sample because only available data can be statistically analysed.

We already know that the accuracy of an estimation increases with the sample size. However, each additional sampled unit costs extra time and money. Given the conflicting aims of accuracy and effort, how do we determine an appropriate sample size?

Textbooks provide a variety of suggestions. Fowler (2002) and Oppenheim (1992) emphasize that important subgroups in a survey—men or women, say—should have 100 observations each, while less important subgroups should have between 20 and 50 observations. Lewin (2005) suggests at least 30 subjects per group. Note that these represent net values. Such recommendations from experienced scientists are helpful, of course, but other considerations go into determining the best sample size as well: the special features of the population under examination, the number of collected variables, the methods of evaluation, the sample sizes used in similar studies in the past, etc. Ultimately, the appropriate sample size depends on multiple factors and cannot be determined by blanket generalizations. When calculating confidence intervals, we learned that the accuracy of the estimation depends on the distribution of the population. Homogenous populations have narrow distributions with small standard deviations that require smaller sample sizes, while heterogeneous populations have broader distributions with large standard deviations that require larger sample sizes for the same level of accuracy. We also learned that sample sizes are shaped by the accuracy of the estimation needed for a particular study. The higher the confidence level $(1 - \alpha)$ that a researcher requires, the larger the sample size has to be.

What we know about confidence intervals can now be used to calculate appropriate sample sizes for a given application. The following example serves to illustrate. A market research agency conducts a survey on the average family income. The goal is to estimate a two-sided confidence interval for average income with a confidence level of 95% whose length amounts to no more than €100 (see Fig. 8.8). A preliminary study yielded a standard deviation of $S_{\text{emp}} = €500$.

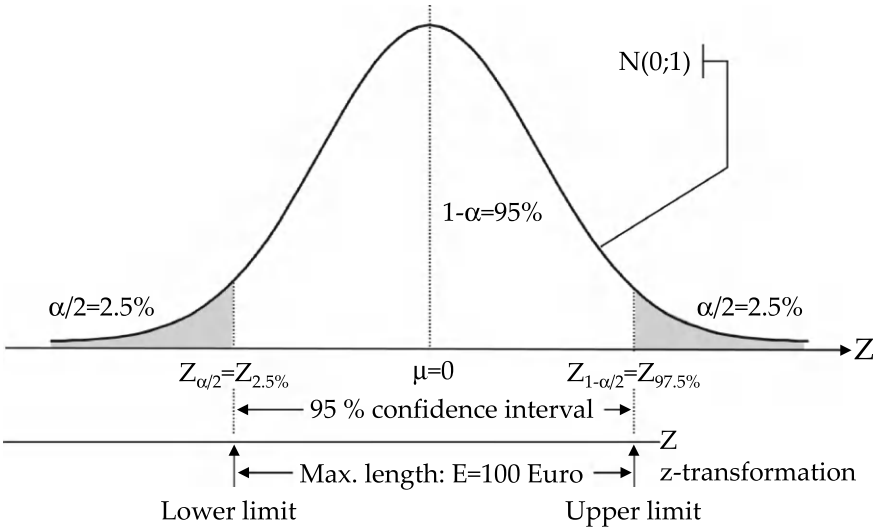


Fig. 8.8 Length of a two-sided confidence interval for means

The length of the confidence interval represents a deviation of the mean from the upper limit plus the same deviation from the lower limit. The deviations represent the standard error $\hat{\sigma}_{\bar{x}}$ weighted with $z_{(1-\alpha/2)}$ for infinite samples with $n \geq 30$. For the length of the interval, this yields (see Fig. 8.8):

$$E = 2 \cdot z_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{\bar{x}} = 2 \cdot z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} = 2 \cdot z_{1-\frac{\alpha}{2}} \frac{S_{\text{emp}}}{\sqrt{n-1}} \Leftrightarrow E = 100 \tag{8.26}$$

Solved for n, this results in a general formula for determining net sample size:

$$n = \frac{2^2 \cdot z_{1-\frac{\alpha}{2}}^2 \cdot S^2}{E^2} = \frac{2^2 \cdot z_{1-\frac{\alpha}{2}}^2 \cdot S_{\text{emp}}^2}{E^2} + 1 \tag{8.27}$$

Plugging the numbers from the example into the formula yields a recommended net sample size of 386 observations:

$$n = \frac{4 \cdot 1.96^2 \cdot 500^2}{100^2} + 1 = 385.15 \rightarrow 386 \tag{8.28}$$

Practical questions also arise when planning of sample size for one-sided confidence intervals. Say a tire manufacturer knows from a previous study that its tires have an average lifespan of 41,000 km with a standard deviation of $S =$ 8000 km. The manufacturer wants to ensure that its tires have a lifespan of at least 40,000 km with 99% confidence. How large must the sample? Figure 8.9 illustrates the question:

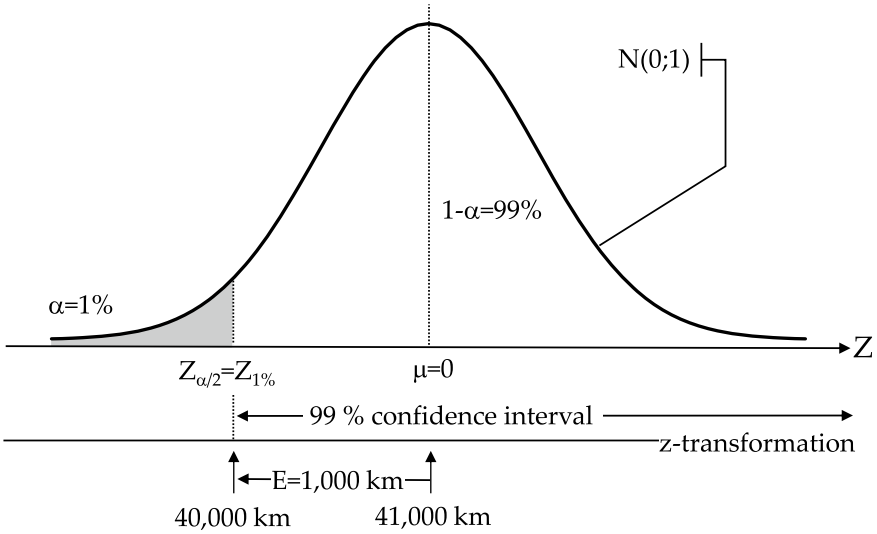


Fig. 8.9 Length of a one-sided confidence interval up to a restricted limit

The calculation is performed by inserting the lower limit of $\mu_{\text{lower}} = 40{,}000$ km, the sample mean $\bar{x} = 41{,}000$ km, and the standard deviation of the sample $S = 8000$ km into the following formula:

$$\mu_{\text{lower}} = \bar{x} - z_{1-\alpha} \cdot \hat{\sigma}_{\bar{x}} = \bar{x} - z_{1-\alpha} \frac{S}{\sqrt{n}} \Rightarrow \frac{\mu_{\text{lower}} - \bar{x}}{z_{1-\alpha}} = -\frac{S}{\sqrt{n}} \Rightarrow n = \left(\frac{S}{\frac{\bar{x} - \mu_{\text{lower}}}{z_{1-\alpha}}} \right)^2 \tag{8.29}$$

After inserting these figures, we arrive at a net sample size of 348 observations:

$$n = \left(-\frac{8000}{\frac{41{,}000 - 40{,}000}{2.33}} \right)^2 = 347.4 \rightarrow 348 \tag{8.30}$$

8.2.3 Confidence Intervals for Proportions

Practical empirical situations often involve the special case in which a confidence interval must be determined for a proportion of the population. Consider small parties in Germany. Because at least 5% of the vote is required to secure representation in the German Bundestag, these parties are very keen on knowing whether election day polling data can produce an accurate one-sided confidence interval

whose lower limit is above the 5% threshold. Economic and business studies frequently investigate average proportions as well. How large is the share of specific products or product lines in a given market segment?

Answering questions like these (and many others) requires datasets with a dichotomous variable. Take the polling example. Say the variable *PartyX* assumes a value of one if a voter chooses party x; in all other cases, the variable receives the value of zero. In the second example, the variable *ProductX* assumes the value of one when a customer purchases product x; in all other cases the variable receives the value of zero. By definition, the average proportions of these variables lie between zero (= 0%) and one (= 100%).

Consider the following example. A company in the food industry wants to know with 95% confidence the market share of a new flavour of yoghurt. The company collects a sample of markets ($n = 300$) in which on average $\overline{p} = 30\%$ of customers decided for this yoghurt. What they want is an interval that covers the actual average market share of π in the population with a confidence level of $(1 - \alpha)$.

As with the confidence interval for the mean, we must first ask, how the average percentage share $\overline{p}$ would be distributed if one drew many samples in succession. For this, we want to make use of two insights from previous sections:

- From Chap. 7 we know that binomial distributions approximate a normal distribution when the condition $n \cdot p \cdot (1 - p) > 9$ is satisfied.
- Moreover, we ascertained in Sect. 8.1 that a normally distributed population always yields a normal sampling distribution of the means.

Hence, provided that $n \cdot p \cdot (1 - p) > 9$, the average percentages $\overline{p}$ of samples have a normal distribution with

$$N\left(\pi; \frac{\sigma}{\sqrt{n}}\right) = N\left(\pi; \sigma_{\overline{p}}\right). \tag{8.31}$$

Analogous to the obtained findings in Sect. 8.2.1 the average proportion π of a population can be inferred by the average proportion $\overline{p}$ of the sample. Accordingly, $(1 - \overline{p})$ is an unbiased estimator for $(1 - \pi)$, so that the standard deviation of a population σ can be estimated by the standard deviation of the sample (see Sect. 7.1.1)

$$S = \sqrt{\overline{p} \cdot (1 - \overline{p})}. \tag{8.32}$$

We can then use the latter to determine the standard error:

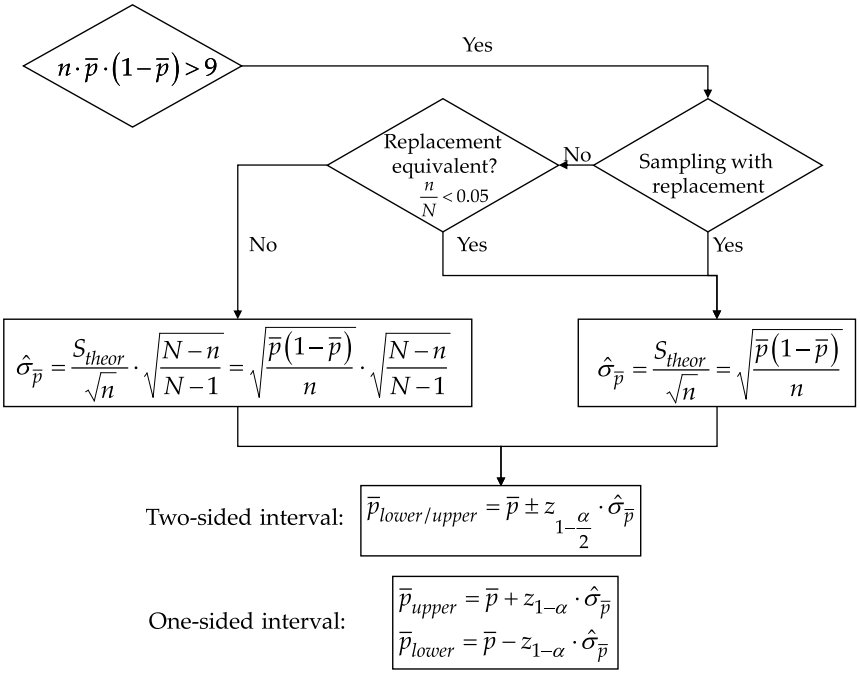
$$\hat{\sigma}_{\overline{p}} = \frac{S}{\sqrt{n}} = \sqrt{\frac{\overline{p}(1 - \overline{p})}{n}} \text{ or} \tag{8.33}$$

$$\hat{\sigma}_{\overline{p}} = \frac{S}{\sqrt{n}} \cdot \sqrt{\frac{N - n}{N - 1}} = \sqrt{\frac{\overline{p}(1 - \overline{p})}{n}} \cdot \sqrt{\frac{N - n}{N - 1}}, \text{ if } n/N \geq 0.05. \tag{8.34}$$

Figure 8.10 illustrates a flow chart for determining confidence intervals for a proportion. For the yoghurt manufacturer example, the chart yields the following calculation of a two-sided confidence interval:

$$\begin{aligned}\overline{p}_{\text{lower/upper}} &= \overline{p} \pm z_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{\overline{p}} = 0.3 \pm 1.96 \cdot \sqrt{\frac{0.3 \cdot (1-0.3)}{300}} \\ &= [0.2481;\, 0.3519] \end{aligned} \tag{8.35}$$

The average proportion in the population is between 24.81 and 35.19% with 95% confidence.



Note:
Two-sided intervals: In two-sided intervals, forms the lower and upper limits of the confidence interval. With a confidence level of (1-α), the average of the population is contained in this interval. **One-side intervals**: To determine the **minimum** average proportion (equal to or greater than a given condition) of the population with a given confidence level of (1-α), form the confidence interval using the lower limit and the upper limit infinite. To determine the **maximum average proportion** (equal to or less than a given condition) of the population with a given confidence level of (1-α), form the confidence interval using the lower limit minus infinite and the upper limit.

Fig. 8.10 Calculating confidence intervals for proportions

8.2.4 Planning Sample Sizes for Proportions

In Sect. 8.2.2 we derived a formula for determining appropriate sample sizes using estimations of metric variable means. When the variable to be estimated is a proportion, this formula can be modified as follows:

$$n = \frac{2^2 \cdot z^2_{1-\frac{\alpha}{2}} \cdot S^2}{E^2} = \frac{2^2 \cdot z^2_{1-\frac{\alpha}{2}} \cdot \overline{p} \cdot (1-\overline{p})}{E^2} \tag{8.36}$$

E represents the target length of the confidence interval. Here too the values assumed for $\overline{p}$ from previous studies are used. Assume that the yoghurt manufacturer from the previous section wants a two-sided confidence interval with a deviation of no more than 2.5 percentage points in either direction with a 95% confidence level. From a previous study, the manufacturer knows the average proportion for $\overline{p} = 30\%$. The confidence interval must have a length of $E = 2 \cdot 2.5$ percentage points $= 5$ percentage points (see Fig. 8.11). The following calculation yields the necessary sample size:

$$n = \frac{2^2 \cdot z^2_{1-\frac{0.05}{2}} \cdot 0.3 \cdot (1-0.3)}{0.05^2} = \frac{4 \cdot 1.96^2 \cdot 0.21}{0.0025} = 1290.76 \rightarrow 1291 \tag{8.37}$$

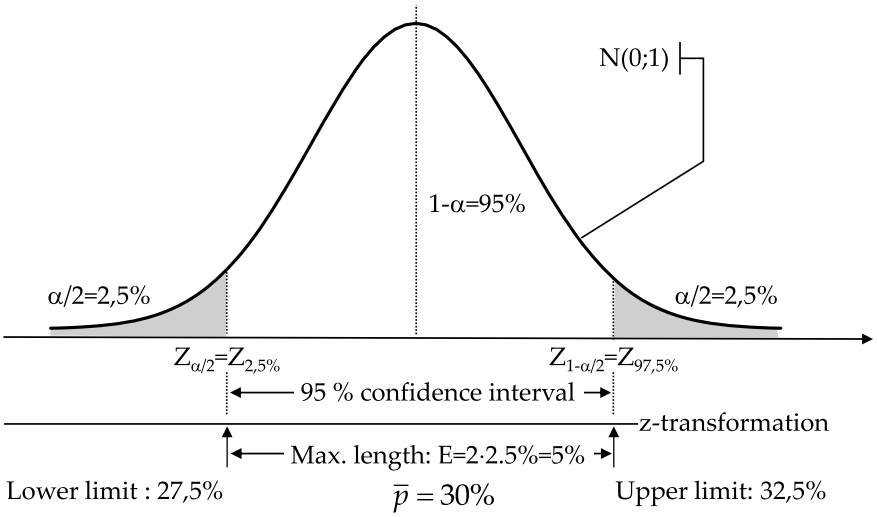


Fig. 8.11 Length of a two-sided confidence interval for a proportion

8.2.5 The Confidence Interval for Variances

Most empirical studies require confidence intervals for mean values or average proportions. In rare cases, they need to determine them for population variance, so we also provide a brief discussion of this topic here. For a population that approximates a normal distribution ($n \geq 30$)

1. the two-sided confidence interval for variance is:

$$P\left(\frac{(n-1)\cdot S^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} \leq \sigma^2 \leq \frac{(n-1)\cdot S^2}{\chi^2_{\frac{\alpha}{2};n-1}}\right) = P\left(\frac{n\cdot S^2_{\text{emp}}}{\chi^2_{1-\frac{\alpha}{2};n-1}} \leq \sigma^2 \leq \frac{n\cdot S^2_{\text{emp}}}{\chi^2_{\frac{\alpha}{2};n-1}}\right) = 1-\alpha \tag{8.38}$$

2. the lower limit of a one-sided confidence interval for variance is:

$$P\left(\frac{(n-1)\cdot S^2}{\chi^2_{1-\alpha;n-1}} = \frac{n\cdot S^2_{\text{emp}}}{\chi^2_{1-\alpha;n-1}} \leq \sigma^2\right) = 1-\alpha \tag{8.39}$$

3. and the upper limit of a one-sided confidence interval for variance is:

$$P\left(\sigma^2 \leq \frac{(n-1)\cdot S^2}{\chi^2_{\alpha;n-1}} = \frac{n\cdot S^2_{\text{emp}}}{\chi^2_{\alpha;n-1}}\right) = 1-\alpha \tag{8.40}$$

In all three cases, calculation requires the sample size n, the variance of the sample and the values from the chi-square table (see Chap. 7) for a given significance level and a given number of degrees of freedom $\left(\chi^2_{\alpha;i}\right)$.

Here is an example of how to apply these formulas. A sample with $n = 20$ is gathered from a normally distributed population. One identifies a variance of the sample of $S^2 = 8$.

(a) Identify a two-sided confidence interval for the variance of the population with a confidence level of 95%.
(b) What is the maximum variance if a confidence level of 95% is assumed?

For part (a), apply the formula for the two-sided interval for variance:

$$\begin{aligned} &P\left(\frac{(n-1)\cdot S^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} \leq \sigma^2 \leq \frac{(n-1)\cdot S^2}{\chi^2_{\frac{\alpha}{2};n-1}}\right) = 1-\alpha \\ &\Rightarrow P\left(\frac{19\cdot 8}{\chi^2_{0.975;19}} \leq \sigma^2 \leq \frac{19\cdot 8}{\chi^2_{0.025;19}}\right) = 0.95 \end{aligned} \tag{8.41}$$

Inserting the values from the chi-square table produces the following result:

$$P\left(\frac{152}{32.852} \leq \sigma^2 \leq \frac{152}{8.907}\right) = 0.95 \Rightarrow P\left(4.63 \leq \sigma^2 \leq 17.07\right) = 0.95 \quad (8.42)$$

Based on the sample we can conclude that the variance of the population lies between 4.63 and 17.07 with 95% confidence.

For part (b), the calculation is:

$$P\left(\sigma^2 \leq \frac{(n-1)\cdot S^2}{\chi^2_{\alpha;n-1}}\right) = 1-\alpha \Rightarrow P\left(\sigma^2 \leq \frac{19\cdot 8}{\chi^2_{0.05;19}}\right) = 0.95 \quad (8.43)$$

From this we get:

$$P\left(\sigma^2 \leq \frac{152}{10.117}\right) = 0.95 \Rightarrow P\left(\sigma^2 \leq 15.02\right) = 0.95 \quad (8.44)$$

Hence, the variance is no more than 15.02 with 95% confidence.

8.2.6 Calculating Confidence Intervals with the Computer

8.2.6.1 Calculating Confidence Intervals with Excel

For calculating confidence intervals, Excel has two pre-programmed functions. *CONFIDENCE.NORM(alpha, standard_dev, size)* and *CONFIDENCE.T(alpha, standard_dev, size)* calculate a confidence interval for the expected value using the normal distribution or the *t*-distribution. After entering the value for α, the standard deviation of the sample, and the sample size, Excel calculates the length of the half confidence band for

$$z_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{\bar{x}}. \quad (8.45)$$

The upper and lower limits of the confidence intervals can be obtained by subtracting this value from the sample mean for the lower limit or adding it to the mean of the sample for the upper limit.

At the Springer website for this book the reader can find a link to a template programmed by the author for calculating one-sided and two-sided confidence intervals. Unlike the integrated routines of SPSS and Stata, the calculation with smaller samples (< 30) is automated using the *t*-distribution. The correction of the sample variance is also automated if the population size N is known and the factor $n/N \geq 0.05$. In addition to confidence intervals for mean values, average proportions, and variances, appropriate sample sizes can be determined for estimating means and proportions.

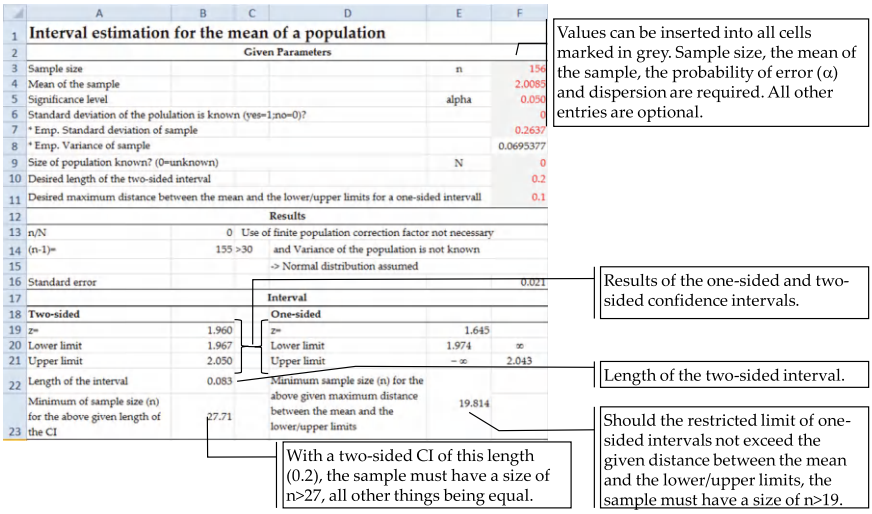
First, let us look closer at an example that determines the confidence interval for a mean value. Open the worksheet *confidence interval mean* in the template.

Enter values in the cells marked in grey. The sample size, the mean value of the sample, the value for α, and the dispersion—either the standard deviation of the population or the standard deviation of the sample—must be indicated. All other entries are optional.

The mean value (= AVERAGE(A2:A157)) and the standard deviation of the sample (= STDEVA(A2:A157)) are calculated for the price variable in *Chapter8_salad_dressing.xls* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_8), and then entered into the template with a value of $\alpha = 0.05$ (see Fig. 8.12). The results of the one-sided and two-sided confidence intervals automatically appear in the bottom half of the template. The standard error is €0.021. Based on the sample, the mean of the price in the population lies between €1.967 and €2.050 with 95% confidence. The length of the two-sided confidence interval is €0.083. For the maximum length of the two-sided interval indicated in the upper part of the template—€0.20—the sample must contain $n > 27.71$ observations, all other parameters being equal.

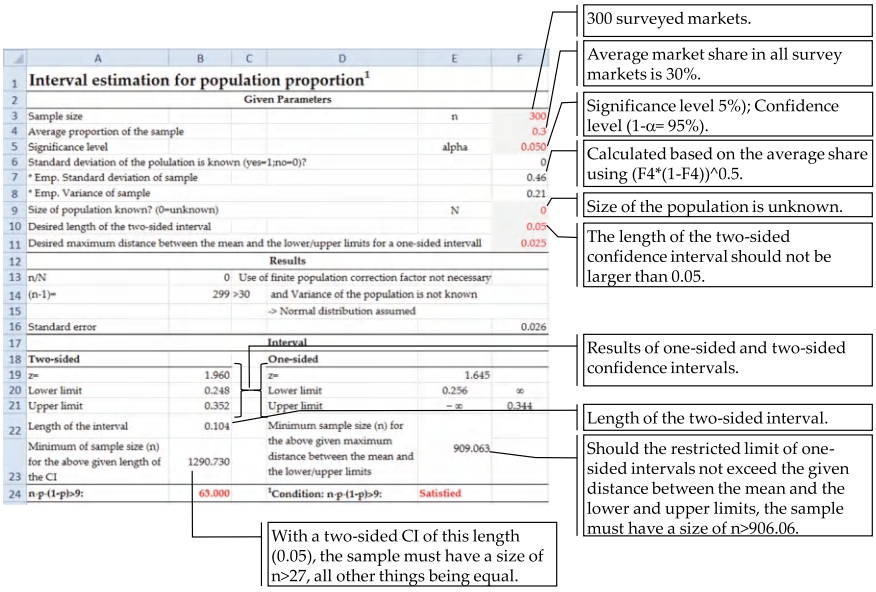
For a given $\alpha = 0.05$ and a one-sided interval

1. the mean of the population is at least €1.97 with 95% confidence and
2. no more than €2.04 with 95% confidence.



Excel formula:
Cell B13: =IF($F$9=0;0;F3/F9); Cell B14: =F3-1; Cell C14: =IF(F3-1>30;">30";"<=30"); Cell F16: =IF(B13<0.05;(F7/((F3-(1-F6))^0.5));(F7/((F3-(1-F6))^0.5))*(((F9-F3)/F9)^0.5)); Cell B19: =IF(($F$3-1)+(30*F6)>30;NORMINV(1-($F5)/2;0;1);TINV(F5;F3-1)); Cell B20: =F4-B19*F16; Cell B21: =F4+B19*F16; Cell B22: =B21-B20; Cell B23: =(2*NORMINV(1-($F5)/2;0;1)*F7/F10)^2+(1-F6); Cell E19: =IF(($F$3-1)+(30*F6)>30;NORMINV(1-($F5)/1;0;1);TINV(F5*2;F3-1)); Cell E20: =F4-E19*F16; Cell F21: =F4+E19*F16; Cell: E22: =(NORMINV(1-($F5);0;1)*F7/F11)^2+(1-F6)

Fig. 8.12 One-sided and two-sided confidence intervals for means with Excel



	A	B	C	D	E	F
1	**Interval estimation for population proportion[1]**					
2	**Given Parameters**					
3	Sample size				n	300
4	Average proportion of the sample					0.3
5	Significance level				alpha	0.05
6	Standard deviation of the polulation is known (yes=1;no=0)?					0
7	* Emp. Standard deviation of sample					0.46
8	* Emp. Variance of sample					0.21
9	Size of population known? (0=unknown)				N	0
10	Desired length of the two-sided interval					0.05
11	Desired maximum distance between the mean and the lower/upper limits for a one-sided intervall					0.025
12	**Results**					
13	n/N	0	Use of finite population correction factor not necessary			
14	(n-1)=	299	>30 and Variance of the population is not known			
15			-> Normal distribution assumed			
16	Standard error					0.026
17				**Interval**		
18	**Two-sided**			**One-sided**		
19	z=	1.960		z=	1.645	
20	Lower limit	0.248		Lower limit	0.256	∞
21	Upper limit	0.352		Upper limit	− ∞	0.344
22	Length of the interval	0.104		Minimum sample size (n) for the above given maximum distance between the mean and the lower/upper limits	909.063	
23	Minimum of sample size (n) for the above given length of the CI	1290.730				
24	**n·p·(1-p)>9:**	63.000		**[1]Condition: n·p·(1-p)>9:**	Satisfied	

Excel formula:
Cell B13: =IF($F$9=0;0;F3/F9); Cell B14: =F3-1; Cell C14: =IF(F3-1>30;">30";"<=30"); Cell F16: =IF(B13<0.05;(F7/((F3-(0-F6))^0.5));(F7/((F3-(0-F6))^0.5))*(((F9-F3)/F9)^0.5)); Cell B19: =IF(($F$3-1)+(30*F6)>30;NORMINV(1-($F5)/1;0;1);TINV(F5*2;F3-1)); Cell B20 =F4-B19*F16; Cell B21: =F4+B19*F16; Cell B22: =B21-B20; Cell B23: =(2*NORMINV(1-($F5)/2;0;1)*F7/F10)^2+(0-F6); Cell E19: =IF(($F$3-1)+(30*F6)>30;NORMINV(1-($F5)/1;0;1);TINV(F5*2;F3-1)); Cell E20: =F4-E19*F16; Cell F21: =F4+E19*F16; Cell: E22: =(NORMINV(1-($F5);0;1)*F7/F11)^2+(0-F6)

Fig. 8.13 One-side and two-sided confidence intervals for proportions with Excel

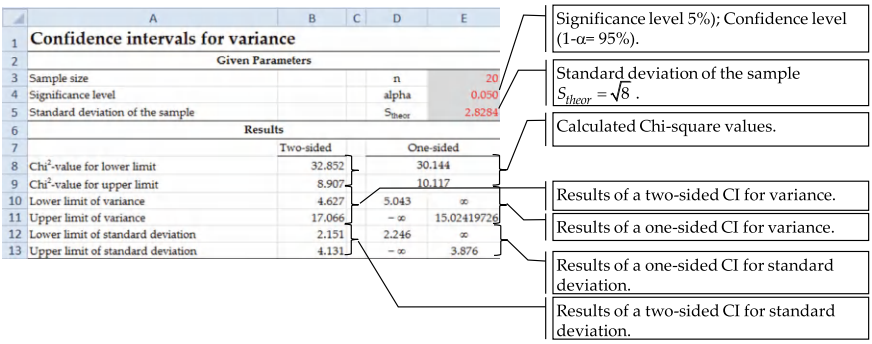
If the restricted limit of one-sided intervals does not exceed the distance between the mean and the lower and upper limits of €0.10 given in the upper part of the template, the sample must have a minimum size of $n > 19.814$ observations, all other parameters being equal.

To determine the confidence interval for proportions, open the worksheet "confidence interval proportion" in the template. Once again enter the values in the cells marked in gray. Figure 8.13 shows the results for the yoghurt manufacturer example in Sects. 8.2.3 and 8.2.4.

Finally, to determine a confidence interval for a variance, open the worksheet "confidence interval variance" in the template. You only need to enter the sample size, the value for α, and the standard deviation of the sample. Figure 8.14 shows the results for the yoghurt manufacturer example in Sects. 8.2.3 and 8.2.4.

8.2.6.2 Calculating Confidence Intervals with SPSS

In SPSS, one-sided or two-sided confidence intervals can be calculated for the mean using the following menu functions. Open the file *Chapter8_salad_dressing.sav* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_8). Select *Analyze* → *Descriptive*



	A	B	C	D	E
1	**Confidence intervals for variance**				
2	**Given Parameters**				
3	Sample size			n	20
4	Significance level			alpha	0.050
5	Standard deviation of the sample			S_{theor}	2.8284
6	**Results**				
7		Two-sided		One-sided	
8	Chi^2-value for lower limit	32.852		30.144	
9	Chi^2-value for upper limit	8.907		10.117	
10	Lower limit of variance	4.627		5.043	∞
11	Upper limit of variance	17.066		− ∞	15.0241972
12	Lower limit of standard deviation	2.151		2.246	∞
13	Upper limit of standard deviation	4.131		− ∞	3.876

Excel formula:
Cell B8: =CHIINV((E4/2);E3-1); Cell B9: =CHIINV(1-(E4/2);E3-1); Cell B10: =(E3-1)*(E5^2)/B8; Cell B11: =(E3-1)*(E5^2)/B9; Cell B12: =B10^0.5; Cell B13: =B11^0.5; Cell D8: =CHIINV((E4);E3-1); Cell D9: =CHIINV(1-(E4);E3-1); Cell D10: =(E3-1)*(E5^2)/D8; Cell E11: =(E3-1)*(E5^2)/D9; Cell D12: =D10^0.5; Cell E13: =E11^0.5

Fig. 8.14 One-side and two-sided confidence intervals for variance with Excel

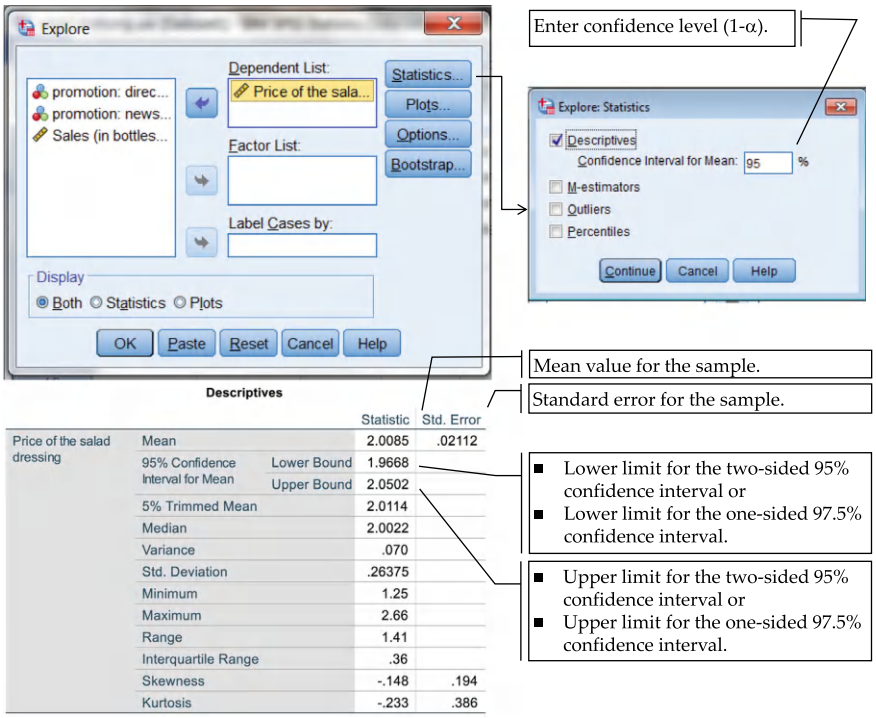
Statistics → *Explore* to open the *Explore* window. Mark the variables for calculating the confidence interval and move to the field *Dependent Lists* by clicking the middle arrow. Next, open the window *Explore: Statistics* by clicking the *Statistics* button, select the field *Descriptives*, and indicate the desired confidence level $(1 - \alpha)$ under *Confidence Interval for Mean.* If, say, you want to calculate a two-sided 95% confidence interval, enter the value 95. If you want to calculate both one-sided 95% confidence intervals, enter the value 90. This is because α is allocated only on one side of the normal curve and on one side of the confidence interval. Since SPSS only calculates two-sided intervals, the value for α must be doubled so that the complete $\alpha = 5\%$ is calculated on the restricted side of a one-sided interval. Generally, with one-sided confidence intervals that have a confidence level of $(1 - \alpha)$ enter the value for $(1 - 2 \cdot \alpha)$. Click *Continue* to close the window. Click *OK* to perform the desired calculation.

Figure 8.15 shows a sample calculated of a two-sided 95% confidence interval for the variable price from the file *Chapter8_salad_dressing.sav*. The sample indicates that the mean of the price in the population lies between €1.97 and €2.05 with 95% confidence. The standard error is €0.02. Both one-sided 97.5% confidence intervals can also be taken from the figure with the altered confidence level:

1. The mean of the population is at least €1.97 with 97.5% confidence.
2. The mean of the population is no higher than €2.05 with 97.5% confidence.

8.2.6.3 Calculating Confidence Intervals with Stata

Stata features two ways to calculate confidence intervals for means and proportions. Open the dataset *Chapter8_salad_dressing.dta* (see supplementary material



Descriptives

			Statistic	Std. Error
Price of the salad dressing	Mean		2.0085	.02112
	95% Confidence Interval for Mean	Lower Bound	1.9668	
		Upper Bound	2.0502	
	5% Trimmed Mean		2.0114	
	Median		2.0022	
	Variance		.070	
	Std. Deviation		.26375	
	Minimum		1.25	
	Maximum		2.66	
	Range		1.41	
	Interquartile Range		.36	
	Skewness		-.148	.194
	Kurtosis		-.233	.386

Fig. 8.15 One-sided and two-sided confidence intervals with SPSS

at DOI: https://doi.org/10.1007/978-3-031-78070-7_8). One assumes that basic parameters such as sample size, mean, proportion, and the standard deviation of the sample are available. To calculate a confidence interval for the mean (see the upper dialogue box in Fig. 8.16 for the example from file *Chapter8_salad_dressing.dta*), select *Statistics* → *Summaries, tables, and tests* → *Summary and descriptive statistics* → *Normal CI Calculator*. To calculate a confidence interval for a proportion (see the lower dialogue box in Fig. 8.16 for the yoghurt manufacturer example from Sect. 8.2.3), select *Statistics* → *Summaries, tables, and tests* → *Summary and descriptive statistics* → *Binomial CI Calculator*. Click *OK* to perform the calculation.

But only in the rarest of situations are the parameters available in advance. For the vast majority of cases, we have to use the *second* technique in Stata for calculating confidence intervals. This technique requires that we use the raw dataset. Here too we will use the *Chapter8_salad_dressing.dta* file as an example. Select *Statistics* → *Summaries, tables, and tests* → *Summary and descriptive statistics* → *Confidence intervals*. Under *Variables: (leave empty for all variables)* indicate the variables in need of a confidence interval (see Fig. 8.17). In our case, this is the metric variable *price*. In the dialogue box under *Confidence level*, enter the desired confidence level ($1 - \alpha$) for a 95% confidence interval. If you want to

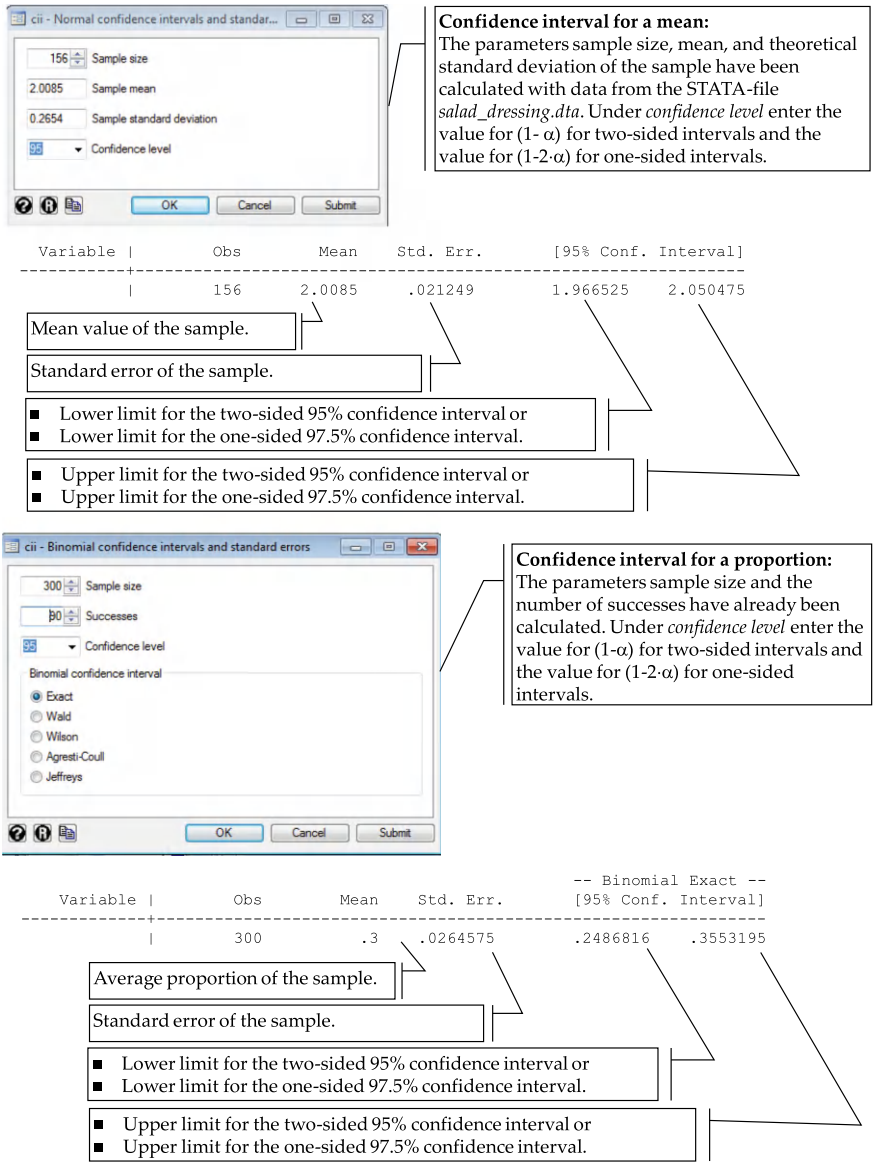


Fig. 8.16 Confidence interval calculation using the Stata CI Calculator

calculate one-sided intervals, enter the value for $(1 - 2 \cdot \alpha)$. The explanation for this is in Sect. 8.2.6.2. Click the button *Continue* to close the dialogue box. Click *OK* to perform the calculation.

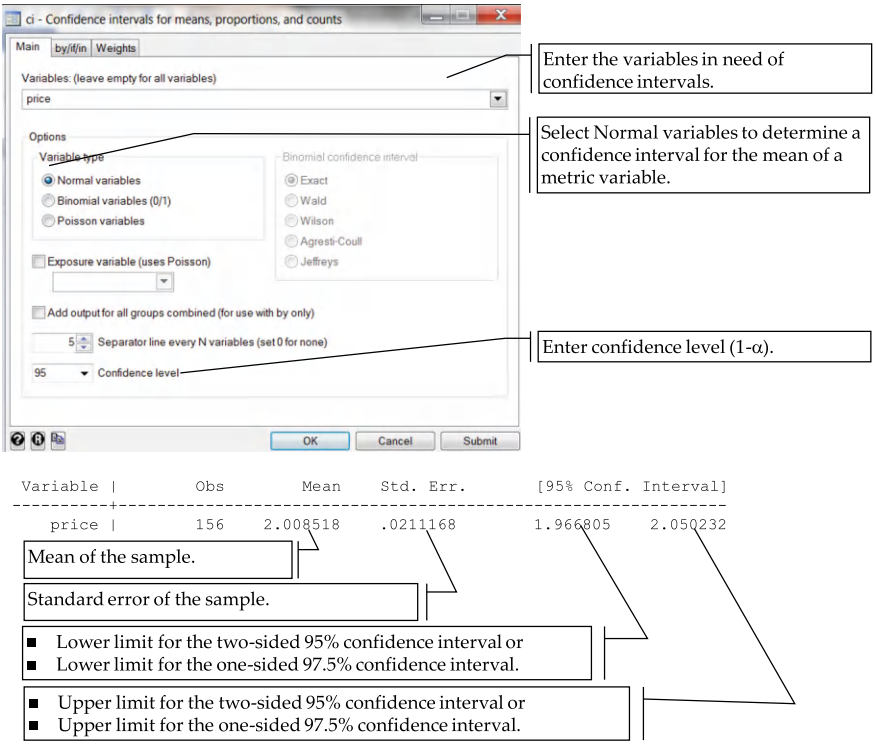


Fig. 8.17 One-sided and two-sided confidence intervals for means with Stata

We interpret the result as we did with SPSS:

1. For two-sided confidence intervals, the mean price in the population is between €1.97 and €2.05 with 95% confidence. The standard error is €0.02.
2. For a one-sided confidence interval restricted on the left side, the mean of the population is at least €1.97 with 97.5% confidence.
3. For a one-sided confidence interval restricted on the right side, the mean of the population is no more than €2.05 with 97.5% confidence.

If the variable that needs a confidence interval is not the metric variable price but the binary variable *prom_new* ("Was the product purchased in week X advertised in the daily newspaper?" Yes = 1; No = 0), the approach would have been similar. You would only have had to select the option *Binomial variables (0/1)* in the dialogue box (see Fig. 8.18). The proportion of weeks with product advertising in a daily newspaper lies between 3.12 and 11.47% with 95% confidence.

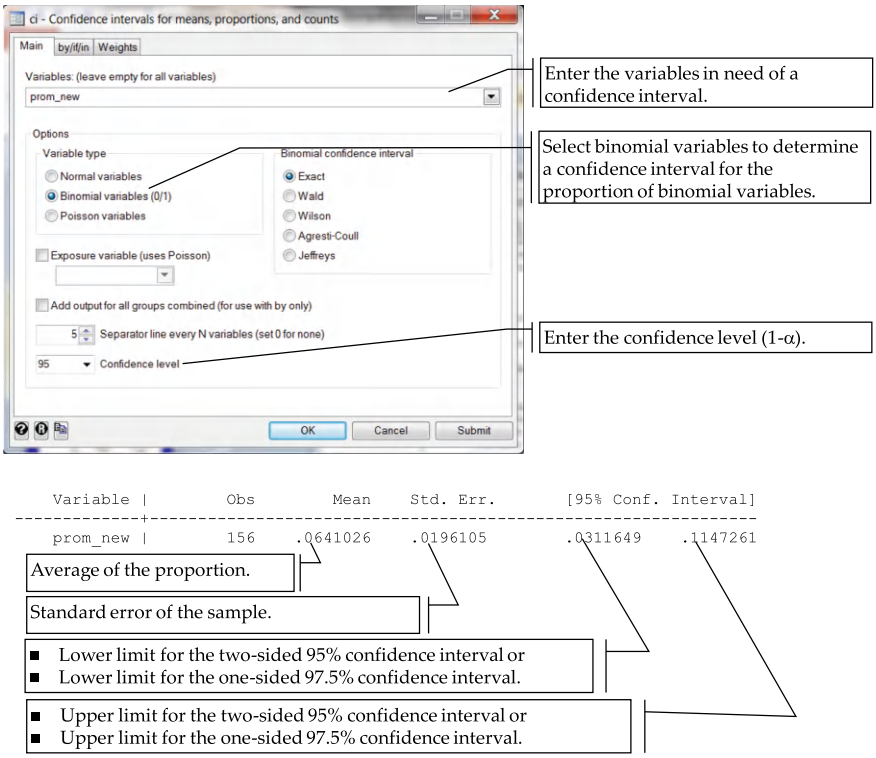


Fig. 8.18 One-sided and two-sided confidence intervals for a proportion value with Stata

8.2.6.4 Calculating Confidence Intervals with R Commander

There are several ways to calculate confidence intervals in R. The simplest way to do so is to use the *t*-test function t.test(). This test is discussed in detail in the next chapter of this book. However, you do not need to know about *t*-tests to use this function to compute confidence intervals. First, load the *Chapter4_salad_dressing.RData* dataset (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_8). Select *Statistics* → *Means* → *Single-sample t-test*. Under *Variables*: indicate the variable in need of a confidence interval (see Fig. 8.19). In our case, this is the metric variable *price*. In the dialogue box under *Confidence level*, enter the desired confidence level $(1 - \alpha)$ for a 95% confidence interval. If you want to calculate one-sided intervals, choose the *Alternative Hypothesis*. Click *OK* to perform the calculation. We interpret the result as we did with SPSS (see Fig. 8.19):

1. For two-sided confidence intervals, the mean price in the population is between €1.97 and €2.05 with 95% confidence. The standard error is €0.02.
2. For a one-sided confidence interval restricted on the left side, the mean of the population is at least €1.97 with 97.5% confidence.

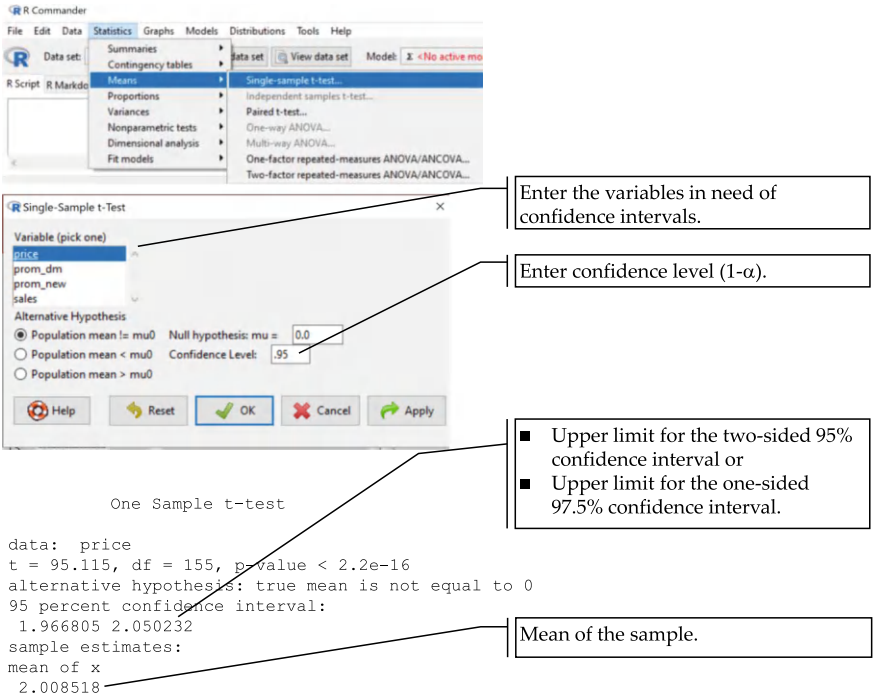


Fig. 8.19 One-sided and two-sided confidence intervals for means with R

3. For a one-sided confidence interval restricted on the right side, the mean of the population is no more than €2.05 with 97.5% confidence.

8.3 Chapter Exercises

Exercise 1

A metal cutter produces metal cuts from a running metal band. The pieces must have a certain length μ, but the cutter is somewhat unreliable. Even when it is set to cut the desired length, fluctuations in the length of the metal band may occur. Assume that the length has a normal distribution.

(a) You take a sample of 17 cuts from the metal band and determine an average value of 150 cm with an empirical variance of $S^2 = 5.76\ \text{cm}^2$. Determine the 80% confidence interval for the average width of the metal band.
(b) What is the smallest possible sample size so that a 90% two-sided confidence interval of the average length of the metal pieces still has a length of 1.2 cm?
(c) What is the smallest variance that still permits us to assume a 90% confidence interval?

Exercise 2

(a) How can we reduce error when using a sample to estimate the parameter of a total population?
(b) How can we improve accuracy when using a sample to estimate μ?

Exercise 3
More than 10,000 customers shop at your store each day. A sample of customers ($n = 20$) taken from the total population with a normal distribution results in an empirical variance of $S^2 = 10^2$ min^2 for the amount of time a customer spends in your store.

(a) Calculate the 99% confidence interval for the empirical variance.
(b) Calculate the two-sided 95% confidence interval for the average time a customer stays in your business. A sample size of $n = 20$ results in an average stay of 55 min.
(c) How large must the sample be so that a two-sided confidence interval of 95% of the average stay is no more than 4 min?

Exercise 4
The company *Tasty Market* wants to find new packaging for its pralines. The management is still unsure about which colour to use. Previously, the pralines were packaged in yellow. Now a blue colour is in discussion.

(a) As part of a market research study in 25 of the 200 *Tasty Market* stores, you must investigate the effect of packaging colour on weekly sales. Assume that sales follow a normal distribution. For the blue packaging, you identify average weekly sales of 9982 packages and an empirical standard deviation of 410 packages. Calculate the 90% confidence interval for the expected sales.
(b) How does the confidence interval change if you select a higher confidence level? Explain your answer.
(c) What is the true empirical standard deviation with a probability of at least 99.5%?

Exercise 5
A car dealer wants to know with 95% confidence what percentage of the 1500 customers who bought a vehicle from him five years ago are still driving it today. In a random sample of 250 customers, it turns out that 207 currently drive a different vehicle.

(a) Identify the correct confidence interval.
(b) What would the confidence interval have to be if the population had been 10,000 customers?
(c) What must the minimum sample size from part b) be, so that the length of the confidence interval is no more than six percentage points?
(d) What is the maximum number of customers who drive the same car with 95% confidence?

(e) How many customers drive the same car with at least 95% confidence if only ten of sixty respondents reported doing so?

8.4 Exercise Solutions

Solution 1

(a) The limits of the confidence interval with unknown variance can be calculated by the following formula:

$$\mu_{\text{upper/lower}} = \overline{x} \pm t_{1-\frac{\alpha}{2}}^{n-1} \hat{\sigma}_{\overline{x}}. \tag{8.46}$$

Because a correction term is not necessary, the assumption of a t-distribution (since $n < 30$) results in the following:

$$\mu_{\text{upper/lower}} = \overline{x} \pm t_{1-\frac{\alpha}{2}}^{n-1} \hat{\sigma}_{\overline{x}} = \overline{x} \pm t_{1-\frac{0.2}{2}}^{17-1} \hat{\sigma}_{\overline{x}} = 150 \pm 1.337 \cdot 0.6 \tag{8.47}$$

$$\rightarrow \mu_{\text{lower}} = 149.1978; \;\; \mu_{\text{upper}} = 150.8022 \tag{8.48}$$

(b) The t-distribution cannot be used to determine the sample size, as the t-values themselves depend on the sample size. For the sake of approximation, a normal distribution can be assumed if the calculated sample size exceeds the value of 30. Hence, for the length of the confidence interval with unknown variance of the total population we get:

$$e = 2 \cdot z_{1-\frac{\alpha}{2}} \sigma_{\overline{x}} = 2 \cdot z_{1-\frac{\alpha}{2}} \frac{S_{\text{emp}}}{\sqrt{n-1}} = 2 \cdot 1.65 \cdot \frac{2.4}{\sqrt{n-1}} = 1.2 \tag{8.49}$$

$$e = \sqrt{n-1}^2 = \left(\frac{2 \cdot 1.65 \cdot 2.4}{1.2} \right)^2 \Rightarrow n = \left(\frac{2 \cdot 1.65 \cdot 2.4}{1.2} \right)^2 + 1$$
$$= 44.56 \approx 45 \tag{8.50}$$

Since the sample size is well over the value of 30, a normal distribution is permissible.

(c) The one-sided confidence interval (the lower limit) can be calculated by the following formula:

$$P\left(\frac{n S_{\text{emp}}^2}{\chi_{1-\alpha;n-1}^2} \leq \sigma^2 \right) = 1 - \alpha \Rightarrow P\left(\frac{17 \cdot 2.4^2}{\chi_{1-0.1;17-1}^2} \leq \sigma^2 \right) = 1 - 0.1$$
$$\Rightarrow P\left(\frac{17 \cdot 2.4^2}{23.542} \leq \sigma^2 \right) = 0.9 \tag{8.51}$$

$$\Rightarrow P\left(4.16 \leq \sigma^2\right) = 0.9 \tag{8.52}$$

Solution 2

(a) By increasing the sample size (increasing n).
(b) By increasing the sample size (increasing n).

Solution 3

(a) The two-sided confidence interval of the variance can be calculated by the following formula:

$$P\left(\frac{nS_{\text{emp}}^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} \leq \sigma^2 \leq \frac{nS_{\text{emp}}^2}{\chi^2_{\frac{\alpha}{2};n-1}}\right) = 1-\alpha$$

$$\Rightarrow P\left(\frac{20\cdot 10^2}{\chi^2_{1-\frac{0.01}{2};20-1}} \leq \sigma^2 \leq \frac{20\cdot 10^2}{\chi^2_{\frac{0.01}{2};20-1}}\right) = 1-0.01 \tag{8.53}$$

$$\Rightarrow P\left(\frac{2000}{38.582} \leq \sigma^2 \leq \frac{2000}{6.844}\right) = 0.99 \Rightarrow P\big(51.837 \leq \sigma^2 \leq 292.228\big) = 0.99 \tag{8.54}$$

(b) The limits of the confidence interval with an unknown variance is given by:

$$\mu_{\text{upper/lower}} = \overline{x} \pm t^{n-1}_{1-\frac{\alpha}{2}}\hat{\sigma}_{\overline{x}}. \tag{8.55}$$

Because the correction term is not necessary, the assumption of the t-distribution (since $n < 30$) results in the following:

$$\mu_{\text{upper/lower}} = \overline{x} \pm t^{n-1}_{1-\frac{\alpha}{2}}\hat{\sigma}_{\overline{x}} = \overline{x} \pm t^{20-1}_{1-\frac{0.05}{2}}\hat{\sigma}_{\overline{x}} = 55 \pm 2.093\cdot 2.2942 \tag{8.56}$$

$$\rightarrow \mu_{\text{lower}} = 50.198;\ \ \mu_{\text{upper}} = 59.802 \tag{8.57}$$

(c) The t-distribution cannot be used to determine the sample size, since the t-values depend on the sample size. For the sake of approximation, the normal distribution can be assumed if the calculated sample size exceeds the value of 30. Hence, for the length of the confidence interval with unknown variance of the total population we get:

$$e = 2\cdot z_{1-\frac{\alpha}{2}}\sigma_{\overline{x}} = 2\cdot z_{1-\frac{\alpha}{2}}\cdot\frac{S_{\text{emp}}}{\sqrt{n-1}} = 4 \tag{8.58}$$

$$\Rightarrow \sqrt{n-1}^2 = \left(\frac{2 \cdot 1.96 \cdot 10}{4}\right)^2 \Rightarrow n = \left(\frac{2 \cdot 1.96 \cdot 10}{4}\right)^2 + 1$$
$$= 97.04 \approx 98 \tag{8.59}$$

Since the sample size is well over 30, the use of a normal distribution is permissible.

Solution 4

(a) The limits of the confidence interval with unknown variance of the total population and $(n-1) \leq 30$ as well as $n/N > 0.05$ can be calculated by the following formula:

$$\mu_{\text{upper/lower}} = \bar{x} \pm t_{1-\frac{\alpha}{2}}^{n-1} \hat{\sigma}_{\bar{x}} \cdot \sqrt{\frac{N-n}{N}} = \bar{x} \pm t_{1-\frac{\alpha}{2}}^{n-1} \frac{S_{\text{emp}}}{\sqrt{n-1}} \cdot \sqrt{\frac{N-n}{N}} \tag{8.60}$$

$$\mu_{\text{upper/lower}} = 9982 \pm 1.711 \frac{410}{\sqrt{25-1}} \cdot \sqrt{\frac{200-25}{200}} \tag{8.61}$$

$\rightarrow \mu_{\text{lower}} = 9848.06$; $\mu_{\text{upper}} = 10{,}115.94$ (Rounding can lead to minor deviations.)

(b) The interval is longer. The reason: the t-value increases and larger confidence requires a greater field of action.

The lower limit of the confidence interval of empirical variance can be calculated as follows:

$$P\left(\frac{nS_{\text{emp}}^2}{\chi_{1-\alpha;n-1}^2} \leq \sigma^2\right) = 1 - \alpha \tag{8.62}$$

$$\rightarrow P\left(\frac{25 \cdot 410^2}{\chi_{99.5\%;n-1}^2} \leq \sigma^2\right) = 99.5\% \rightarrow P\left(92{,}244.014 \leq \sigma^2\right) = 99.5\% \tag{8.63}$$

(c) For the standard deviation we arrive:

$$P\left(\sqrt{92{,}244.014} \leq \sqrt{\sigma^2}\right) = P(303.717494 \leq \sigma) = 99.5\% \tag{8.64}$$

Solution 5
The share of car owners who are still driving the same car today is:

$$\overline{p} = 1 - \frac{207}{250} = 0.172 \tag{8.65}$$

(a) Since $n/\mathrm{N} = 0.167 > 0.05$, the confidence interval can be calculated as follows:

$$\begin{aligned}\mu_{\text{lower/upper}} &= \overline{p} \pm z_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{\overline{p}} \cdot \sqrt{\frac{N-n}{N}} \\ &= 0.172 \pm 1.96 \cdot \sqrt{\frac{0.172 \cdot (1-0.172)}{250}} \cdot \sqrt{\frac{1500-250}{1500-1}}\end{aligned} \tag{8.66}$$

The average share is between 12.9 and 21.5% of the population with 95% confidence.

(b) The confidence interval is calculated without the correction term as follows:

$$\mu_{\text{lower/upper}} = \overline{p} \pm z_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{\overline{p}} = 0.172 \pm 1.96 \cdot \sqrt{\frac{0.172 \cdot (1-0.172)}{250}} \tag{8.67}$$

The average percentage of the population lies between 12.5 and 21.9% with 95% confidence.

(c) The minimum size of the sample would have to be $n = 608$:

$$\begin{aligned}n &= \frac{2^2 \cdot z_{1-\frac{\alpha}{2}}^2 \cdot \overline{p} \cdot (1-\overline{p})}{E^2} = \frac{2^2 \cdot 1.96^2 \cdot 0.172 \cdot (1-0.172)}{0.06^2} \\ &= 607.9 \rightarrow 608\end{aligned} \tag{8.68}$$

(d) The maximum proportion is 21.1% of customers:

$$\mu_{\text{upper}} = \overline{p} + z_{1-\alpha} \cdot \hat{\sigma}_{\overline{p}} = 0.172 + 1.96 \cdot \sqrt{\frac{0.172 \cdot (1-0.172)}{250}} = 21.1\% \tag{8.69}$$

(e) Since the condition $n \cdot \overline{p} \cdot (1-\overline{p}) > 9$ does not hold, it is impossible to calculate the confidence interface by approximation through the normal distribution.

References

Fowler, F. J. (2002). *Survey research methods* (3rd ed.). Sage.
Lewin, C. (2005). Elementary quantitative methods. In B. Somekh & C. Lewin (Eds.), *Research methods in the social sciences* (pp. 215–225). Sage.
Oppenheim, A. N. (1992). *Questionnaire design, interviewing and attitude measurement, continuum.* Wiley.



Hypothesis Testing

9

9.1 Fundamentals of Hypothesis Testing

Among the most important techniques in statistics is hypothesis testing. A hypothesis is a supposition about a certain state of affairs. It does not spring from a sudden epiphany or a long-standing conviction; rather, it offers a testable explanation of a specific phenomenon. A hypothesis is something that we can accept (verify) or reject (falsify) based on empirical data.

Intuitively, one might think that the best way to prove a hypothesis is to try to verify it. But the twentieth-century philosopher of science Karl Popper argued that falsification is the better approach. In his book *The Logic of Scientific Discovery*, Popper (1934) asked how we can empirically prove the hypothesis that all swans are white. We can go out and start taking note of all the swans we observe. We can talk with people in other regions and countries about the swans they have seen. But according to Popper even if every swan we observe or hear about is white, it is impossible for us to know for certain that no black swans exist. By contrast, if we observed a single black swan we would know for certain that the hypothesis "all swans are white" is false. In other words, no amount of empirical testing can establish a hypothesis as true, but it can establish a hypothesis as false.

In statistical hypothesis testing, as well, falsification plays a central role. In their pathbreaking work, Neyman and Pearson (1928a, 1928b) and Neyman (1937) developed a testing model consisting of a null hypothesis H_0 and an alternative hypothesis H_1. The null hypothesis H_0 is statistically tested based on the results of a random sample to gain information about the population given a maximum probability error of α. This comparison can yield two results:

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78070-7_9.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*,
Springer Texts in Business and Economics,
https://doi.org/10.1007/978-3-031-78070-7_9

1. The sample falsifies the null hypothesis H_0. By finding the proverbial black swan in our sample, we can assume with a maximum probability error of α that H_1 is true. Usually, H_1 hypotheses are about relationships between variables or differences between groups.
2. The sample does not falsify the null hypothesis H_0, i.e. one finds no black swans in the sample, only white ones. As Popper argued, this does not prove the validity of H_0. It only permits us to conclude that H_0 has not been falsified.

Unlike Popper's swans, the unique feature of statistical testing is that the decision for H_1 is always accompanied by a maximum probability error α determined by the investigator. While the existence of a black swan disproves the null hypothesis that all swans are white (H_0) with 100% certainty and a probability error of $\alpha = 0\%$, the assumption of H_1 in statistical testing comes with a confidence level of $(1 - \alpha)$, which in practice is always less than 100% (though frequently close). Strictly speaking, there is no such thing as a statistical proof in empirical research. We only have findings that support a hypothesis or that force us to reject the null hypothesis. It is especially important, therefore, that investigators disclose the assumptions they make about, say, the confidence level $(1 - \alpha)$ and the maximum probability of error of α.

This brings us to the subject of decision risk. It might first appear that by setting the probability error of α close to zero, one could successfully manage the risk of a wrong decision. But this is not the case. Though the α error—frequently called a *type I error* or the significance level of a test—sets an upper limit for falsely rejecting H_0, this error requires the acceptance of H_1. Likewise, the decision not to reject H_0 can also be true or false. The risk of failing to reject a false H_0 is called a *type II error*, or β error. Each decision carries its own risks. If one rejects H_0, then the maximum probability that this decision is in error is α. If one does not reject H_0, then the maximum probability that this decision is in error is β. Figure 9.1 shows the relationship between test decisions and reality. The probability that an investigator will correctly reject H_0 refers to the power of a test (power $= (1 - \beta)$).

Let us consider an example that illustrates the difference between the two error types. A student must take an HIV test as a part of his visa application for a study abroad programme. After three weeks he finds out that he tested positive. When he asks about the accuracy of the test, the doctor who administered it—a general practitioner not well-versed in statistics—reads in the instruction leaflet that the test accurately identifies infected patients in $(1 - \alpha) = 99.99\%$ of cases. The likelihood of the student's having HIV, the doctor reasons, must also be 99.99%.

		Actual Situation	
		H_0 is true [Patient is infected with HIV]	H_0 is false [Patient is not infected with HIV]
Statistical Decision	Do not reject H_0 [Positive HIV Diagnosis]	**Confidence (1-α) / No Error** *[Likelihood of producing a positive HIV diagnosis when the patient is infected (1-α)=99.99%]*	**Type II Error (β)** *[Likelihood of producing a positive HIV diagnosis when the patient is not infected]*
	Reject H_0 / decision for H_1 [Negative HIV Diagnosis]	**Type I Error (α)** *[Likelihood of producing a negative HIV diagnosis when the patient is infected (α=0.01%)]*	**Power (1-β) / No Error** *[Likelihood of producing a negative HIV diagnosis when the patient is not infected]*

Fig. 9.1 Probabilities of error for hypotheses testing

Highly distressed by the news, the student decides to get a second opinion at a local HIV centre. There a counsellor explains that the significance level α describes the probability of a false negative, i.e. producing a negative HIV diagnosis (H_1) when the patient is infected with HIV (H_0), so that

$$\alpha = P(H_1 \text{ is assumed} | H_0 \text{ is true}) = 0.01\%. \tag{9.1}$$

Since the student tested positive for HIV (H_0), the α error can no longer occur, just the β error, but the instruction leaflet that comes with the test does not mention the size of the type II error (β). The counsellor points out that while the significance level α of the diagnostic HIV test is 0.01%, the β is 50%! This means there is a 50% likelihood of producing a positive HIV diagnosis when the patient is not infected (known as a false positive). To reduce his uncertainty significantly, he would have to take another HIV test: the probability of being infected with HIV after testing positive two times in succession is around 99.8%.

What is the relationship between the size of an α-error and the size of a β-error? The size of the α-error does not provide direct information about the risk of failing to reject an H_0 that is false. Nevertheless, the level of the α-error automatically defines the level of the β-error. For example, Fig. 9.2 shows blood test results from sick and healthy patients. We see that the level of a certain marker in sick patients is higher than that in healthy patients, hence the right-shift in the results of the former. But some healthy patients have elevated levels of this marker as well, which is why the curves partly overlap. A line indicating the threshold above which sick patients are diagnosed is located in this area of intersection. As a result, some healthy patients will receive an erroneous diagnosis (β-error), and some sick patients will go undiagnosed (α-error). We could reduce the α-error by shifting the diagnosis threshold to the left, but this would automatically increase the β-error. The smaller the α-error is, the larger the β-error is, and vice versa. There is no simple way to reduce both risks at the same time. One approach is to increase the

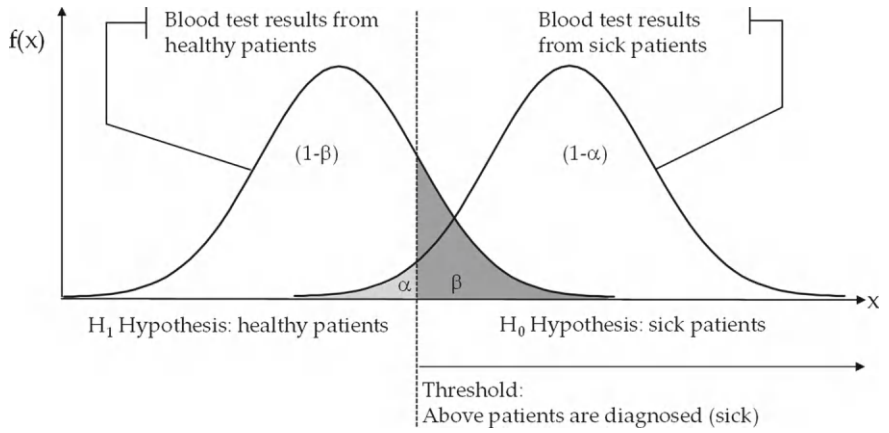


Fig. 9.2 Error probabilities for diagnosing a disease

sample size. This improves the accuracy of the estimate, so that the area in which the distributions overlap becomes smaller.

Now that we have defined the hypotheses, we must decide which statistical method to apply. Ultimately, the decision depends on the type of research one is conducting. If the objective is to discover a significant relationship between two variables, we will need to test *relationship hypotheses*. Say we want to find out if there is a relationship between the variables gender and income. First, we define H_0 and H_1 as follows:

H_0: There is no correlation between gender and income.
H_1: There is a correlation between gender and income.

Some statistical software provides *p*-values (see Sect. 9.2.3), which indicate the exact probability of error when assuming a relationship between variables. In practice, however, hypotheses are more commonly designed to test differences, so we will concentrate on these below.

Generally, one must distinguish between tests for dependent samples and tests for independent samples. The first type of test tells us whether the values of two or more variables (characteristics) differ. The individual values of the two or more variables describe different characteristics of a person or an object. If two variables are involved, one speaks of paired samples. A typical example of a paired sample is when measuring peoples' preferences before and after watching an ad for a certain product. In this case, two paired values exist for each person (i.e. for each row) in the dataset. The interesting question is whether the average values of the variables differ significantly (left table in Fig. 9.3). In this case, we define H_0 and H_1 as follows:

H_0: The preferences of the target group do not change after watching the ad.
H_1: The preferences of the target group change after watching the ad.

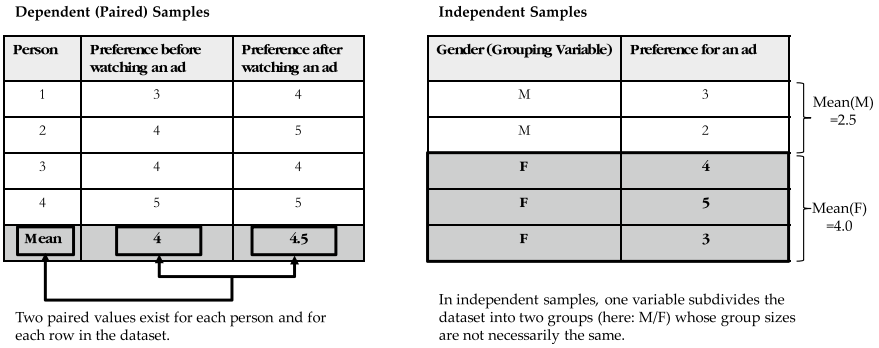


Fig. 9.3 The data structure of independent and dependent samples

A test for independent samples can tell us whether, say, the mean values of the groups differ. In independent samples, one variable subdivides the dataset into two (or more) groups whose group sizes are not necessarily the same. At the same time, group-specific mean values for a second variable are calculated and checked for significant differences. Such a procedure is called an independent test because the total population can be separated into independent subgroups using grouping variables. With an independent test, we can check whether, say, men and women have different preferences for an ad (right table in Fig. 9.3). Here we define H_0 and H_1 as follows:

H_0: The preferences for an ad do not differ between men and women.
H_1: The preferences for an ad differ between men and women.

Data that satisfy certain distribution assumptions—most commonly those of a normal distribution—permit the use of parametric tests. Data that do not, require us to employ nonparametric tests. If the data are nominally or ordinally scaled, then in most cases only nonparametric tests are applicable.

In sum, then, there is no universally valid answer to the question of which test to use. The test depends on the objective, on the scale, and on the distribution of the data. The most important tests investigate whether parameters for central tendencies differ. The most important values for describing the central tendency are the arithmetic mean, the modal value, and the median. Which test should be used when is shown in Fig. 9.4. The tests will be discussed individually in the following sections.

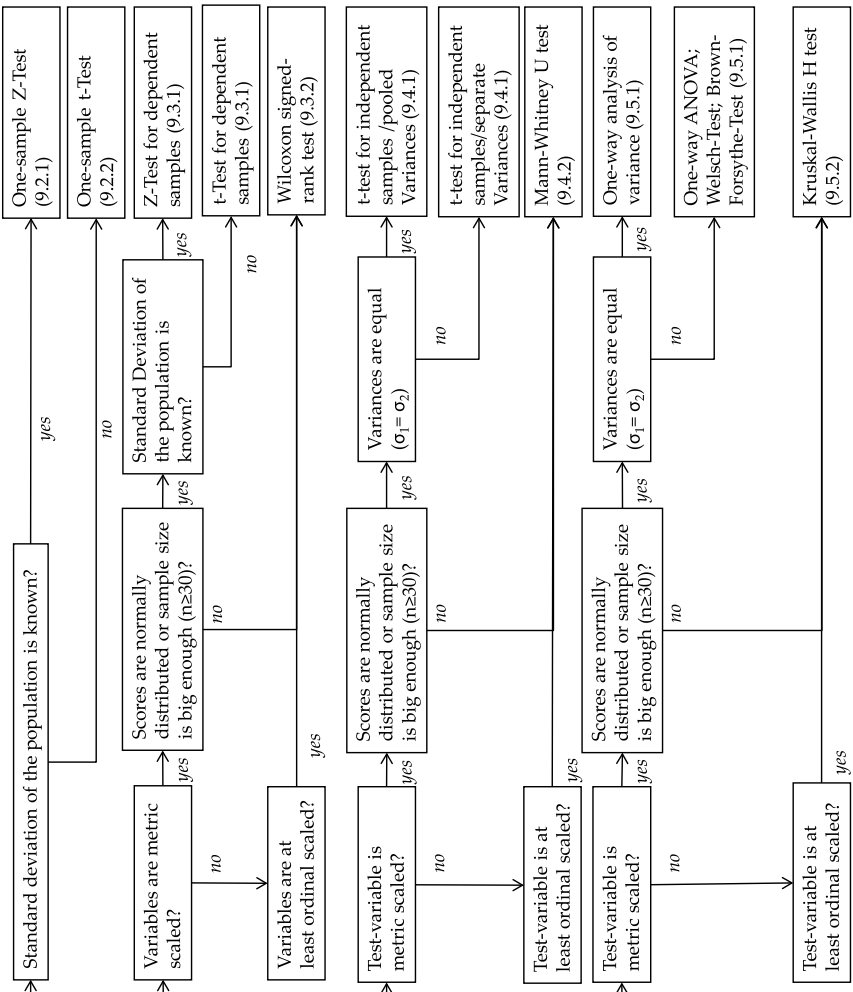


Fig. 9.4 Tests for comparing the parameters of central tendency

9.2 One-Sample Tests

9.2.1 One-Sample *Z*-Test (When σ Is Known)

Using the one-sample *Z*-test we want to discuss the steps of a hypothesis test in more detail. Let us first consider the following example. A tyre manufacturer decided to change the rubber formulation of its tyres. The previous model had an average lifespan of $\mu_0 = 40{,}000$ km with a standard deviation of $\sigma_0 = 2000$ km. The question is whether the new rubber formulation significantly changed the average lifespan of the tyres.

Step 1: Formulate the Hypotheses
In a first step we must determine whether the question is about a two-tailed or one-tailed hypothesis test. Since the lifespan may have either increased or decreased, the test is two-tailed. The hypotheses are as follows:

H_0: $\mu = \mu_0 = 40{,}000$ km
H_1: $\mu \neq \mu_0 = 40{,}000$ km

If the question was whether the lifespan increased significantly, we would need a one-tailed test. The hypotheses would then be:

H_0: $\mu \leq \mu_0 = 40{,}000$ km
H_1: $\mu > \mu_0 = 40{,}000$ km

Following Popper's idea, the hypothesis we want to prove—that lifespan increases—is formulated in the alternative hypothesis H_1.

To test the opposite case—that lifespan decreases—the hypotheses would have to look like this:

H_0: $\mu \geq \mu_0 = 40{,}000$ km
H_1: $\mu < \mu_0 = 40{,}000$ km

Step 2: Determine the Significance Level α

Next, we must define the significance level α, i.e. the maximum permissible probability that we reject an H_0 that is true. The significance level α is usually set at 1, 5, or 10%, though 5% is the most common value, which is what we will use here ($\alpha = 0.05$).

Step 3: Draw a Sample
In this example we draw a random sample of $n = 100$ observations.

Step 4: Check the Test Requirements

• Interval or ratio scale of measurement (approximately interval)	✓
• Random sampling from a defined population	✓
• Characteristic is normally distributed in the population or sample size is big ($n - 1 > 30$)	✓
• Variance of the population is known	✓

Step 5: Determine the Critical Test Values
If it is true that H_0: $\mu = \mu_0 = 40{,}000$, then the mean of the sample must lie within the following interval with a confidence level of $(1 - \alpha)$:

$$P\left(c_{\text{lower}} \leq \overline{x} \leq c_{\text{upper}}\right) = 1 - \alpha. \tag{9.2}$$

As Fig. 9.5 shows, the upper and lower limits of this interval correspond to the critical values at the edges of the normal distribution. If we standardize the distribution, the quantiles $(-\ z_{(1-\alpha/2)})$ and $(+\ z_{(1-\alpha/2)})$ represent the upper and lower limits of the transformed Z value:

$$P\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\overline{x} - \mu_0}{\sigma_{\overline{x}}} \leq +z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha. \tag{9.3}$$

Through a simple conversion we get:

$$P\left(\mu_0 - z_{1-\frac{\alpha}{2}}\sigma_{\overline{x}} \leq \overline{x} \leq \mu_0 + z_{1-\frac{\alpha}{2}}\sigma_{\overline{x}}\right) = 1 - \alpha. \tag{9.4}$$

Now we insert the known values from the hypothesis and the sample:

- μ_0, the hypothetical value assumed in H_0 ($\mu_0 = 40{,}000$);
- $\sigma_{\overline{x}}$, the standard error that results from the standard deviation of the population:

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2000}{\sqrt{100}}; \tag{9.5}$$

- the value for the probability error of α ($\alpha = 5\%$). Because the test is two-tailed, the probability error of α is equally distributed on the right and left sides of the distribution, so that:

$$P\left(40{,}000 - z_{0.975}\frac{2000}{\sqrt{100}} \leq \overline{x} \leq 40{,}000 + z_{0.975}\frac{2000}{\sqrt{100}}\right) = 0.95 \tag{9.6}$$

$$P\left(40{,}000 - 1.96\frac{2000}{\sqrt{100}} \leq \overline{x} \leq 40{,}000 + 1.96\frac{2000}{\sqrt{100}}\right) = 0.95 \tag{9.7}$$

$$P(39{,}608 \leq \overline{x} \leq 40{,}392) = 0.95. \tag{9.8}$$

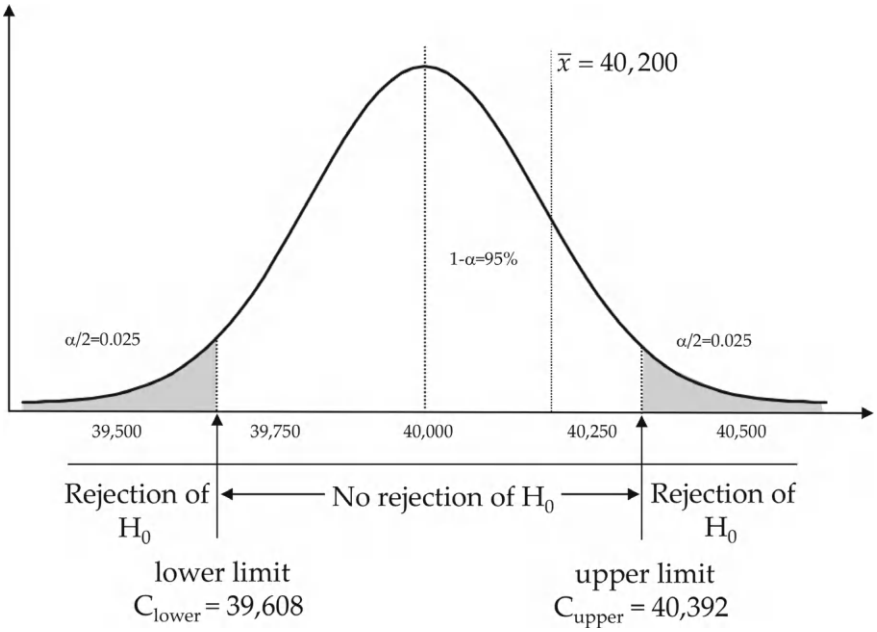


Fig. 9.5 Rejection regions for H_0

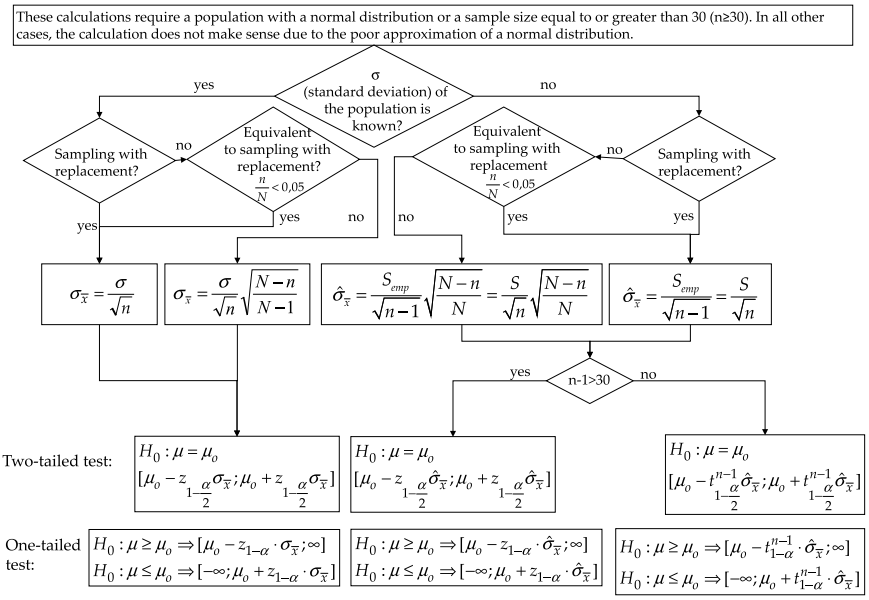
Step 6: Determine the Empirical Test Value
The average life of the tyres in this sample is $\bar{x} = 40{,}200$ km and is assumed to have a normal distribution.

Step 7: Make a Test Decision
If H_0 is true—i.e. if the mean of the population is $\mu_0 = 40{,}000$—then in 95% of cases the sample mean lies between 39,608 and 40,392 km. Accordingly, the rejection regions for H_0 are the intervals $[-\infty; 39{,}608]$ and $[40{,}392; \infty]$. Now, say the sample we draw results in an $\bar{x}$ value of 40,200. This means that we cannot reject H_0.

Figure 9.6 provides a flow chart summarizing the calculation of one-sample tests under various conditions. Again, one must carefully distinguish between cases where the standard deviation of the population is known from the beginning (left branch in Fig. 9.6) and those where it is not (right branch in Fig. 9.6). The first is known as a *Z*-test; the second is called a *t*-test. We used the former in our example.

9.2.2 One-Sample *t*-Test (When σ Is Not Known)

To understand the *t*-test, let us once again consider an example. A research team has developed a new propulsion system for heavy-duty trucks that is designed to use less diesel fuel than conventional engines. To calculate real-life fuel savings,



Note:
If the mean value from the sample lies within the given interval, H_0 cannot be rejected. The size of the β error (failing to reject H_0 when it is false) is unknown. If the mean value from the sample lies outside the given interval, H_0 can be rejected with a probability error of α (or with a confidence of 1-α). The actual mean value of the population differs significantly from the assumed value μ_0.

Fig. 9.6 The one-sample *Z*-test and the one-sample *t*-test

the researchers test 20 vehicles outfitted with the new system. It turns out that the trucks consume on average 15 L of diesel for every 100 km they travel, with a variance of $S^2 = 25\ \mathrm{L}^2$. In contrast to the example from Sect. 9.2.2, the investigators here know the variance of the sample but not the variance of the population, which is why they have to run a *t*-test and not a *Z*-test. Fuel consumption can be assumed to follow a normal distribution. The company's chief executives will introduce the new technology only if they can assume with 99.5% confidence that fuel consumption will be less than with the old technology. Vehicles with the old technology use on average 17.5 L for every 100 km they travel. The research team now wants to determine whether they satisfy the condition stipulated by the chief executives. The steps the researchers follow are similar to the *Z*-test:

Step 1: Formulate the Hypotheses
The research team decides for a one-tailed test because the company's chief executives plan to introduce the new technology only if it uses demonstrably less fuel than the old one. They must formulate the hypothesis so that the α error represents the gravest error in the decision-making process. The possible hypothesis combinations are:

$$H_0\colon \mu < \mu_0 = 17.5\ \mathrm{L} \text{ and } H_1\colon \mu \geq \mu_0 = 17.5\ \mathrm{L}$$

In this case, an α error means keeping the old technology (H_1: $\mu \geq \mu_0 = 17.5$ L) even though the new one is more efficient (H_0: $\mu < \mu_0 = 17.5$).

H_0: $\mu \geq \mu_0 = 17.5$ L and H_1: $\mu < \mu_0 = 17.5$ L

Here, an α error means introducing a new technology (H_1: $\mu < \mu_0 = 17.5$ L) even though the new one is not more efficient (H_0: $\mu \geq \mu_0 = 17.5$).

From the perspective of the executives, the second α error is more serious, because in the best case the new technology is just as good as the old one. Executives will thus want to control for this error and determine its probability with α. The research team must, therefore, examine the following:

H_0: $\mu \geq \mu_0 = 17.5$ L and H_1: $\mu < \mu_0 = 17.5$ L

Step 2: Determine the Significance Level α

The company's senior management has decided that $\alpha = 0.5\%$.

Step 3: Draw a Sample

The researchers draw a sample of $n = 20$. The data approximately follow a normal distribution.

Step 4: Check the Test Requirements

• Interval or ratio scale of measurement (approximately interval)	✓
• Random sampling from a defined population	✓
• Characteristic is normally distributed in the population or sample size is big ($n - 1 > 30$)	✓
• Variance of the population is unknown	✓

Step 5: Determine the Critical Test Values

The calculation of the one-sample *t*-test is shown in Fig. 9.6. The variance of the population is unknown and the data approximately follow a normal distribution. The sampling is done without replacement, but the size of the population N must be assumed to be very large so that the standard error can be calculated as follows on the basis of $\frac{n}{N} < 0.05$:

$$\hat{\sigma}_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{5}{\sqrt{20}} = 1.118 \tag{9.9}$$

Since $(n - 1)$ is no larger than 30, H_0 cannot be rejected if $\bar{x}$ lies within the interval:

$$\left[\mu_0 - t_{1-\alpha}^{n-1}\hat{\sigma}_{\bar{x}}; \infty\right]. \tag{9.10}$$

After inserting and calculating these values, we get:

$$\left[17.5 - t_{99.5}^{19} \cdot 1.118; \infty\right] \tag{9.11}$$

$$= [17.5 - 2.861 \cdot 1.118; \infty] \tag{9.12}$$

$$= [14.3; \infty]. \tag{9.13}$$

Step 6: Determine the Empirical Test Value
The sample produces an average value of $\overline{x} = 15$ L (test statistic) with a variance of $S^2 = 25\ \mathrm{L}^2$.

Step 7: Make a Test Decision
Assuming a confidence level of $(1 - \alpha) = 99.5\%$, we cannot reject H_0: $\mu \geq \mu_0 = 17.5$ L because the mean value of the sample ($\overline{x} = 15$ L) lies in a region that does not permit a rejection of H_0. Statistically, the test does not prove that the new propulsion system uses less than 17.5 L per 100 km.

9.2.3 Probability Value (*p*-Value)

At this stage we can now ask, what is the exact probability of erroneously assuming H_1? This is called the probability value, or *p*-value, and it describes the likelihood of the empirical test value—in this case, $\overline{x}$—under the assumption of the null hypothesis. For a two-tailed hypothesis test, the *p*-value is calculated by

$$p = 2 \cdot \left(1 - P\left(t^{n-1} \leq t_{\text{critical}} = \left|\frac{\overline{x} - \mu_0}{\sigma_{\overline{x}}}\right|\right)\right) \tag{9.14}$$

For the one-tailed hypotheses, use the following for the *p*-values:

$$p_{\text{left}} = P\left(t^{n-1} \leq t_{\text{critical}} = \frac{\overline{x} - \mu_0}{\sigma_{\overline{x}}}\right) \quad \text{for } H_1 : \mu < \mu_0 \tag{9.15}$$

$$p_{\text{right}} = \left(1 - P\left(t^{n-1} \leq t_{\text{critical}} = \frac{\overline{x} - \mu_0}{\sigma_{\overline{x}}}\right)\right) \quad \text{for } H_1 : \mu > \mu_0. \tag{9.16}$$

For our one-tailed example, this means:

$$p_{\text{left}} = P\left(t^{19} \leq t_{\text{critical}} = \frac{15 - 17.5}{1.118}\right) \tag{9.17}$$

$$p_{\text{left}} = P\left(t^{19} \leq t_{\text{critical}} = -2.236\right) \tag{9.18}$$

The t-table in the appendix (see Appendix C) produces for $n = 19$ degrees of freedom and a critical value of $t_{\text{critical}} = -2.236$ a p-value that lies between $(1 - 0.975) = 2.5\%$ and $(1 - 0.99) = 1\%$. With Excel, we can determine the exact value using the command $= 1 - T.DIST(-2.236; 19; 1)$. For the p-value of the left-tailed test, we get:

$$p_{\text{left}} = P\left(t^{19} \leq t_{\text{critical}} = -2.236\right) = 1.88\% \tag{9.19}$$

This means that one errs in $p = 1.88\%$ of cases when assuming that the new propulsion system is more efficient (H_1: $\mu < \mu_0 = 17.5$ L). This means that we must reject H_0 if the significance level has been set at 0.05. In contrast, we cannot reject H_0 if the significance level was fixed below 1.88% (see example above with $\alpha = 0.5\%$). Statistical software packages indicate the p-value in their results. If the p-value is smaller than the significance level α, we reject H_0, and vice versa.

9.2.4 One-Sample *t*-Test with SPSS, Stata, R, and Excel

A market researcher would like to say with 95% confidence whether the weekly sales figures of three companies differ significantly from €40,000. He collects sales data for 20 weeks. This data (see right part of Fig. 9.7) is now presented in four files: an R data file (*Chapter9_t_test.RData*), an SPSS file (*Chapter9_t_test.sav*), a Stata file (*Chapter9_t_test.dta*), and an Excel file (*Chapter9_t_test.xls*) (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9).

In SPSS, select *Analyze* → *Compare Means* → *One-Sample T Test*… to open the window in Fig. 9.7 on the left. Under *Test Value*, enter 40,000. Under *Test Variable(s)*, indicate the test variables—in our case, the product sales of the three companies. SPSS does not indicate the significance of the one-tailed test. When testing $\bar{x} < \mu_0$ and H_1: $\mu < \mu_0$, we must halve the two-tailed significance value. If we were to ask whether, say, the sales of company two are smaller than €42,000, the software first produces a two-tailed p-value of $p = 13.1\%$. For $\bar{x} = 41{,}941.65 < \mu_0 = 42{,}000$, we err in 6.55% of the cases when assuming that sales are smaller than €42,000. The one-tailed test (H_1: $\mu < \mu_0 = 40{,}000$) would not be statistically significant when $\alpha = 5\%$. We can proceed in similar fashion with H_1: $\mu > \mu_0$ and $\bar{x} > \mu_0$.

Stata offers two options for performing one-sample t-tests. If the parameters μ_0, $\bar{x}$, and α are already calculated, select *Statistics* → *Summaries, tables, and tests* → *Classical tests of hypotheses* → *One-Sample mean comparison calculator* to calculate the p-values of the one-tailed and both two-tailed tests (left window in Fig. 9.8). If the parameters are calculated from a dataset before carrying out the test, select *Statistics* → *Summaries, tables, and tests* → *Classical tests of hypotheses* → *One-sample mean comparison test* (right window in Fig. 9.8). Both approaches produce identical results (see Fig. 9.8).

In the R Commander, select *Statistics* → *Means* → *Single-sample t-test*… to open the window shown in Fig. 9.9 on the left. Select the variable you want to

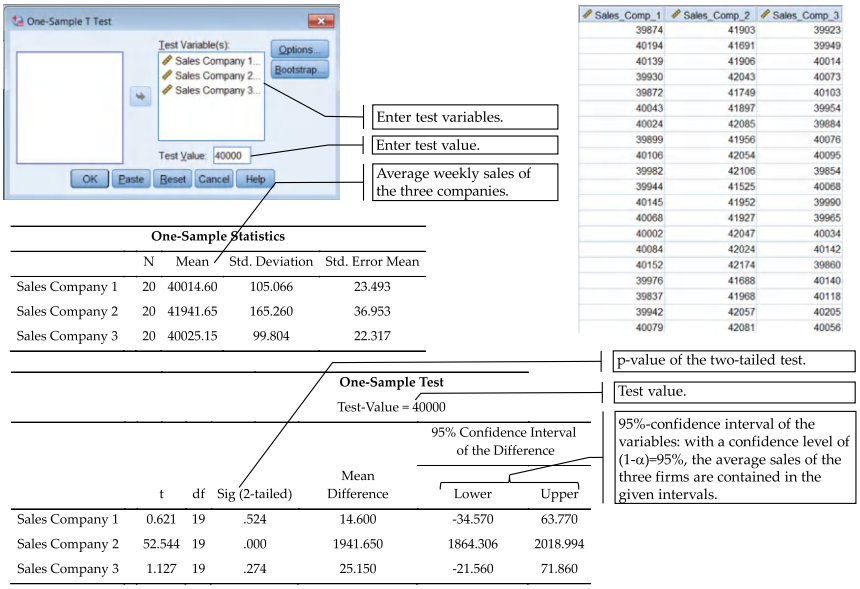


Fig. 9.7 The one-sample *t*-test with SPSS

test from the drop-down menu and enter the hypothesized mean value (this is the population mean you are testing against). Specify any other options as needed (e.g. confidence level). See the results at the bottom of Fig. 9.9.

In Excel, use the add-ins manager to make sure that the *Analysis ToolPak* and the *Analysis ToolPak VBA* are activated.[1] Then under the Data tab click the *Data Analysis* command button and select the *t-test: Paired two sample for means* entry from the list.

Though we want to perform a one-sample *t*-test (and not the two-sample test!), Excel does not offer this function specifically, so we have to use a little trick. In the field *Variable 1 Range*, indicate the data range of the test variables—e.g. the weekly sales of company one. For each of the test values, enter the test value μ_0 in one of the other columns; it will now appear in the field *Variable 2 Range*. In our example, all values for $\mu_0 = 40{,}000$ go in the column D (see Fig. 9.10).

The mean values of the three companies are 40,114.60, 41,941.65, and 40,025.15. In 52.4% of the cases, we err when assuming that the sales of company one differ from €40,000.00. The *p*-value of company three is 27.4%. Both values exceed the limit $\alpha = 0.05$, so that H_0 cannot be rejected. Neither the sales of company one nor those of company three are statistically different from €40,000.

[1] In Excel this can be reached by clicking *File* → *Options* → *Add-ins* → *Go*.

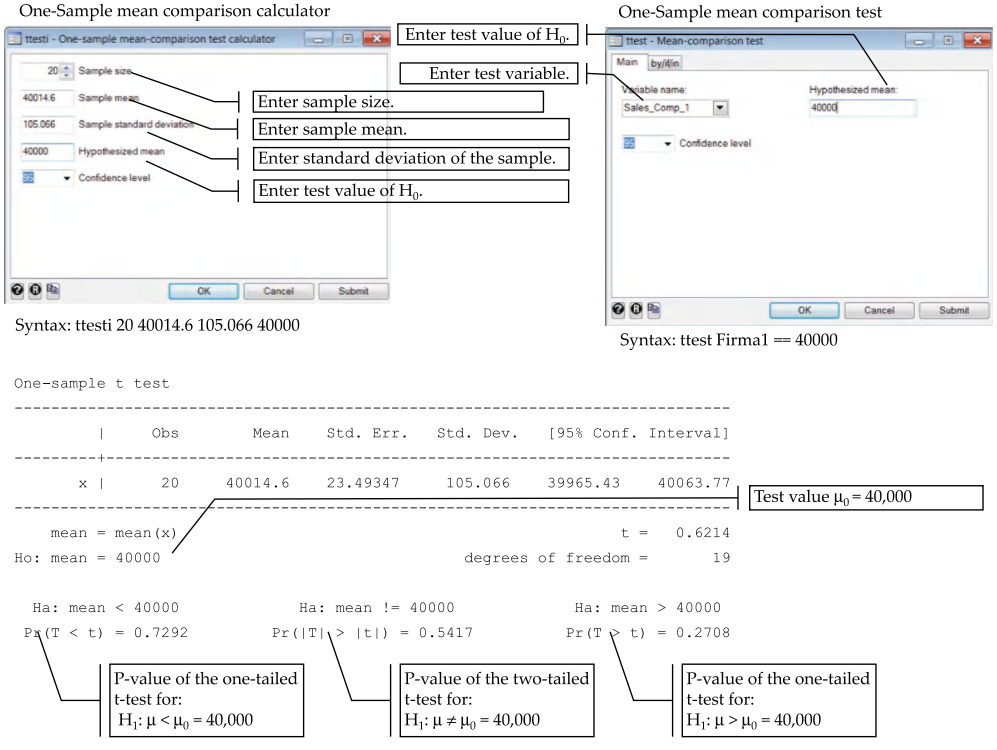


Fig. 9.8 The one-sample *t*-test with Stata




```
Result:
One Sample t-test data: sales_company_1
t = 0.62145, df = 19, p-value = 0.5417
alternative hypothesis: true mean is not equal to 40000
95 percent confidence interval: 39965.43 40063.77
sample estimates: mean of x 40014.6
```

Fig. 9.9 The one-sample *t*-test with R Commander

The sales of company two, however, are statistically different from €40,000. The two-tailed significance is 0.000, so that the *p*-value is smaller than 0.05.

9.3 Tests for Two Dependent Samples

9.3.1 The *t*-Test for Dependent Samples

The previous section was about whether the expected value of a population μ differs significantly from a hypothetical value μ_0 with H_0: $\mu = \mu_0$ and H_1: $\mu \neq \mu_0$. In practice, though, situations often arise in which researchers must compare mean values between two or more characteristics of an observation. A frequent case involving dependent samples is the before-after analysis with repeated measurements. Here are two examples:

- Customers are asked about their preference for a certain product before and after a promotional campaign. Researchers want to know whether average customer preference for the product before (μ_1) the promotional campaign changes after the promotional campaign (μ_2), and whether that change is statistically significant.
- Workers are asked to perform a certain task with two different machines. Does the average time it takes to perform the task with each machine differ significantly ($\mu_1 \neq \mu_2$)?

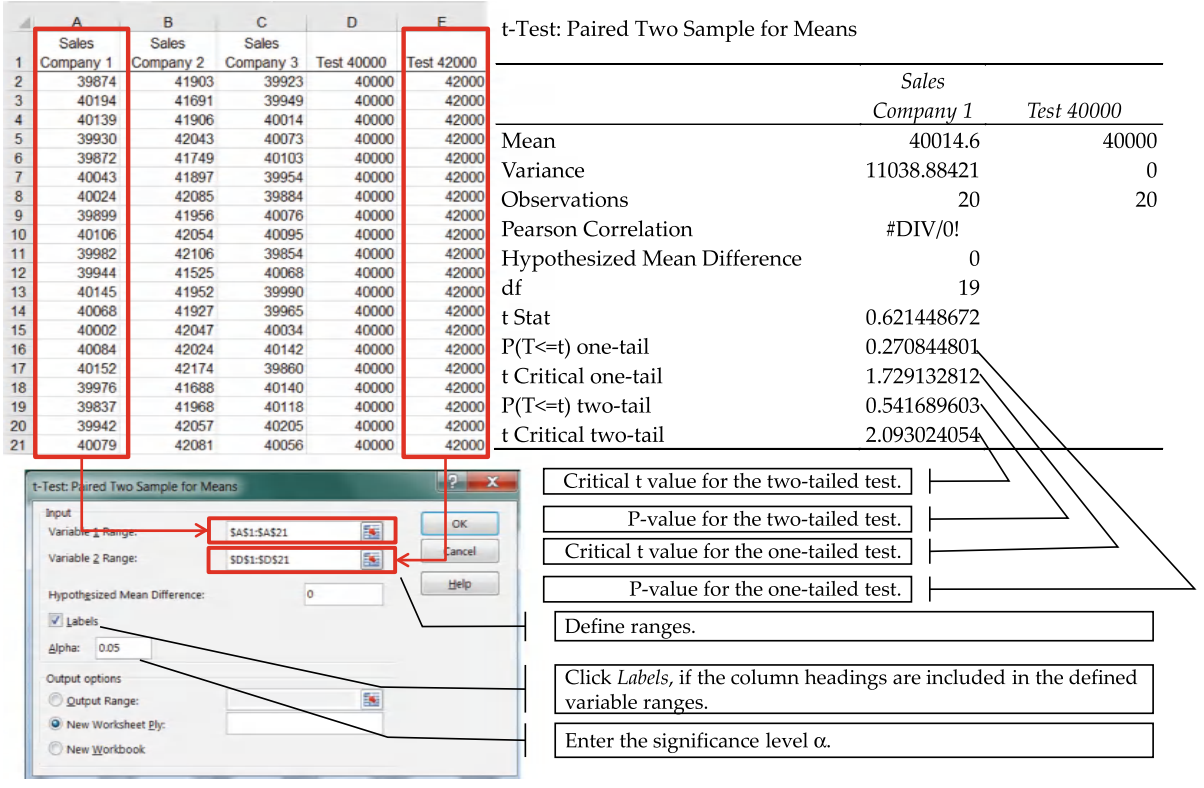


Fig. 9.10 The one-sample *t*-test with Excel

Sometimes, however, two groups are compared whose subjects or objects are paired. For instance, one can ask whether the average height of romantic partners differs. Each couple observation contains two uniquely matched data points arranged side by side in the dataset. In each of these cases, the goal is to verify the hypothesis that the mean value of an observed sequence of values of a given variable differs significantly from the mean value of a sequence of values from another variable. Let us look more closely at another example. Say market researchers want to know whether the average price of two coffee brands differs significantly across 32 test markets. They collect data in each of the markets and then consider their findings, which are shown in Fig. 9.11.

The researchers then perform the following steps to determine whether the average prices of the brands differ:

Step 1: Formulate the Hypotheses
The researchers' question can be expressed as a two-tailed test to identify the difference of the average prices for the coffee brands. The hypotheses are:

H_0: $\mu_1 = \mu_2 \rightarrow H_0$: $\mu_1 - \mu_2 = 0$
H_1: $\mu_1 \neq \mu_2 \rightarrow H_1$: $\mu_1 - \mu_2 \neq 0$

Again, the H_1 hypothesis describes the situation that researchers want to prove. If the researchers wanted to find out if the first coffee brand was on average more expensive than the second coffee brand, they would have to use a one-tailed test. The hypotheses would then look like this:

H_0: $\mu_1 \leq \mu_2$: The average price of coffee brand #1 is equal to or less than the average price of coffee brand #2 and
H_1: $\mu_1 > \mu_2$: The average price of coffee brand #1 is greater than the average price of coffee brand #2.

Step 2: Determine the Significance Level α

Next, we must define α, i.e. the maximum permissible probability that we reject an H_0 that is true. The significance level α is usually set at 1, 5, or 10%, though 5% is the most common value, which is what we will use here ($\alpha = 0.05$).

Step 3: Draw a Sample
The researchers draw a random sample of $n = 32$. As with a one-sample t-test, the paired t-test requires that the average price differences follow a normal distribution. This assumption need not be satisfied if a sample size of $n \geq 30$ suffices. If neither of the conditions is satisfied, the researchers would need to use a Wilcoxon signed-rank test (see Sect. 9.3.2). In our example, the sample with $n = 32$ observations is large enough to use the t-test.

Store	(2) Price Brand #1	(3) Price Brand #2	(4) Price difference (d_i) (Price #1 - Price #2)
1	3.22	3.01	0.21
2	3.09	3.16	-0.07
3	3.09	2.99	0.10
4	3.40	3.39	0.01
5	3.09	3.23	-0.14
6	3.19	3.05	0.14
7	2.99	2.88	0.11
8	2.65	2.57	0.08
9	3.09	2.99	0.10
10	3.09	3.21	-0.12
11	3.19	3.05	0.14
12	3.09	3.06	0.03
13	2.89	2.81	0.08
14	2.99	2.96	0.03
15	2.99	2.88	0.11
16	2.97	2.94	0.03
17	2.97	2.94	0.03
18	3.09	3.20	-0.11
19	2.59	2.73	-0.14
20	2.99	2.87	0.12
21	2.89	2.89	0.00
22	3.09	3.13	-0.04
23	2.59	2.59	0.00
24	2.89	2.93	-0.04
25	2.79	2.80	-0.01
26	2.99	2.96	0.03
27	2.99	3.07	-0.08
28	2.79	2.80	-0.01
29	2.91	2.90	0.01
30	3.04	3.02	0.02
31	2.69	2.57	0.12
32	2.72	2.50	0.22
Mean	2.9700	2.9400	0.0300
Standard deviation	0.1872	0.2040	0.0940

Fig. 9.11 Prices of two coffee brands in 32 test markets

Step 4: Check the Test Requirements

• Interval or ratio scales of measurement (approximately interval) for both variables	✓
• Random sampling from a defined population	✓
• Characteristics and difference scores are normally distributed in the population or sample size is big enough ($n \geq 30$)	✓

Step 5: Determine the Critical Test Value
The difference of the mean values follows a t-distribution with $n - 1$ degrees of freedom. When calculating the two-tailed test by hand, the researchers must first determine the theoretical t-value with a given probability error of α. With $(n - 1) = 31$ degrees of freedom, a two-tailed test—$\alpha/2$ on each side of the distribution—has a conservative theoretical t-value of $t^{31}_{0.025} \approx 2042$ (see Appendix C). A more precise table would have yielded a value of $t^{31}_{0.025} = 2.0395$.

Step 6: Determine the Empirical Test Value
The following formula calculates the empirical t-value of the sample:

$$t^{n-1} = \frac{\left(\overline{d} - \mu_{\mathrm{d}}\right)}{S_{\mathrm{d}}/\sqrt{n}} \tag{9.20}$$

The value for μ_{d} corresponds to that of the expected price difference. If the researchers wanted to check whether the mean values have the same size, then the expected difference would be zero ($\mu_{\mathrm{d}} = 0$). The calculation of the empirical t-value can then be simplified as follows:

$$t^{n-1} = \frac{\overline{d}}{S_{\mathrm{d}}/\sqrt{n}} \tag{9.21}$$

The sample from Fig. 9.11 yields an average price difference of 0.03 monetary units (MU):

$$\overline{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{0.96}{30} = 0.03\,\mathrm{MU}. \tag{9.22}$$

The same result occurs when subtracting the mean value of the first variable (price brand #1) from the mean value of the second variable (price brand #2):

$$\overline{d} = \bar{x}_1 - \bar{x}_2 = 0.03\,\mathrm{MU}. \tag{9.23}$$

The standard deviation of the price difference is 0.09 MU:

$$S_{\mathrm{d}} = \sqrt{\frac{\sum_{i=1}^{n}\left(d_i - \overline{d}\right)^2}{n-1}} = \sqrt{\frac{0.2742}{31}} = 0.0940\,\mathrm{MU}. \tag{9.24}$$

These values produce the following empirical t-value:

$$t^{31} = \frac{\left(\overline{d} - \mu_{\mathrm{d}}\right)}{S_{\mathrm{d}}/\sqrt{n}} = \frac{(0.03 - 0)}{0.0940/\sqrt{32}} = 1.804 \tag{9.25}$$

Step 7: Make a Test Decision
If the critical t-value exceeds the absolute empirical t-value from the dataset, the researchers cannot reject H_0. If it does not, they must reject H_0. In this study, H_0 cannot be rejected because the absolute empirical t-value is |1.804|, which is smaller than the theoretical minimum value $t^{31}_{0.025} = 2.0395$. Given a probability error of $\alpha \leq 0.05$, the price difference would not be statistically significant.

Calculating the paired t-test with statistical software indicates the p-value of the empirical data. The p-value reveals the probability of the empirical t-value when assuming the null hypothesis H_0, i.e. the probability error when assuming a price difference between the brands. If the p-value is below the fixed significance level α, the researchers must reject H_0. Below we will explain how to calculate paired t-tests using different statistical software packages.

9.3.1.1 The Paired *t*-Test with SPSS

Using the sample file *Chapter9_coffee.sav* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9) we will now calculate the paired t-test in SPSS. Select *Analyze* → *Compare Means* → *Paired-Samples T Test*… to open the window shown in Fig. 9.12. Then use drag and drop to move the variables being compared—price_1 and price_2—into the field for Variable1 and Variable2. Clicking *OK* creates the tables in Fig. 9.12. The first shows the descriptive parameters of the variables along with their mean values for variable 1 (price for brand #1) and variable 2 (price for brand #2), $\bar{x}_1 = 2.97$ and $\bar{x}_2 = 2.94$. The second table indicates that the size of the price difference is 0.03 euro and that the empirical t-value is 1.804. Accordingly, we can say with 95% confidence that the real price difference in the population lies between − 0.00391 and 0.06391 euro. The crucial result appears in the column *Sig. (2-tailed)*, where the p-value for a two-tailed test is 0.081. This means that we err in 8.1% of cases when assuming an average price difference. This value is larger than the significance level of $\alpha = 5\%$, however.

Were the hypotheses different, namely if

H_0: $\mu_1 \leq \mu_2$: product price #1 is equal to or lesser than product price #2 and
H_1: $\mu_1 > \mu_2$: product price #1 is greater than product price #2,

then the p-value would have to be divided by two. In this case, we would err in only 4% of the cases when assuming that product #1 is on average more expensive than product #2. In this case, the result would be statistically significant.[2]

9.3.1.2 The Paired *t*-Test with Stata

Using the sample file *Chapter9_coffee.dta* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9) we will now calculate the paired t-test in Stata. Select *Statistics* → *Summaries, tables, and tests* → *Classical tests of*

[2] For a very good explanation of how to perform this test in SPSS, see https://www.youtube.com/watch?v=MJGk2sg4EZU on the *how2stats* YouTube channel.

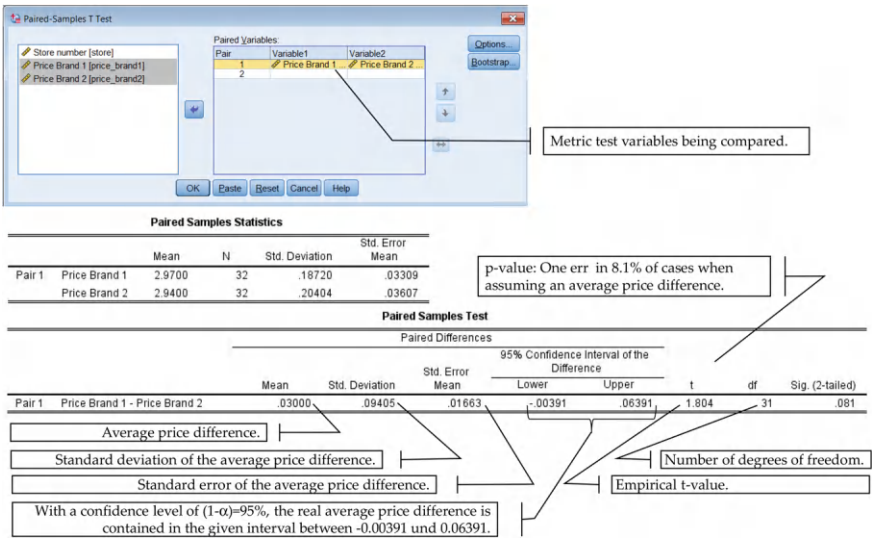


Fig. 9.12 The paired *t*-test with SPSS

hypotheses → *Mean-comparison test, paired data* to open the window shown in Fig. 9.13. Then move the variables being compared—price_brand1 and price_brand2—into the field for *First variable* and *Second variable.*

Clicking *OK* creates the table in Fig. 9.13. It shows the descriptive parameters of the variables along with their mean values for the first variable (price for brand #1) and the second variable (price for brand #2), $\bar{x}_1 = 2.97$ and $\bar{x}_2 = 2.94$. It also indicates that the size of the price difference is 0.03 euro and that the empirical *t*-value is 1.8044. Accordingly, we can say with 95% confidence that the real price difference in the population lies between − 0.0039082 and 0.0639082 euro. The crucial result appears at the bottom of Fig. 9.13, where the *p*-value for a two-tailed test is $\Pr(|T| > |t|) = 0.0809$. This means that we err in 8.1% of cases when assuming an average price difference. This value is larger than the significance level of $\alpha = 5\%$, however.

Were the hypotheses different, namely if

H_0: $\mu_1 \leq \mu_2$: product price #1 is equal to or lesser than product price #2 and
H_1: $\mu_1 > \mu_2$: product price #1 is greater than product price #2,

then the *p*-value would have been $\Pr(T > t) = 0.0404$. In this case, we would err in only 4% of the cases when assuming that product #1 is on average more expensive than product #2. In this case, the result would be statistically significant.

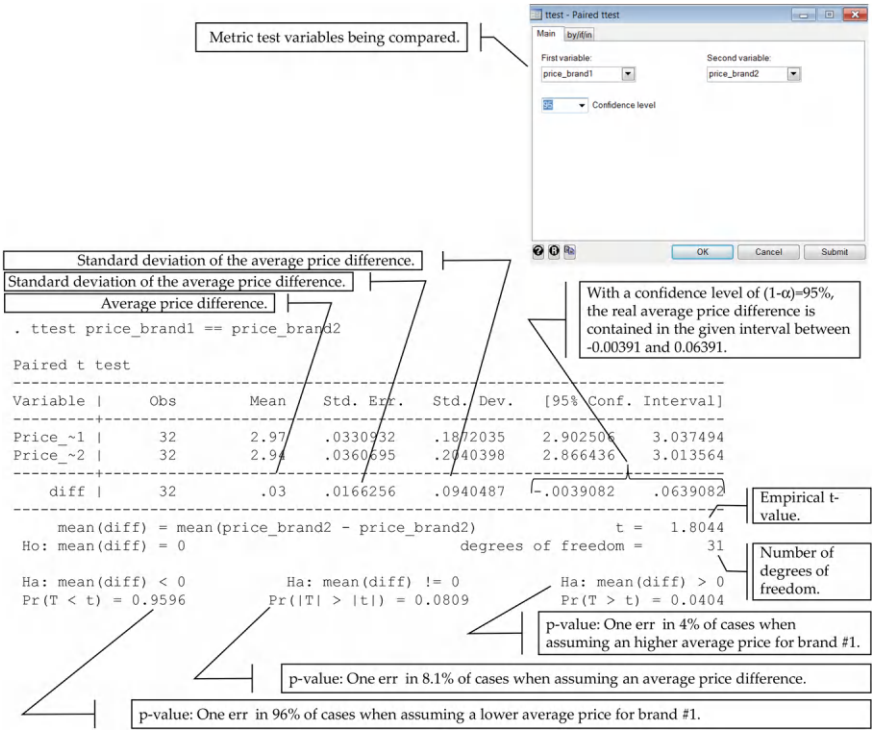


Fig. 9.13 The paired *t*-test with Stata

If the above hypotheses were to be reversed, namely if

H_0: $\mu_1 \geq \mu_2$: product price #1 is equal to or higher than product price #2 and
H_1: $\mu_1 < \mu_2$: product price #1 is smaller than product price #2,

then the *p*-value would have been $\Pr(T < t) = 0.9596$. In this case, we would err in only 96% of the cases when assuming that product #1 is on average cheaper than product #2. In this case, the result would not be statistically significant.[3]

9.3.1.3 The Paired *t*-Test with R Commander

Using the sample file *Chapter9_coffee.RData* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9) we will calculate the paired *t*-test with the R Commander. Select *Statistics* → *Means* → *Paired t-test*… to open the

[3] For a very good explanation of how to perform this test in Stata, see https://www.youtube.com/watch?v=ajzMeANAMzI on the *Stata Learner* YouTube channel.

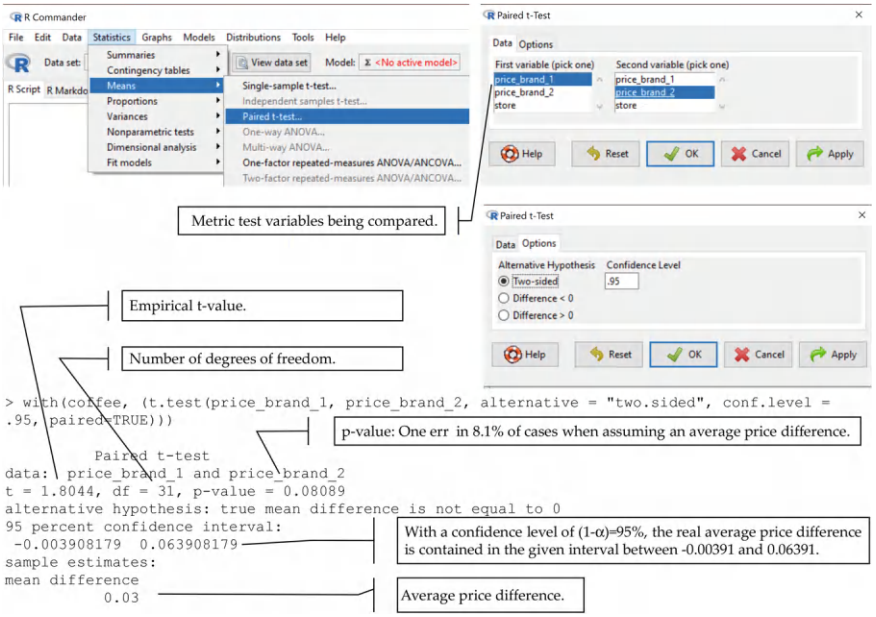


Fig. 9.14 The paired *t*-test with R Commander

window shown in Fig. 9.14. Choose the variables being compared—price_brand1 and price_brand2—in the fields for *First variable* and *Second variable*. *Options* allow you to select between one-sided or two-sided tests. Clicking *OK* creates the table in Fig. 9.14.

It is shown that the price difference amounts to 0.03 euro and that the empirical *t*-value is 1.8044. Accordingly, we can say with 95% confidence that the real price difference in the population lies between (− 0.003908179) and 0.063908179 euro (confidence interval). The *p*-value for a two-tailed test is $p = 0.08089$. This means that we are wrong in 8.1% of cases if we assume an average price difference. This value is above the significance level $\alpha = 5\%$, meaning that the mean difference is not significant at this level.

9.3.1.4 The Paired *t*-Test with Excel

In Excel (*Chapter9_coffee.xls*; see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9), go to *Data* and select *Data Analysis* and *t-Test: Paired Two Sample for Means*. This opens the window shown in Fig. 9.15. Here you can enter the ranges of the variables in *Variable 1 Range* and *Variable 2 Range*. If the column headings are included in the defined variable ranges, *Labels* needs to be clicked. To test the equality of the mean values of the

variables, enter the value zero under *Hypothesized Mean Difference*. Clicking *OK* produces the results shown in Fig. 9.15.[4]

9.3.2 The Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test analyses two paired and ordinally scaled random variables for equality of central tendency. Since this test requires no assumptions about the distributions of the two variables, it is considered a nonparametric alternative to the paired *t*-test. It is primarily for ranked data, but it can also be used if the requirements of the *t*-test are not satisfied. Even in cases in which the requirements of a parametric test are satisfied, the Wilcoxon signed-rank test has some very efficient properties.

In constructing his test, Wilcoxon (1945, 1947) drew on a basic idea conceived by Spearman (1904). Spearman followed the Pearson correlation when calculating the ranked correlation coefficients that would later be named after him but used ranked data instead of the initial metric data. Likewise, Wilcoxon transformed metric data into ranked data for the paired *t*-test calculation. We will now use his approach to re-examine our example of coffee brands (see Fig. 9.16) and see whether their prices differ in central tendency.

Step 1: Formulate the Hypotheses
The researchers' question can be expressed as a two-tailed test to identify the difference of the mean values for the two coffee brands. The hypotheses are:

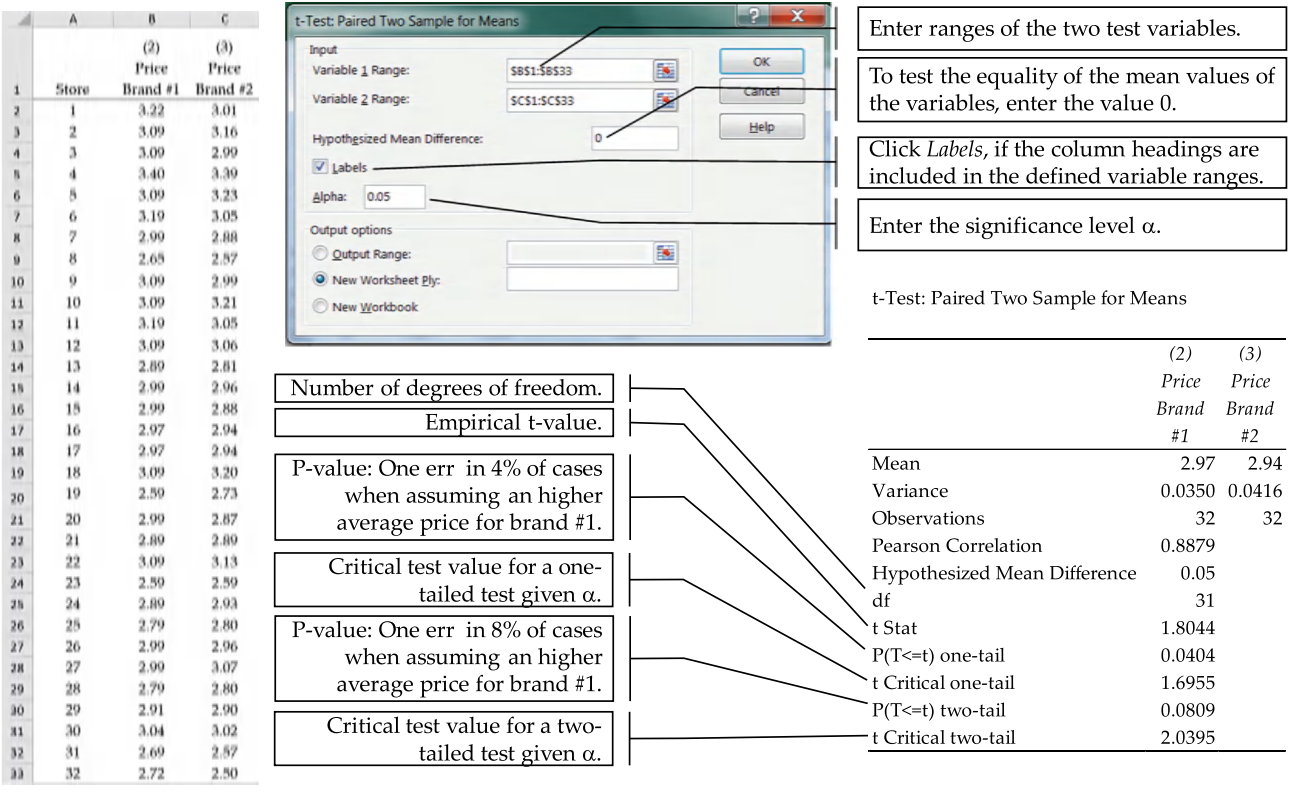
H_0: $M_1 - M_2 = 0$
H_1: $M_1 - M_2 \neq 0$.

If the researchers wanted to find out if the first coffee brand was more expensive than the second coffee brand, they would have to use a one-tailed test. The hypotheses would then look like this:

H_0: $M_1 \leq M_2$: The average price of coffee brand #1 is equal to or less than the average price of coffee brand #2 and
H_1: $M_1 > M_2$: The average price of coffee brand #1 is greater than the average price of coffee brand #2.

Again, the H_1 hypothesis describes the situation that researchers want to prove.

[4] For a very good explanation of how to perform this test in Excel, see https://www.youtube.com/watch?v=wy8GVt7Ityk on the YouTube channel of https://alphabench.com.



	A	B	C
1	Store	(2) Price Brand #1	(3) Price Brand #2
2	1	3.22	3.01
3	2	3.09	3.16
4	3	3.09	2.99
5	4	3.40	3.39
6	5	3.09	3.23
7	6	3.19	3.05
8	7	2.99	2.88
9	8	2.65	2.57
10	9	3.09	2.99
11	10	3.09	3.21
12	11	3.19	3.05
13	12	3.09	3.06
14	13	2.89	2.81
15	14	2.99	2.96
16	15	2.99	2.88
17	16	2.97	2.94
18	17	2.97	2.94
19	18	3.09	3.20
20	19	2.59	2.73
21	20	2.99	2.87
22	21	2.89	2.89
23	22	3.09	3.13
24	23	2.59	2.59
25	24	2.89	2.93
26	25	2.79	2.80
27	26	2.99	2.96
28	27	2.99	3.07
29	28	2.79	2.80
30	29	2.91	2.90
31	30	3.04	3.02
32	31	2.69	2.57
33	32	2.72	2.50

t-Test: Paired Two Sample for Means

	(2) Price Brand #1	*(3) Price Brand #2*
Mean	2.97	2.94
Variance	0.0350	0.0416
Observations	32	32
Pearson Correlation	0.8879	
Hypothesized Mean Difference	0.05	
df	31	
t Stat	1.8044	
P(T<=t) one-tail	0.0404	
t Critical one-tail	1.6955	
P(T<=t) two-tail	0.0809	
t Critical two-tail	2.0395	

Fig. 9.15 The paired *t*-test with Excel

	A	B	C	D	F	G	H	I
1	Store	(2) Price Brand #1	(3) Price Brand #2	(4) Price difference (d_i) (Price #1 - Price #2)	(5) Ranks of the differences	(6) "Positive Ranks"	(7) "Negative Ranks"	(8) No differences
2	1	3.22	3.01	0.21	29	29		
3	2	3.09	3.16	-0.07	13		13	
4	3	3.09	2.99	0.10	17.5	17.5		
5	4	3.40	3.39	0.01	2.5	2.5		
6	5	3.09	3.23	-0.14	26.5		26.5	
7	6	3.19	3.05	0.14	26.5	26.5		
8	7	2.99	2.88	0.11	20	20		
9	8	2.65	2.57	0.08	15	15		
10	9	3.09	2.99	0.10	17.5	17.5		
11	10	3.09	3.21	-0.12	23		23	
12	11	3.19	3.05	0.14	26.5	26.5		
13	12	3.09	3.06	0.03	8	8		
14	13	2.89	2.81	0.08	15	15		
15	14	2.99	2.96	0.03	8	8		
16	15	2.99	2.88	0.11	20	20		
17	16	2.97	2.94	0.03	8	8		
18	17	2.97	2.94	0.03	8	8		
19	18	3.09	3.20	-0.11	20		20	
20	19	2.59	2.73	-0.14	26.5		26.5	
21	20	2.99	2.87	0.12	23	23		
22	21	2.89	2.89	0.00				X
23	22	3.09	3.13	-0.04	11.5		11.5	
24	23	2.59	2.59	0.00				X
25	24	2.89	2.93	-0.04	11.5		11.5	
26	25	2.79	2.80	-0.01	2.5		2.5	
27	26	2.99	2.96	0.03	8	8		
28	27	2.99	3.07	-0.08	15		15	
29	28	2.79	2.80	-0.01	2.5		2.5	
30	29	2.91	2.90	0.01	2.5	2.5		
31	30	3.04	3.02	0.02	5	5		
32	31	2.69	2.57	0.12	23	23		
33	32	2.72	2.50	0.22	30	30		
34					465	313	152	

K	L	M
Two-tailed test		
Positive Ranks:	313	=G34
Negative Ranks:	152	=H34
W	152	=MIN(L2;L3)
a	0.05	
n*	30	=COUNTIF(F2:F33;">0.0")
Z	-1.65575	=(G39-((G41*(G41+1)/4)))/(((G41*(G41+1)*(2*G41+1)/24)^0.5)
p-value	0.098	=NORM.S.DIST(L7;1)
Ho not rejected		=IF(L16>L19; "H_o rejected"; "Ho not rejected")
One-tailed test		
Positive Ranks:	313	=G34
Negative Ranks:	152	=H34
W	152	=MIN(L2;L3)
a	0.05	
n*	30	=COUNTIF(F2:F33;">0.0")
Z	-1.655751	=(G39-((G41*(G41+1)/4)))/(((G41*(G41+1)*(2*G41+1)/24)^0.5)
p-value	0.049	=NORM.S.DIST(L7;1)
Ho rejected		=IF(L16>L19; "H_o rejected"; "Ho not rejected")

Fig. 9.16 Data for the Wilcoxon signed-rank test

Step 2: Determine the Significance Level α

Next, we must define α, i.e. the maximum permissible probability that we reject an H_0 that is true. The significance level α is usually set at 1, 5, or 10%, though 5% is the most common value, which is what we will use here ($\alpha = 0.05$).

Step 3: Draw a Sample

The random sample has a size of $n = 32$. Both of the variable sequences being compared are ordinal-scaled.

Step 4: Check the Test Requirements

• Interval or ordinal scales of measurement for both variables	✓
• Random sampling from a defined population	✓
• The distribution of the two related test variables needs to be symmetrical	✓

Step 5: Determine the Critical Test Value

First, we calculate the price difference for each test market (price_1 − price_2, see column (4) in Fig. 9.16). Observations without price differences are omitted from further calculations. The new sample size is now $n^* = 30$.[5]

Now we order and rank them by the size of their differences regardless of their signs (column (5) in Fig. 9.16). The larger the absolute price difference is, the higher its rank. The average rank score is calculated for identical price differences (tied ranks).

Next, we add up all twenty rank scores associated with a positive price difference in the original dataset (column (6) in Fig. 9.16).

$$W^+ = \sum_{i=1}^{n} \text{positive rank scores} = 313 \tag{9.26}$$

We do the same for the ten negative rank scores (column (7) in Fig. 9.16):

$$W^- = \sum_{i=1}^{n} \text{negative rank scores} = 152 \tag{9.27}$$

It is striking that the sum of the positive ranks is much larger than the sum of the negative ranks. It appears that product #1 is frequently more expensive than product #2. This is because, were the prices the same, one half of the total rank sum would consist of positive differences and the other half, of negative differences. In both cases, we would expect a rank sum of 232.5:

$$E(W^+) = E(W^-) = \frac{n^* \cdot (n^* + 1)}{4} = \frac{30 \cdot (30 + 1)}{4} = 232.5 \tag{9.28}$$

[5] Most statistical software packages perform this step automatically.

This is not the case, however. The more W^+ and W^- diverge, the greater the likelihood that the central tendencies of the two variables (prices) differ. In other words, the more W^+ and W^- diverge, the greater the likelihood that H_1—that the ordinal variables (prices) show a difference with regard to central tendency—is correct. But is this difference statistically significant?

To answer this question, we will henceforth distinguish between small samples and large samples. For small samples of up to 25 observations ($n^* \leq 25$), the critical values are available in the table in Appendix D. Our example yields a critical test value of $W_c = 137$ with a two-tailed hypothesis and a significance level of $\alpha = 0.05$.

Step 6: Determine the Empirical Test Value

We will now compare this value with the empirical test value W. First, we need to determine the minimum from the rank sums W^+ and W^-:

$$W = \min\left(W^+; W^-\right) = \min(313; 152) = 152 \tag{9.29}$$

Step 7: Make a Test Decision

If the minimum W is larger than the critical value, we cannot reject H_0; otherwise we can. The test logic here is the exact opposite of the other test procedures: H_0 is rejected if the empirical test value W is smaller than the critical value in the table, because the minimum from W^+ and W^- becomes smaller as the price difference grows larger.

In our example, W has a value of 152, which is larger than the critical value in the table, where $W_c = 137$. Accordingly, we cannot reject H_0 and we cannot assume that the price difference with an α of 0.05 is statistically significant.

$$W = \min\left(W^+; W^-\right) = \min(313; 152) = 152 > 137 = W_c\left(0.05; n^* = 30\right) \tag{9.30}$$

This procedure describes the exact Wilcoxon signed-rank test. The procedure is based on the Wilcoxon table values for W_c in Appendix D for samples of up to 50 observations. With samples larger than 25 ($n^* > 25$) and with no tied ranks, the asymptotic Wilcoxon signed-rank test can be applied. The following Z value has an asymptotic standard normal distribution. At first glance, this formula might seem complicated, but it contains only the parameters W and n^*:

$$Z = \frac{W - \left(\frac{n^* \cdot (n^*+1)}{4}\right)}{\sqrt{\frac{n^* \cdot (n^*+1) \cdot (2 \cdot n^*+1)}{24}}} \sim N(0; 1) \tag{9.31}$$

If tied ranks exist, we must use the following somewhat more complicated formula in which k is the number of tied ranks and t_i corresponds to the length of the ith tied rank:

$$Z = \frac{W - \left(\frac{n^* \cdot (n^*+1)}{4}\right)}{\sqrt{\frac{n^* \cdot (n^*+1) \cdot (2 \cdot n^*+1) - \sum_i^k \frac{t_i^3 - t_i}{2}}{24}}} \sim N(0;1) \tag{9.32}$$

In our example, we have one tied rank of length two (rank score: 17.5), five tied ranks of length three (rank scores: 2.5; 5; 8; 15; 20), and two tied ranks of length four (rank score: 23; 26). This yields an empirical Z value of:

$$Z = \frac{152 - \left(\frac{30 \cdot (30+1)}{4}\right)}{\sqrt{\frac{30 \cdot (30+1) \cdot (2 \cdot 30+1) - \left(\left(\frac{2^3-2}{2}\right) + 5 \cdot \left(\frac{3^3-3}{2}\right) + 2 \cdot \left(\frac{4^3-4}{2}\right)\right)}{24}}} = (-1.658) \tag{9.33}$$

This formula looks intimidating, but in practice researchers typically use computer software to calculate the Z value and to identify the associated p-value. We will learn more about this in the following sections.

Those who have to calculate the test by hand must check whether the empirical Z value lies in the H_0 rejection area of the normal distribution or not (see Fig. 9.17). A two-tailed hypothesis and an α of 0.05 from the normal distribution table for the "non-rejection-area" of H_0 yields results within the limits of

$$Z_{1-\frac{\alpha}{2};\frac{\alpha}{2}} = (\pm 1.96). \tag{9.34}$$

At 1.658, the empirical Z value lies in this area. Hence, we cannot reject H_0.

9.3.2.1 The Wilcoxon Signed-Rank Test with SPSS

Let us perform the Wilcoxon signed-rank test in SPSS using the sample file *Chapter9_coffee.sav* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9). Select *Analyze* → *Nonparametric Tests* → *Legacy Dialogs* → *2 Related Samples …* to open the window shown in Fig. 9.18. Here you can drag and drop the variables to be compared (*price for brand #1* and *price for brand #2*) into the fields for *Variable1* and *Variable2*. Click the button *Exact…* to create an exact or asymptotic Wilcoxon signed-rank test. An exact test should be used only with small samples ($n^* \leq 25$). Clicking *Ok* produces the tables in Fig. 9.18. The table on the left provides the rank sums W^+ and W^- and the absolute frequencies of positive and negative ranks and the number of identically priced observations. The table on the right gives the p-values of one-tailed and two-tailed hypothesis tests. From this we see that one errs in 9.9% of the cases when assuming a price difference. This value lies above the significance level of $\alpha = 5\%$. The difference is thus not significant.

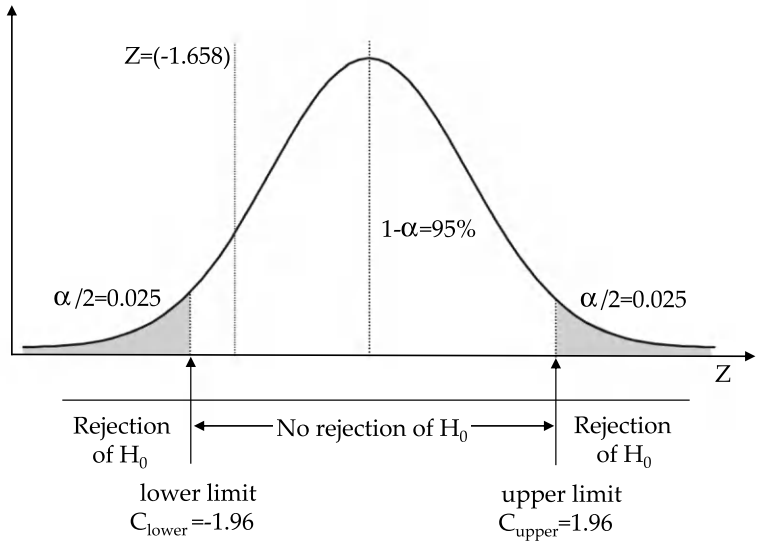


Fig. 9.17 Rejection area of the Wilcoxon signed-rank test

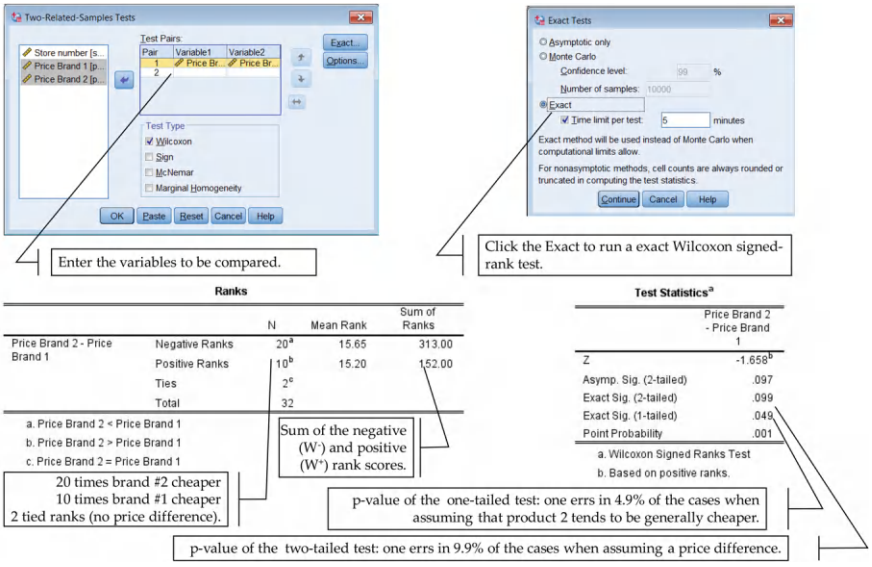


Fig. 9.18 The Wilcoxon signed-rank test with SPSS

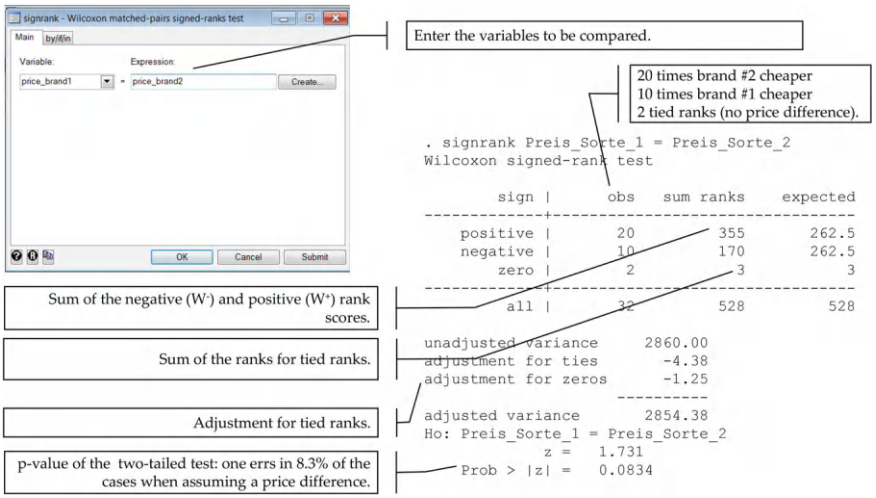


Fig. 9.19 The Wilcoxon signed-rank test with Stata

With a one-tailed hypothesis test, we err in 4.9% of the cases when assuming that product #2 tends to be generally cheaper. This result is statistically significant when $\alpha = 0.05$ is assumed.[6]

9.3.2.2 The Wilcoxon Signed-Rank Test with Stata

Using the sample file *Chapter9_coffee.dta* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9) we want to calculate the Wilcoxon signed-rank test in Stata. Selecting *Statistics* → *Summaries, tables, and tests* → *Nonparametric tests of hypotheses* → *Wilcoxon matched-pairs signed-rank test* opens the window shown in Fig. 9.19. Enter the variables to be compared, *price_brand1* and *price_brand2*, in the corresponding fields *Variable* and *Expression*. Clicking Ok produces the tables in Fig. 9.19. The table provides the positive and negative rank sums W^+ and W^- and the absolute frequencies of positive and negative ranks and the number of observations with the same prices. Below the *p*-value is given for the two-tailed hypothesis test. One errs in 8.3% of cases if one assumes a price difference. This value lies above the fixed significant level of 0.05. Therefore, the result is not statistically significant.[7]

[6] For a very good explanation of how to do this in SPSS, see https://www.youtube.com/watch?v=dkobjvhxTro on the YouTube channel of *Dr. Todd Grande*.

[7] For a very good explanation of how to do this in Stata, see https://www.youtube.com/watch?v=2oJxerMCwIE and https://www.youtube.com/watch?v=NIwtaZqNFs8 on the *Stata Learner* YouTube channel.

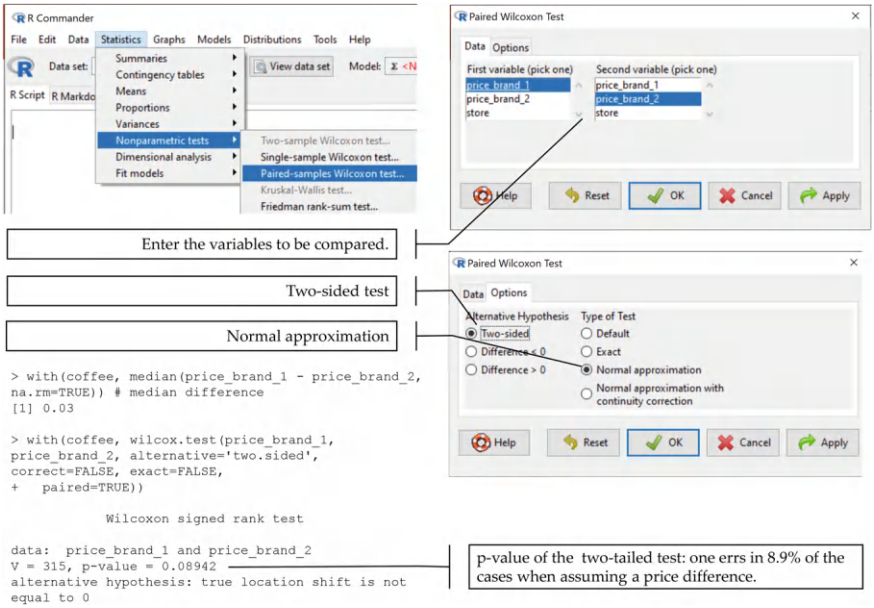


Fig. 9.20 The Wilcoxon signed-rank test with R Commander

9.3.2.3 The Wilcoxon Signed-Rank Test with R Commander

Using the sample file *Chapter9_coffee.RData* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9) we are going to calculate the Wilcoxon signed-rank test with R Commander. Selecting *Statistics* → *Nonparametric tests* → *Paired-samples Wilcoxon tests*… opens the window shown on the right in Fig. 9.20. Enter the variables to be compared, *price_brand1* and *price_brand2*, in the corresponding fields *First variable* and *Second variable*. *Options* allows you to select between one-sided and two-sided tests. Clicking Ok produces the tables in Fig. 9.20. The p-value for a two-tailed test is $p = 0.08942$. This means that we are wrong in 8.9% of cases if we assume a price difference. This value is above the significance level $\alpha = 5\%$, meaning that the mean difference is not significant at this level.

9.3.2.4 The Wilcoxon Signed-Rank Test with Excel

The Wilcoxon signed-rank test is not implemented as a function in Excel. There is a way to calculate the asymptotic p-value, as Fig. 9.16 shows using our sample file *Chapter9_coffee.xls* (see supplementary material at DOI:

https://doi.org/10.1007/978-3-031-78070-7_9). But this procedure will not correct for any tied ranks. For smaller samples ($n \leq 25$), use the Wilcoxon table.[8]

9.4 Tests for Two Independent Samples

9.4.1 The *t*-Test of Two Independent Samples

The *t*-test of two independent samples identifies the difference between the mean values in two different groups. The sample can be broken down into two subsamples using grouping variables. For both groups, we assume the expected values of their respective populations differ by Δ with regard to a metric variable.

Step 1: Formulate the Hypotheses

$$H_0: \mu_1 - \mu_2 = \Delta$$

The most frequently used value for the difference is $\Delta = 0$, i.e. that there is no difference between the mean values μ_1 and μ_2 of the two populations. The hypotheses of a two-tailed test are then:

$$H_0: \mu_1 = \mu_2 \text{ and } H_1: \mu_1 \neq \mu_2$$

If the question was whether the mean value of group #1 is significantly larger than the mean value of group #2, we would need a one-tailed test. The hypotheses would then be:

$$H_0: \mu_1 \leq \mu_2 \text{ and } H_1: \mu_1 > \mu_2$$

Step 2: Determine the Significance Level α

Next, we must define α, i.e. the maximum permissible probability that we reject an H_0 that is true. The significance level α is usually set at 1, 5, or 10%, though 5% is the most common value, which is what we will use here ($\alpha = 0.05$).

Step 3: Draw a Sample

Population #1 has a mean of μ_1 with a standard deviation of σ_1. Population #2 has a mean of μ_2 and a standard deviation of σ_2. Now we draw a sample with a size n_1 from population #1 and a sample with a size n_2 from population #2. For each sample, we determine the mean values ($\bar{x}_1$ and $\bar{x}_2$) and the standard deviations ($\hat{\sigma}_1 = S_1$ and $\hat{\sigma}_2 = S_2$).

[8] For a very good explanation of how to do this in Excel, see https://www.youtube.com/watch?v=xlgeta9FivI on the YouTube channel of *Matthias Kullowatz* and https://www.youtube.com/watch?v=mJtbhGETU88 on the YouTube channel of *Dr. Todd Grande*.

Step 4: Check the Test Requirements

• Interval or ratio scales of measurement (approximately interval) of the test variable	✓
• Random sampling from two defined populations	✓
• Samples are independent and there is no overlap between group members	✓
• Test variable is normally distributed in the population or sample size is big enough ($n \geq 30$)	✓

Step 5: Determine the Critical Test Value
The difference between the mean values of the samples follows a *t*-distribution with $(n_1 + n_2 - 2)$ degrees of freedom. The critical *t*-value for a two-tailed *t*-test is defined as follows:

$$t^{\text{critical}}_{1-\frac{\alpha}{2};n_1+n_2-2} \tag{9.35}$$

The *t*-value for a one-tailed *t*-test would be:

$$t^{\text{critical}}_{1-\alpha;n_1+n_2-2} \tag{9.36}$$

Step 6: Determine the Empirical Test Value
From the samples, we can calculate the values for $\bar{x}_1$, $\bar{x}_2$, n_1, n_2, $\hat{\sigma}$ and determine the empirical test value:

$$t^{\text{df}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{9.37}$$

If we are to calculate the value exactly, we must check in advance whether the variances of the two samples are of the same size or not. The first is called homoscedastic variance and the second, heteroscedastic variance. If the variances are homoscedastic, then $\sigma_1 = \sigma_2$, so that:

$$\hat{\sigma} = \hat{\sigma}_1 = \hat{\sigma}_2 = \frac{(n_1 - 1) \cdot \hat{\sigma} + (n_2 - 1) \cdot \hat{\sigma}}{n_1 + n_2 - 2} = \frac{(n_1 - 1) \cdot S + (n_2 - 1) \cdot S}{n_1 + n_2 - 2} \tag{9.38}$$

The empirical test value can be converted as follows:

$$t^{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \cdot \frac{(n_1-1)\cdot\hat{\sigma}+(n_2-1)\cdot\hat{\sigma}}{n_1+n_2-2}}} \tag{9.39}$$

Step 7: Make a Test Decision
If the absolute empirical test value from the dataset exceeds the critical t-value, H_0 has to be rejected:

$$t^{\text{critical}}_{1-\frac{\alpha}{2};n_1+n_2-2} < \left|t^{n_1+n_2-2}\right| = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\cdot\frac{(n_1-1)\cdot\hat{\sigma}+(n_2-1)\cdot\hat{\sigma}}{n_1+n_2-2}}} \rightarrow \text{reject}\, H_0 \tag{9.40}$$

If it does not, H_0 cannot be rejected. As we will see later in an example (see Sect. 9.4.1.1), the absolute empirical test value from the dataset exceeds the critical t-value. We must therefore reject H_0:

$$t^{\text{critical}}_{0.975;238} = 1.97 < \left|t^{238}\right| = |-6.61| \tag{9.41}$$

Next we can evaluate the likelihood of the empirical t-value at $(n_1 + n_2 - 2)$ degrees of freedom when assuming the validity of the null hypothesis. If the p-value lies below the significance level α, we must reject H_0. See the software calculations in the following sections.

Both the F-test and Levene's test can check whether it is fair to assume that the variances are really equal (see assumption in Eq. 9.38). When discussing the software applications below, these tests will be described in more detail. If the equality of variance cannot be assumed, we must use a different approach to calculate the t-value. Welch (1947) and Satterthwaite (1946) each proposed an approach for determining the critical t-value in the heteroscedastic case of $\sigma_1 \neq \sigma_2$. In many instances, the calculations of these test statistics are complex enough that we must rely on software to do the work.

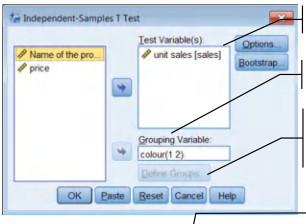
Nevertheless, one virtue of the t-test is that it produces robust results even when the test requirements in step 4 are not satisfied, provided that the samples are approximately the same size. This is especially the case for unimodal distributions. However, even significant differences in sample sizes have little effect on the precision of the t-test if the variances of the populations are roughly the same. But significant differences in sample sizes and in population variances will bias the results. In this case, therefore, the nonparametric U test should be used (see Sect. 9.4.2).

Let us use the following example to show how to calculate a t-test for two independent samples with Excel, SPSS, and Stata. A company wants to introduce a new type of chocolate praline, but marketing executives have been unable to agree on whether the packaging should be blue or yellow. To aid their decision, they commission a research study measuring the effect of colour on sales in 240 comparable test markets. They want to find out whether the sales attained by the two packaging colours differ significantly on a 5% significance level.

9.4.1.1 The *t*-Test for Two Independent Samples with SPSS

To run the *t*-test for two independent samples with SPSS based on the sample file *Chapter9_chocolatepraline_colour_name_price.sav* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9), select *Analyze* → *Compare Means* → *Independent-Samples T Test*. This opens the window in Fig. 9.21, which shows the test variable (sales) and the grouping variable (colour). Under *Define Groups*, select the group numbers for comparison—one for blue packaging and two for yellow packaging. Now click *OK* to produce the table in Fig. 9.21.

The resulting mean values—$\bar{x}_{\text{blue}} = 9982.4$ for the blue packaging and $\bar{x}_{\text{yellow}} = 9606.1$ for the yellow packaging—indicate a difference in sales based on colour. To know for certain, we first need to check whether the equality of variance assumption has been satisfied. Levene's test produces a *p*-value of $p = 0.000$ (< 5%). Therefore, we cannot assume the equality of variance. For the heteroscedastic case, SPSS provides a second line (*Equal variances not assumed*) that contains the results of the Welch's Test for Unequal Variances. The *p*-value of 0.000 in the *Sig. (2-tailed)* column indicates that the mean values of sales for the two packaging colours differ significantly. Indeed, we can be 95% confident that the sales difference lies between 264.146 and 488.454 packages.



Select the (metric) test variables *(sales)*.

Select the grouping variable *(colour)*.

Select the group numbers for comparison: 1 for blue packaging and 2 for yellow packaging.

	Colour	N	Mean	Std. Deviation	Std. Error
unit sales	blue	120	9982.4	385.15	35.16
	yellow	120	9606.1	490.29	44.76
	Total	240			

Number of observations for the colour *blue*.

Average *sales* for the colour *blue*.

Standard deviation of the variable *sales* for the colour *blue*.

Standard error of the variable *sales* for the colour *blue*.

Total number of observations.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
unit sales	Equal variances assumed	13.666	.000	6.612	238	.000	376.30	56.915	264.178	488.422
	Equal variances not assumed			6.612	225.364	.000	376.30	56.915	264.146	488.454

Levene's test produces a p-value of p=0.000. Therefore, we cannot assume the equality of variance. For the heteroscedastic case, SPSS provides in the second line (Equal variances not assumed) the results of the Welch's Test for Unequal Variances.

We can be 95% confident that the difference lies between 264.146 and 488.454.

The mean values of *sales* for the two packaging colours differ significantly (p≤0.05).

Fig. 9.21 The *t*-test for two independent samples with SPSS

Though significant deviations from the normal distribution exist[9] for the package colours, the sample size is large (240 observations) and both groups are of the same size and the test variable is distributed roughly unimodal. Hence, it is very unlikely that the results of the *t*-test are biased.

9.4.1.2 The *t*-Test for Two Independent Samples with Stata

Before calculating the *t*-test for two independent samples with Stata based on the sample file *Chapter9_chocolatepraline_colour_name_price.dta* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9), we must first check the requirements for applying the *t*-test. Select *Statistics → Summaries, tables, and tests → Classical tests of hypotheses → Robust equal-variance test*, and then enter the variable *sales* under *Variable* and the variable *colour* under *Variable defining comparison groups*. Clicking *OK* produces results from three statistical tests on the equality of variances: Levene's test (W0) and two variations of the Brown–Forsythe *F*-test (W50 and W10):

- $W0 = 13.666002\,\mathrm{df}(1, 238)\ \ \Pr > F = 0.00027109$
- $W50 = 9.140792\,\mathrm{df}(1, 238)\ \ \Pr > F = 0.00277339$
- $W10 = 13.408026\,\mathrm{df}(1, 238)\ \ \Pr > F = 0.00030872.$

All *p*-values—designated here as Pr > F—lie below the usual threshold of $\alpha = 5\%$, which means that the variances are unequal. As with SPSS, though significant deviations from the normal distribution arise for the package colours,[10] the sample size is large (240 observations) and both groups are the same size and the test variable is distributed roughly unimodal, so it is unlikely that the results are biased.

Now we can calculate the *t*-test for two independent samples. Select *Statistics → Summaries, tables, and tests → Classical tests of hypotheses → Two-group mean-comparison test* to open the window in Fig. 9.22, which shows the test variable (sales) and the grouping variable (colour). If the variances are unequal, select *Unequal Variances*, which calculates the *t*-test using a robust approach proposed by Satterthwaite (1946). If you select *Welch's approximation*, the *t*-test will be calculated using the approach proposed by Welch (1947). Each differs only in the way that it approximates the degrees of freedom.

The resulting mean values are $\bar{x}_{\text{blue}} = 9982.367$ for the blue packaging and $\bar{x}_{\text{yellow}} = 9606.067$ for the yellow packaging. The difference between these values is statistically significant, since the *p*-value—$\Pr(|T| > |t|) = 0.0000$—lies below the usual threshold of the significance level of $\alpha = 5\%$. Hence, we can be 95% confident that the sales difference lies between 264.146 and 488.454 packages.

[9] Based on the results of the Kolmogorov–Smirnov test and the Shapiro–Wilk test (see Sect. 9.6.2), we must reject the hypothesis of a normal distribution.

[10] Based on the results of the Kolmogorov–Smirnov test and the Shapiro–Wilk test (see Sect. 9.6.2), we must reject the hypothesis of a normal distribution.

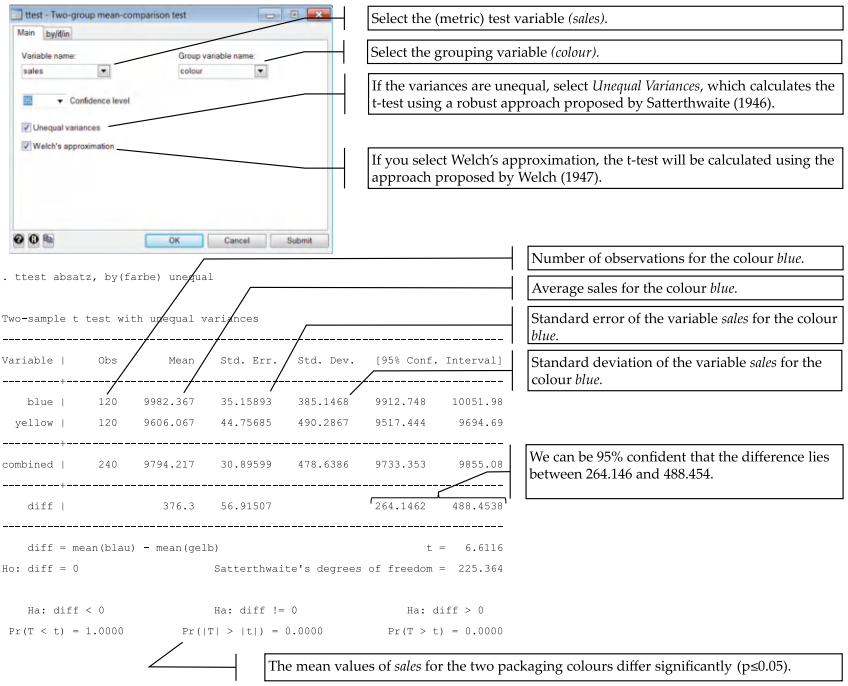


Fig. 9.22 The *t*-test for two independent samples with Stata

9.4.1.3 The *t*-Test for Two Independent Samples with R Commander

Using the example file *Chapter9_chocolatepraline_colour_name_price.RData* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9), we have to verify that the assumption of equal variances is fulfilled before calculating the two-sample *t*-test with R Commander. Select *Statistics* → *Variances* → *Levene's test*…

In the Levene's test window, select the *Mean* as the *Center*, the *sales* variable under *Response Variable*, and the *colour* variable under *Factors*. Clicking *OK* produces the following result: The *p*-value ($p = 0.00027$)—designated as Pr > F—lies below the usual threshold of $\alpha = 5\%$, which means that the variances are unequal (see upper part of Fig. 9.23). In this case, the use of the "robust" Welch *t*-test is recommended.

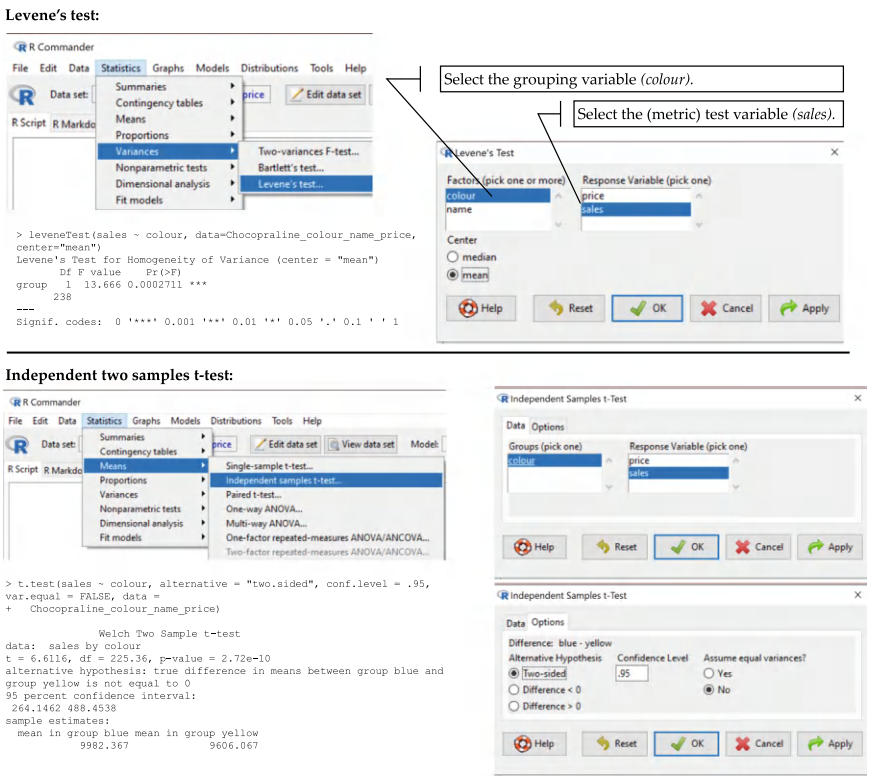
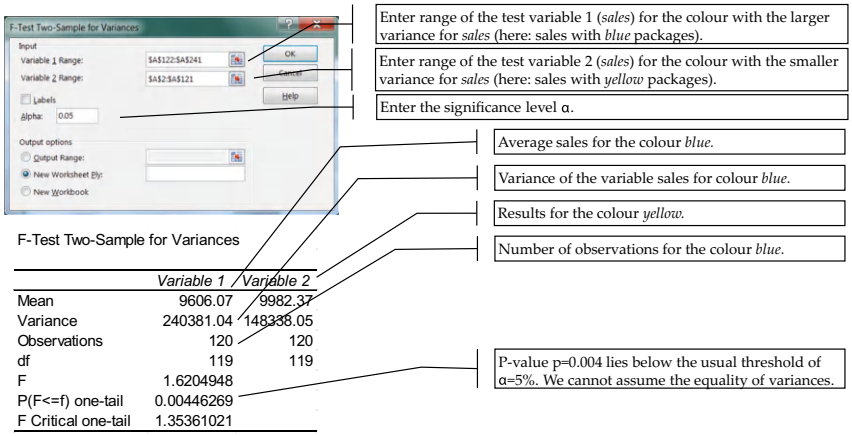


Fig. 9.23 The *t*-test for two independent samples with R Commander

Though significant deviations from the normal distribution arise for the package colours,[11] the sample size is large (240 observations), both groups are the same size and the test variable is distributed roughly unimodal, so it is unlikely that the results are biased.

Now we can calculate the *t*-test for two independent samples. Select *Statistics* → *Means* → *Independent samples t-test*… to open the window in the lower part of Fig. 9.23, which shows the selected *Response Variable* (*sales*) and the *Groups* variable (*colour*). As the variances are unequal in our case, select the option *No* in the option window for the question "Assume equal variances?". The *t*-test is then calculated using a robust approach proposed by Welch (1947). The resulting mean values are $\bar{x}_{\text{blue}} = 9982.367$ for the blue packaging and $\bar{x}_{\text{yellow}} = 9606.067$ for the yellow packaging. The difference between these values is statistically significant, since the *p*-value ($p < 0.00001$) lies below the usual threshold of the significance

[11] Based on the results of the Kolmogorov–Smirnov test and the Shapiro–Wilk test (see Sect. 9.6.2), we must reject the hypothesis of a normal distribution.



F-Test Two-Sample for Variances

	Variable 1	*Variable 2*
Mean	9606.07	9982.37
Variance	240381.04	148338.05
Observations	120	120
df	119	119
F	1.6204948	
P(F<=f) one-tail	0.00446269	
F Critical one-tail	1.35361021	

Fig. 9.24 Testing for equality of variance with Excel

level of $\alpha = 5\%$. Hence, we can be 95% confident that the sales difference lies between 264.1462 and 488.4538 packages.

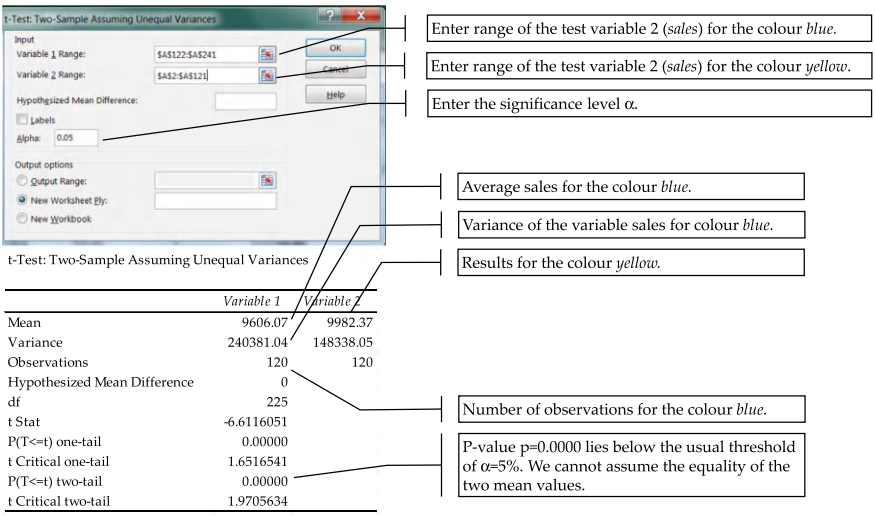
9.4.1.4 The *t*-Test for Two Independent Samples with Excel

Open the dataset *Chapter9_chocolatepraline_colour_name_price.xls* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9). The test with Excel requires that we first arrange the data in cohesive data blocks with regard to their grouping variables. The simplest way to do this is to select the entire dataset and go to *Data* → *Sort*… This arranges the data by grouping variables in ascending or descending order and sequencing the groups successively.

To check the equality of variance assumption, make sure that the add-in modules Analysis-ToolPak and Analysis-ToolPak—VBA are activated[12] and then select *Data* → *Analysis* and the function *F-Test Two-Sample for Variances*.In the window, enter the range of the test variable *sales* of the first colour under *Variable 1 Range* and the range of the test variable *sales* of the second colour under *Variable 2 Range*. Please note that Excel will produce correct values only if the variable with the larger variance is entered under *Variable 1 Range* and the variable with the smaller variance is entered under *Variable 2 Range*. Next to *Alpha* indicate the desired significance level (usually $\alpha = 0.05$). Based on the results in Fig. 9.24, we can see that the empirical test value $F = 1.62$ is larger than the critical F value $F_{\text{crit}} = 1.35$ (calculated based on $\alpha = 5\%$). Because the p-value—$p = 0.004$—lies below the usual threshold of $\alpha = 5\%$, we cannot assume the equality of variances.

To calculate the *t*-test for two independent samples with unequal variances, go to *Data* → *Analysis* → *t-Test: Two-Sample Assuming Unequal Variances*. If the

[12] In Excel this can be reached by clicking *File* → *Options* → *Add-ins* → *Go*.



t-Test: Two-Sample Assuming Unequal Variances

	Variable 1	*Variable 2*
Mean	9606.07	9982.37
Variance	240381.04	148338.05
Observations	120	120
Hypothesized Mean Difference	0	
df	225	
t Stat	-6.6116051	
P(T<=t) one-tail	0.00000	
t Critical one-tail	1.6516541	
P(T<=t) two-tail	0.00000	
t Critical two-tail	1.9705634	

Fig. 9.25 The *t*-test for two independent samples with Excel

variances are equal, select *t-Test: Two-Sample Assuming Equal Variances.*[13] In the window (Fig. 9.25), enter the range of the test variables of the first group under *Variable 1 Range* and the range of the test variables of the second group under *Variable 2 Range*. Leave the space next to *Hypothesized Mean Difference* empty. The resulting mean values are $\bar{x}_{\text{blue}} = 9982.37$ for the blue packaging and $\bar{x}_{\text{yellow}} = 9606.07$ for the yellow packaging. The difference between these values is statistically significant, since the *p*-value—$P(T \leq t) = 0.0000$ on both tails—lies below the usual threshold of the significance level of $\alpha = 5\%$.

9.4.2 The Mann–Whitney *U* Test (Wilcoxon Rank-Sum Test)

The *U* test by Mann–Whitney (1947) is a nonparametric test for checking the differences between two independent samples (or groups) with regard to an ordinal or metric variable. Commonly known as the Wilcoxon rank-sum test (Wilcoxon, 1945), it does not have strict conditions that must be satisfied before being applied. The only requirement is that each comparison group follows roughly a similar distribution for the test variable. The *U* test should be applied if the *t*-test for two

[13] For a very good explanation of how to perform these steps using Excel, see https://www.youtube.com/watch?v=BlS11D2VL_U on the YouTube channel of *Jim Grange* or https://www.youtube.com/watch?v=X14z9r8FUKY on the YouTube channel of *Dr. James Clark* from the Kings College London (Essential Life Science Tutorials).

independent samples does not apply because the test variables violate its requirements (see step 4 in Sect. 9.4.1). Moreover, the Mann–Whitney U test is suited for comparing two groups by means of an ordinal—i.e. nonmetric—test variable.

To see how the Mann–Whitney U test works, consider a study of whether men or women are willing to spend more on a certain bottle of wine. The responses of twelve men and thirteen women are placed on a five-level ordinal scale (1 = €0–5; 2 = €5.01–10; 3 = €10.01–15; 4 = €15.01–20; 5 = €20.01–25). Figure 9.26 shows the results.

Step 1: Formulate the Hypotheses

The question is whether men and women differ with regard to how much they are willing to pay for a bottle of wine. Hence, we need to test the null hypothesis that the average rank scores of the male and female groups do not differ against the alternative hypothesis that they do:

$$H_0 : E(\overline{R}_1) = E(\overline{R}_2)$$

Willingness to pay	Gender
€ [0-5]	female
€ [0-5]	female
€ [0-5]	female
€ [0-5]	female
€]5-10]	female
€]5-10]	female
€]5-10]	female
€]10-15]	female
€]10-15]	male
€]10-15]	female
€]10-15]	male
€]10-15]	male
€]10-15]	male
€]10-15]	female
€]10-15]	female
€]15-20]	female
€]15-20]	male
€]15-20]	male
€]15-20]	female
€]15-20]	male
€]15-20]	male
€]15-20]	male
€]15-20]	male
€]15-20]	male
€]20-25]	male

Step 1: Assign rank scores

Rank willingness to pay
2,5
2,5
2,5
2,5
6
6
6
11,5
11,5
11,5
11,5
11,5
11,5
11,5
11,5
20
20
20
20
20
20
20
20
20
20
25

Step 2: Place rank scores in the appropriate male or female group

male	female
11,5	2,5
11,5	2,5
11,5	2,5
11,5	2,5
20	6
20	6
20	6
20	11,5
20	11,5
20	11,5
20	11,5
25	20
	20
211	114

Step 3: Calculation

	Gender	N	Mean Rank	Sum of Ranks
Willingness to pay	male	12	17.58	211
	female	13	8.77	114
	Total	25		

Test Statistics[a]

	Willingsness to pay
Mann-Whitney U	23.000
Wilcoxon W	114.000
Z	-3.126
Asymp. Sig. (2-tailed)	.002
Exact Sig. [2*(1-tailed Sig.)]	.002[b]

a Grouping Variable: Gender
b Not corrected for ties.

Fig. 9.26 Mann–Whitney U test

$$H_1 : E(\overline{R}_1) \neq E(\overline{R}_2)$$

Step 2: Determine the Significance Level α

Next, we must define α, i.e. the maximum permissible probability that we reject an H_0 that is true. The significance level α is usually set at 1, 5, or 10%, though 5% is the most common value, which is what we will use here ($\alpha = 0.05$).

Step 3: Draw a Sample

Researchers draw a sample with a size $n_1 = 12$ from population #1 (male) and a sample with a size $n_2 = 13$ from population #2 (female).

Step 4: Check the Test Requirements

• Interval or ordinal scale of measurement of the test variable	✓
• Random sampling from two defined populations	✓
• Samples are independent and there is no overlap between group members	✓
• Both groups should have the same shape distributions for the test variable	✓

Step 5: Determine the Empirical Test Value

First, we need to assign rank scores and factor in any tied ranks. In our example, four respondents (all women) selected the lowest payment preference, so that each receives a rank score of $((1 + 2 + 3 + 4)/4) = 2.5$ (see step 1 in Fig. 9.26). Three respondents (again, all women) indicated the second lowest amount. Each of them is assigned a rank score of $6 = ((5 + 6 + 7)/3)$. Eight respondents (four men and four women) provided responses that fall in the third category. Their average rank score is $11.5 = ((8 + 9 + 10 + 11 + 12 + 13 + 14 + 15)/8)$. We proceed in similar fashion for the remaining two payment categories. Then we place the rank scores in the appropriate male or female group and arrange them by size (see step 2 in Fig. 9.26).

Adding up the results, we see that the men have a rank sum of 211 and that the women have a rank sum of 114. Next, we must calculate the average rank score, which takes into account that the number of observations in each group can differ (as in our example). On average, the men have a rank score of 17.6 (= 221/12) and the women have a rank score of 8.8 (= 114/13). If the male and female groups in the study were willing to spend the same on the wine, then their average rank scores would be approximately the same. Obviously, this is not the case, though we still need to test whether the difference between the average rank scores is large enough to be considered statistically significant.

First let us determine the empirical test value U. This is where the Mann–Whitney test comes in. It calculates the frequency with which a rank score in one group is smaller than the rank scores in the other group. Identical rank scores are counted at half their value. We start by comparing the first rank score in the male group—11.5—with all 13 rank scores in the female group. This value is larger in seven instances, identical in four instances, and smaller in two instances. The U

value for the first rank score is therefore $2 + 4/2 = 4$. We proceed in this way for all the other rank scores in the male group, producing a final U value of

$$U = 4 + 4 + 4 + 4 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 0 = 23. \tag{9.42}$$

Hence, in 23 of 156 ($= n_1 \cdot n_2 = 12 \cdot 13$) possible comparisons, the rank scores in the male group are smaller than the female group; and in 133 ($= 156 - 23$) comparisons, the rank scores in the male group are larger than the female group. Briefly, U and U' can be calculated as follows:

$$\begin{aligned} \text{Male}: U &= n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - \text{ranksum}_{\text{male}} \\ &= 12 \cdot 13 + \frac{12 \cdot (12 + 1)}{2} - 211 = 23 \end{aligned} \tag{9.43}$$

$$\begin{aligned} \text{Female}: U\prime &= n_1 \cdot n_2 + \frac{n_2 \cdot (n_2 + 1)}{2} - \text{ranksum}_{\text{female}} \\ &= 12 \cdot 13 + \frac{13 \cdot (13 + 1)}{2} - 114 = 133 \end{aligned} \tag{9.44}$$

Step 6: Determine the Critical Test Value

If men and women gave the same responses, a comparison of the groups would show just as many smaller rank scores as larger ones, and the values for U and U' would be approximately the same, resulting in an expected value of $E(U) = 156/2 = 78$.

Now we can use these calculations to perform a Mann–Whitney U test. In most cases, computer software carries out the exact test only when the samples are small. If each of the group samples are larger than 20, then the software approximates the normal distribution of the U values using the expected value $E(U)$.[14]

Step 7: Make a Test Decision

In our example, the sample size is small. Hence, we see the results of both the exact U test and the asymptotic significance of the approximation in Fig. 9.26. In both cases, we would err in $p = 0.2\%$ of cases when assuming that men and women are willing to spend different amounts on the wine. Accordingly, we must reject the null hypothesis that men and women are willing to spend the same on the wine, as the p-value lies below all usual thresholds (1, 5, or 10%).

Had we used the t-test for independent samples, the results would have been approximately the same at $p = 0.01\%$, though we cannot assume a normal distribution.[15] The reason lies in the fact that the independent t-test responds robustly

[14] See Conover (1980). For more information on accurately calculating the approximated U test, see Bortz et al. (2000, p. 200).

[15] Based on the results of the Kolmogorov–Smirnov test and the Shapiro–Wilk test, we must reject the hypothesis of a normal distribution.

when the normal distribution requirement has not been satisfied. This raises the general question: when should we use a *U* test and when should we use a *t*-test? The following points can help us make this decision:

- If the distributions of the comparison groups are similar, then the results of the *U* test primarily express differences in the central tendency of the test variable for both groups. The more the distributions of the groups differ, the more the results of the *U* test express differences in the distribution shape and central tendency of the test variable.
- With a normally distributed population, the *U* test is relatively inefficient compared with the independent *t*-test. It is asymptotically efficient for 95% of cases (Dixon, 1954), which means that a *U* test requires a sample size of 40 instead of a sample size of 38 for an independent *t*-test.
- When there is no normally distributed population, the *U* test is always more efficient than the *t*-test (Witting, 1960).

When in doubt, I recommend using both tests, since computers make their calculation easy. Each serves as a good backup for drawing conclusions when the results are similar. If the tests come to different conclusions, it is important to think about why this occurred.

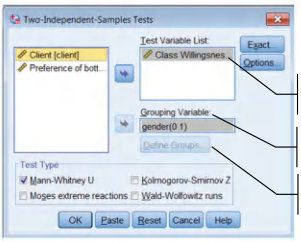
9.4.2.1 The Mann–Whitney *U* Test with SPSS

To perform the Mann–Whitney *U* test for the sample file *Chapter9_rank (Wine Bottle).sav* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9) in SPSS, select *Analyze → Nonparametric Tests → Legacy Dialogs → 2 Independent Samples…*. This opens the window shown on the left in Fig. 9.27. Next add the grouping variable (gender) to the *Grouping Variable* and indicate the group numbers for comparison (group zero for male and one for female) under *Define Groups*. Enter the test variable (Willingness-to-pay) in the *Test Variable List*. Under *Exact…* indicate whether SPSS should calculate an exact or an asymptomatic *U* test. For small samples, the test results for both tests are calculated automatically.

In both cases, we err in $p = 0.2\%$ of cases when we assume that men and women are willing to spend different amounts on the wine. The female group has an average rank score of 8.77, which is lower than the male group, which has an average rank score of 17.58.

9.4.2.2 The Mann–Whitney *U* Test with Stata

To perform the *U* test for the example file *Chapter9_rank (Wine Bottle).dta* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9) in Stata, select *Statistics → Nonparametric analysis → Tests of hypotheses → Wilcoxon rank-sum test*. This opens the window shown in Fig. 9.28. Next enter the grouping variable (*gender*) in the *Grouping variable* field. Grouping variables may have no more than two possible values. Under *Variable* enter the test variable (*willingness-to-pay*).

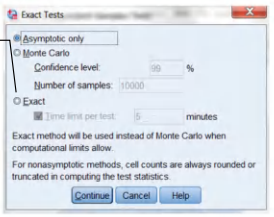


Select asymptotic or exact U test.

Select the (metric or ordinal) test variables *(Willingness to pay)*.

Select the grouping variable *(gender)*.

Select the group numbers for comparison: 0 for *male* and 1 for *female*.



	Gender	N	Mean Rank	Sum of Ranks
Willingness to pay	male	12	17.58	211
	female	13	8.77	114
	Total	25		

Mean rank of the *Willingness to pay* for male participants.

Mean rank of the *Willingness to pay* for female participants.

Test Statistics[a]

	Willingness to pay
Mann-Whitney U	23.000
Wilcoxon W	114.000
Z	-3.126
Asymp. Sig. (2-tailed)	.002
Exact Sig. [2*(1-tailed Sig.)]	.002[b]

a Grouping Variable: Gender
b Not corrected for ties.

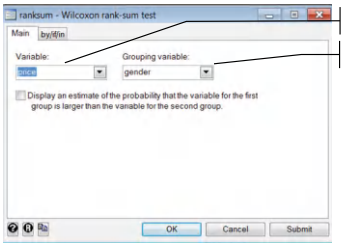
Calculated U value.

Calculated Z-value.

P-value of the asymptotic U test: we err in p=0.2% (≤0.05) of cases when we assume that men and women are willing to spend different amounts on the wine.

P-value of the exact U test not corrected for ties: we err in p=0.2% (≤0.05) of cases when we assume that men and women are willing to spend different amounts on the wine.

Fig. 9.27 The Mann–Whitney *U* test in SPSS



Select the (metric or ordinal) test variables *(Willingness to pay)*.

Select the grouping variable *(gender)*.

```
. ranksum preis_k1, by(gesch)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

      Gender |      obs     rank sum    expected
-------------+---------------------------------
        male |       12          211         156
      female |       13          114         169
-------------+---------------------------------
    combined |       25          325         325

unadjusted variance      338.00
adjustment for ties      -28.34
                         ------
adjusted variance        309.66

Ho: price(gender==male) = price(gender==female)
           z =   3.126
    Prob > |z| =   0.0018
```

Expected rank for male participants if no difference is assumed.

Expected rank for female participants if no difference is assumed.

Calculated Z-value.

P-value of the asymptotic U test corrected for ties: we err in p=0.18% (≤0.05) of cases when we assume that men and women are willing to spend different amounts on the wine.

Fig. 9.28 The Mann–Whitney *U* test with Stata

According to the results, we err in $p = 0.18\%$ of cases when we assume that men and women are willing to spend different amounts on the wine. The actual rank sum of the female group is 114. This is lower than 169, the expected rank sum if men and women had the same preferences. Hence, the female respondents were willing to pay significantly less on the wine than the males.

9.4.2.3 The Mann–Whitney *U* Test with R Commander

Select *Statistics* → *Nonparametric tests* → *Two-sample Wilcoxon test*… to perform the *U* test for the example file *Chapter9_rank (Wine Bottle).RData* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9). This opens the window shown on the right in Fig. 9.29. Next, enter the grouping variable (*gender*) in the *Groups* field. Grouping variables may have no more than two possible values. Under *Response Variable* enter the test variable (*price*).

According to the results, we err in $p = 0.18\%$ of cases when we assume that men and women are willing to spend different amounts of money on the wine. The median willingness-to-pay for group 0 (men) is 4, the median willingness-to-pay for group 1 (women) is 2. Hence, the female respondents were willing to pay significantly less on wine than the males.

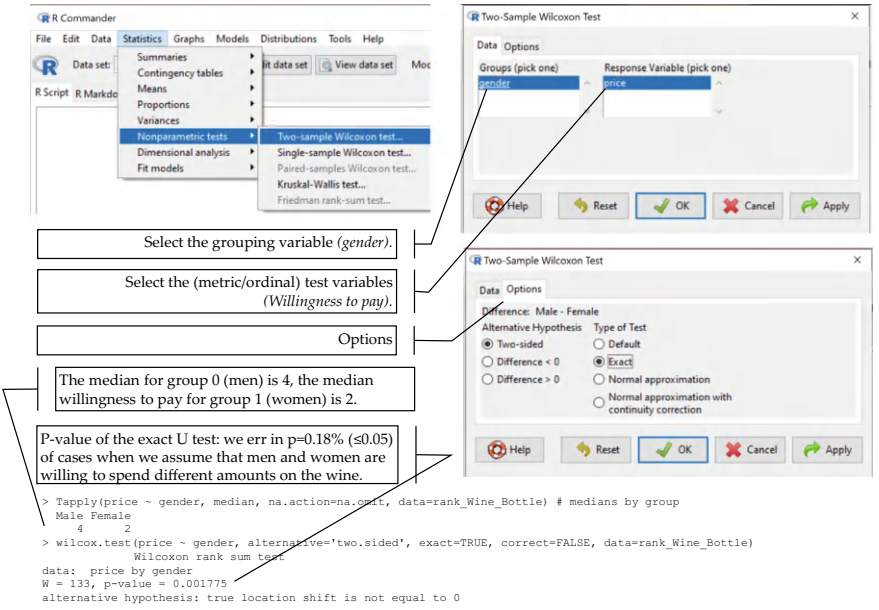


Fig. 9.29 The Mann–Whitney *U* test with R Commander

9.5 Tests for *k* Independent Samples

9.5.1 Analysis of Variance (ANOVA)

As we learned in the previous chapter, the *t*-test for two independent samples allows us to check the significance of the difference between the mean values of two groups. Say we have a dataset that recorded the gender and annual purchase amount for each customer of a certain product. The *t*-test for independent samples can identify whether annual purchase amounts for men and women differ significantly from one another. But what if we want to compare the mean values of more than two groups? How, for instance, would we determine whether average unit sales for three different product variants vary by supermarket?

A first obvious solution could be to perform several *t*-tests for independent samples. For a simple (single factor) example with three groups (product variants), we would need to carry out $\binom{3}{2} = 3$ *t*-tests; with four groups (product variants), we would need $\binom{4}{2} = 6$ *t*-tests; and with five groups (product variants), we would need $\binom{5}{2} = 12$ *t*-tests. If the number of groups to be compared is large, the number of required *t*-tests becomes unmanageable. Moreover, as the number of required *t*-tests increases, so too does the likelihood of accidental significance with one of the tests.

For this reason, comparisons of more than two groups must be performed using Analysis of Variance, or ANOVA for short. There is not one, single form of ANOVA; many exist, and which one to use depends on the situation (see Fig. 9.30).

9.5.1.1 One-Way Analysis of Variance (ANOVA)

Let us take a closer look at single-factor ANOVA (also called one-way ANOVA), the simplest analysis of variance technique. Suppose a company wants to introduce a new type of praline to the market. The management is still undecided about the name. The top three favourites are *Chocolate Dream*, *Choc Champ*, and *Beethoven Bonbons*. A market study in 240 test markets investigates the effect of each name on sales.[16]

First we calculate the mean unit sales for each name and for all the observations (see Fig. 9.31). Average total product sales are 9794 units per market. The product sold under the name *Beethoven Bonbons* had above-average sales (9872 units), while *Chocolate Dream* and *Choc Champ* had below-average sales (9757 and 9753 units, respectively).

[16] See the file *Chapter9_Chocopraline_colour_name_price.sav* for SPSS, *Chapter9_Chocopraline_colour_name_price.RData* for R, and *Chapter9_Chocopraline_colour_name_price.dta* for Stata (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9).

		Number of factors (qualitative)		
		=1	>1	
Number of independent variables (metric)	=1	Single-factor univariate analysis of variance or one-way ANOVA[1)]	Multiple-factor univariate analysis of variance or two-way ANOVA[2)]	Univariate analysis of variance (ANOVA)
	>1	Single-factor multivariate analysis of variance or one-way MANOVA[3)]	Multiple-factor multivariate analysis of variance or two-way MANOVA[4)]	Multivariate analysis of variance (MANOVA)
		Single-factor ANOVA/MANOVA	Multiple-factor ANOVA/MANOVA	

Examples: What are the effects of each product variant (factor #1)...
1) ...on revenue levels (dependent variable #1)?
2) ...and the effects of different locations (factor #2) on revenue levels (dependent variable #1)?
3) ...on revenue levels (dependent variable #1) and unit sales (dependent variable 2)?
4) ...and different locations (factor #2) on revenue levels (dependent variable #1) and unit sales (dependent variable #2)?

Fig. 9.30 Overview of ANOVA

Product name	Mean unit sales	Standard deviation	N
Chocolate Dream	9757.2	593.5	80
Choc Champ	9753.4	425.1	80
Beethoven Bonbons	9872.0	388.4	80
Total	9794.2	478.6	240

Fig. 9.31 ANOVA descriptive statistics

When we present the sales figures graphically (see Fig. 9.32), we see that every point (observation in the markets) can be represented as a linear combination of the total mean ($\overline{Y}$), of the average group-specific deviation for one of the G different names from the total mean value (here product name "g": α_g), and of an observation-specific residual (ε_{gk}):

$$Y_{gk} = \overline{Y} + \alpha_g + \varepsilon_{gk} \tag{9.45}$$

The entire variance of the sales (S_y^2) can be broken down into a share of the variance explained by the factor-level product name (S_X^2) and an unexplained share of variance (S_e^2). According to Backhaus et al. (2021, p. 178), this results in the

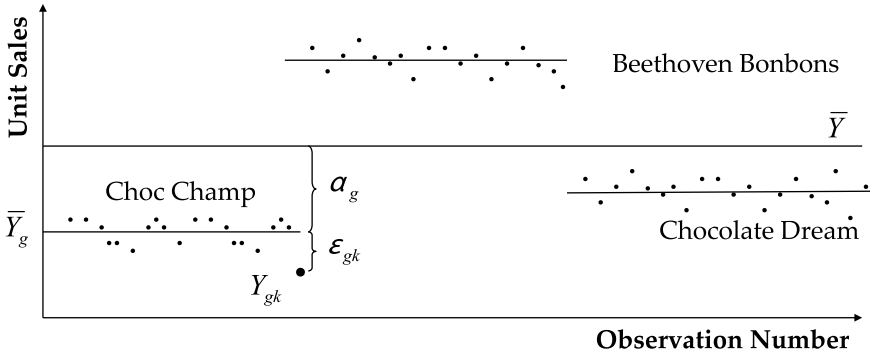


Fig. 9.32 Graphic visualization of a one-way ANOVA

following equation for G factor steps and K observation values within a factor step

$$\frac{\sum_{g=1}^{G}\sum_{k=1}^{K}\left(Y_{gk}-\overline{Y}\right)^2}{G\cdot K-1}=\frac{K\sum_{g=1}^{G}\left(\overline{Y}_g-\overline{Y}\right)^2}{G-1}+\frac{\sum_{g=1}^{G}\sum_{k=1}^{K}\left(Y_{gk}-\overline{Y}_g\right)^2}{G\cdot(K-1)} \tag{9.46}$$

ANOVA determines whether at least one α_g exists that is significantly different than zero, i.e. whether at least one product name differs significantly from the total mean value. Its objective is to test the following hypothesis[17]:

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \ldots = \alpha_G = 0 \text{ versus} \tag{9.47}$$

$$H_1 : \text{ There exists at least one } \alpha_g \neq 0. \tag{9.48}$$

The larger the variance explained by factor levels (S_X^2) is compared with unexplained variance (S_e^2), the greater the likelihood that at least one mean value of an individual group (e.g. the product name) will differ significantly from the mean values of other groups. The quotient of two variances follows an F-distribution:

$$F_{\text{emp}} = \frac{S_X^2}{S_e^2} \tag{9.49}$$

The larger this value tends to be, the larger the influence of factors (e.g. product name) on the dependent variable (e.g. sales)—in dependence on the degrees of freedom (df).

[17] The result of a single-factor univariate ANOVA for a factor with two traits is the same as that of a *t*-test for independent samples.

Source	Type III Sum of squares	df	Mean of square	F	Sig.
Corrected model	726,592.1[a]	2	363,296.1	1.6	.205
Constant term	23,022,403,227.3	1	23,022,403,227.1	100,992	.000
Name	726,592.1	2	363,296.1	1.6	.205
Error	54,027,080.6	237	227,962.4		
Total	23,077,156,900.0	240			
Corrected total	54,753,672.7	239			

a. R-squared = .013 (Adjusted R-squared = .005)

Fig. 9.33 ANOVA tests of between-subjects effects (SPSS)

The results from Fig. 9.33 show that the corrected model is insignificant. This means that the model's variables—up to now the only variable has been *name*—have no influence on sales. The *p*-value of the variable *name*—0.205 < (α = 0.05)—shows this clearly. We would err in 20.5% of cases if we assumed that name has an influence on sales. The differences among the mean values identified above are not large enough to be considered statistically significant; they might have also occurred by chance.

Statistics software such as SPSS or Stata makes the analysis of variance fairly simple. But even if we use software, we must make sure that the following requirements are met before carrying out ANOVA:

- Independent observation values. This is the most important requirement for ANOVA. Dependent observations occur when correlating reactions and answers result (Hair et al., 1998, p. 348)—for instance, as part of a group discussion or interview. Special conditions (noise, unclear instructions, or questions) are responsible for dependent observations.
- Homogeneity of variance. This requirement can be tested using Levene's test or Bartlett's test.[18] If the test is insignificant, it means that the variances of individual groups (e.g. product names) do not differ significantly from each other. If the observed groups are roughly the same size, then the requirement of variance homogeneity need not be met (Bortz, 1999, p. 276). When the groups have different sizes and there is no homogeneity of variance, however, ANOVA may not be applied. In this case, we have to resort to the nonparametric

[18] Traditionally, the *F*-test (or its application to groups, Bartlett's test) is used to measure equal variance. But these tests react very sensitively to deviations from the normal distribution, which is why the more robust Levene's test is preferred. SPSS automatically performs Levene's test for equal variance when performing ANOVA (choose *Options* → *Homogeneity tests*). Stata uses the one-way ANOVA to determine Bartlett's test for equal variances. The Levene's test calculation (w_0) is located under the heading *Hypothesis Tests*. In R Commander, the Levene's test can be found under *Statistics* → *Variances* → *Levene's test*…

Kruskal–Wallis test (see Sect. 9.5.2) or to modified and robust tests such as the Welch test (1947) or the Brown–Forsythe test (1974).

- Normality of the residuals. The residuals must follow a normal distribution with a zero mean and a constant variance in each of the factor categories (e.g. for each of the product names). It should be noted that the ANOVA *F*-Test is fairly robust against the violation of this assumption. As long as the distributions of the error terms are not extremely different from a normal distribution, the *p*-value of the ANOVA is not meaningfully affected. This is particularly true for larger sample. The greater the sample size, the less important the assumption of normality is. With large samples, a normal distribution is assured through the effects of the central limit theorem (regardless of the distribution of the raw data). In the case of low variance in the raw data, ten to twenty observations per cell are sufficient to ensure the effects of the central limit theorem; in the case of very strong variance, some 50 observations per cell are required (Stevens, 1972).

Levene's test or Bartlett's test for the above example shows a significant difference among the variance of unit sales for individual product names. This violation of variance homogeneity indicates that ANOVA should not have been used. Since the group sizes are all the same, however, the violation hardly affects the results. Something similar applies for the violation of the normal distribution. Though the Kolmogorov–Smirnov test produces significant deviations from the normal distribution result for all the names (*Chocolate Dream*, *Beethoven Bonbons*, and *Choc Champ*), the sample size of 240 observations is large enough to refrain from the normal distribution requirement. Nevertheless, the results should be checked again using the Kruskal–Wallis test which, as it happens, confirms the results.

9.5.1.2 Two-Way Analysis of Variance (ANOVA)

As noted, it is rare that a single factor alone influences a particular outcome; usually a whole array of causes plays a role. For this reason, let us expand our example by including information on packaging colour. Say the company ignores the insignificance of product names and chooses *Beethoven Bonbons* based on its years of experience. To see which packaging colour is most useful, we must carry out a multiple-factor or two-way ANOVA (see Fig. 9.33). In this case, Levene's test produces an insignificant result ($p = 0.476$), i.e. our two-factor model does not violate the requirement of homogeneity of variance.

In the two-factor model, individual observations can be represented as a linear combination of the total mean value ($\overline{Y}$), the average name-specific deviation (here product name "*g*": α_g), the average colour-specific deviation for packaging (here colour "*h*": β_h), and an observation-specific residual (ε_{ghk}):

$$Y_{ghk} = \overline{Y} + \alpha_g + \beta_h + \varepsilon_{ghk} \tag{9.50}$$

Yet there is another effect as well: the interaction between product name and product colour ($\alpha\beta_{gh}$). The linear combination exhibits the following form:

$$Y_{ghk} = \overline{Y} + \alpha_g + \beta_h + \alpha\beta_{gh} + \varepsilon_{ghk} \tag{9.51}$$

This produces the *p*-values and mean values in Fig. 9.34.

The results from Fig. 9.34 show that the corrected model is significant. This means that at least one of the independent variables exerts an influence on sales. The *p*-value for the variable *colour* is $p = 0.000 < (\alpha = 0.05)$. This means that we would err in less than 5% of cases if we assumed that packaging colour has an influence on sales. The colour blue, with an average of 9983 units sold, is more successful than the colour yellow, which has 9606. In our two-factor model the product name is significant ($p = 0.000 < (\alpha = 0.05)$), too. The name *Beethoven Bonbons* was best received, with an average of 9872 units sold. In the initial

Product name	Product colour	Mean value	Standard deviation	N
Chocolate Dream	Blue	10,305.8	223.9	40
	Yellow	9,208.6	214.7	40
	Total	9,757.2	593.5	80
Choc Champ	Blue	10,118.7	165.3	40
	Yellow	9,388.2	255.2	40
	Total	9,753.4	425.1	80
Beethoven Bonbons	Blue	9,522.6	178.9	40
	Yellow	10,221.4	152.3	40
	Total	9,872.0	388.4	80
Total	Blue	9,982.4	385.1	120
	Yellow	9,606.1	490.3	120
	Total	9,794.2	478.6	240

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected model	45,242,582.7[a]	5	9,048,517	222.6	.000
Constant term	23,022,403,227.3	1	23,022,403,227	566,416.9	.000
Name	726,592.1	2	363,296	8.9	.000
Colour	8,496,101.4	1	8,496,101	209.0	.000
Name * Colour	36,019,889.2	2	18,009,945	443.1	.000
Error	9,511,090.0	234	40,646		
Total	23,077,156,900.0	240			
Corrected total	54,753,672.7	239			

a. R-squared = .826 (Adjusted R-squared = .823)

Fig. 9.34 ANOVA tests of between-subjects effects and descriptive statistics

"naïve" interpretation, the combination of the average best colour and the average best product name yields an optimal product configuration: blue packaging and *Beethoven Bonbons*. But this naïve interpretation ignores the interaction effects mentioned above.

With multiple-factor (two-way) ANOVA, interaction effects should always be taken into account. With two-factor models, there is one potential interaction effect; with three-factor models, there are up to three; with four-factor models, up to six. ANOVA indicates whether these are significant. But what is the interaction effect, exactly? It is best illustrated by representing the marginal means. The interaction effect—i.e. the interaction between two factors—can take a variety of forms (see Fig. 9.35).

If there is no interaction effect, the plots for the marginal means are nearly parallel (see case 1 in Fig. 9.35). In this hypothetic case, the colour with the largest average value—in our example, blue—lies independent from the name always above yellow. The distance between the parallels equals the difference between the average effect on sales for each colour. The effects of colour and the effects of name on sales are independent of each other. The difference between a *Beethoven Bonbon* packaged in blue and one packaged in yellow is the same as the difference between a *Chocolate Dream* packaged in blue and one packaged in yellow.

If, as in our example, there is a significant interaction effect, then the lines do not run in parallel (see Fig. 9.36). The colours are not only different in themselves; they have different strengths depending on the product name. In our example, it is

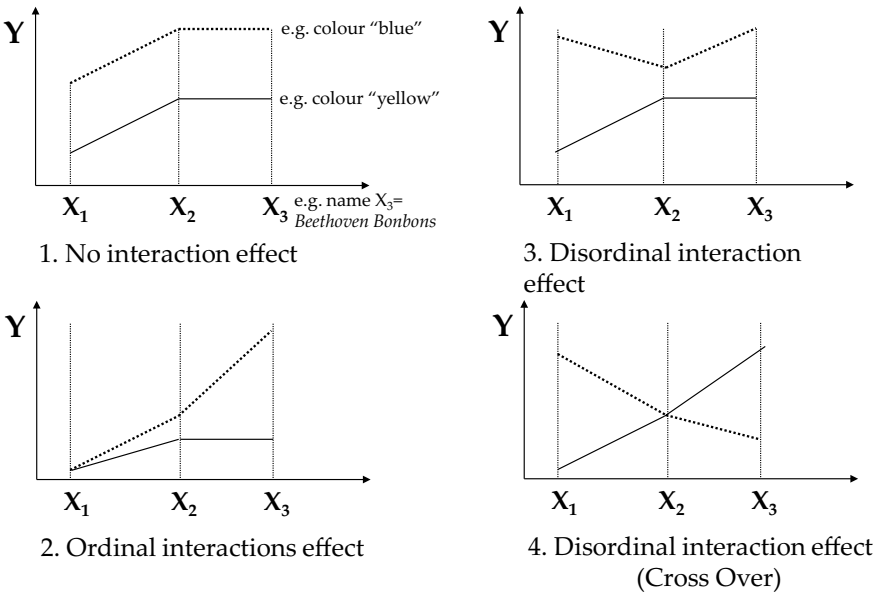


Fig. 9.35 Interaction effects with multiple-factor ANOVA

clear that the colour blue yields better sales for *Chocolate Dream* and *Choc Champ* than the colour yellow. Between the names, blue has a stronger effect on *Chocolate Dream* than it does on *Choc Champ*. With *Beethoven Bonbons*, by contrast, blue has a negative effect. Yellow is the better choice—a different result than suggested to us by our "naïve" assessment of the total mean values. This is the typical *cross-over* case (see case 4 in Fig. 9.35). When interaction effects are significant, then, it is important that we take into account marginal means as well as total mean values.

If factor levels differ from each other significantly with regard to customer preference, we must determine which produce a significantly higher customer preference, which produce a significantly lower customer preference, and which have the same. To use individual *t*-test combinations for this purpose would—as discussed—increase the likelihood of an erroneous assumption of differences known as a Type 1 error. But there are a variety of techniques available for controlling this kind of error. These multiple comparison procedures are frequently referred to as post-hoc methods. In Fig. 9.37, Scheffé's method is carried out for the given example. The product name *Beethoven Bonbons* differs from the other names at

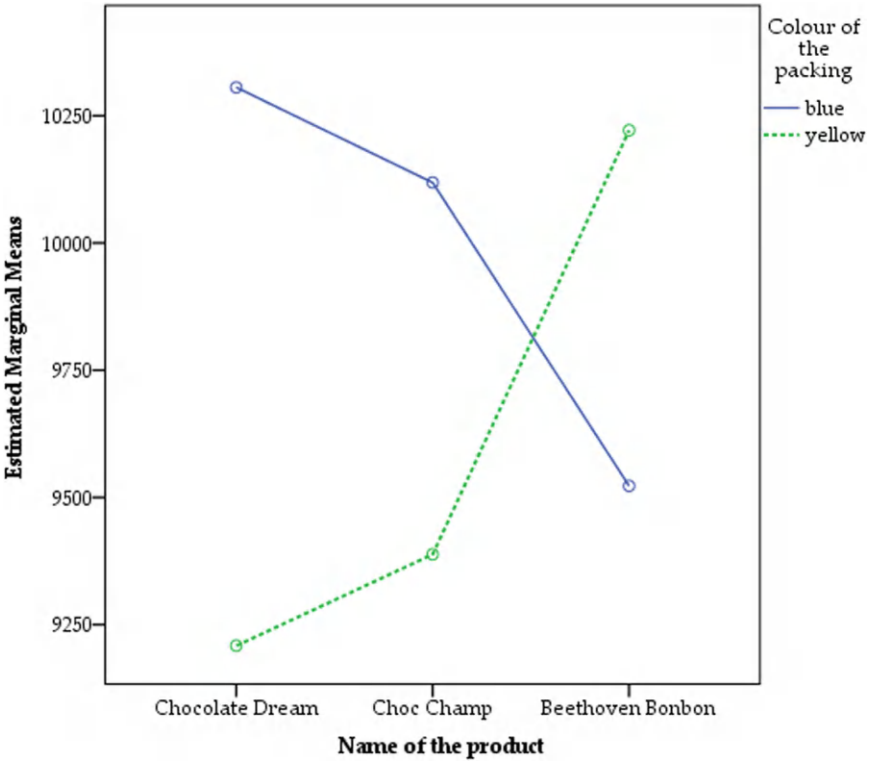


Fig. 9.36 Estimated marginal means of unit sales

(I) Name of the product	(J) Name of the product	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Chocolate Dream	Choc Champ	3,8	31.9	.993	-74.779	82.279
	Beethoven Bonbons	-114.8(*)	31.9	.002	-193.33	-36.271
Choc Champ	Chocolate Dream	-3.8	31.9	.993	-82.279	74.779
	Beethoven Bonbons	-118.6(*)	31.9	.001	-197.079	-40.021
Beethoven Bonbons	Chocolate Dream	114.8(*)	31.9	.002	36.271	193.329
	Choc Champ	118.6(*)	31.9	.001	40.021	197.079

* The mean difference is significant at the .05 level.]

Fig. 9.37 Multiple comparisons with Scheffé's method

the five per cent level (as indicated by the asterisk). The other names exhibit no significant differences between each other. Hence, the significance of the entire factor *product name* can be explained by the significantly higher results of the name *Beethoven Bonbons.*

The special feature of the Scheffé's method is that it does much to limit the occurrence of Type 1 errors. A given factor only turns out to be significant when individual group differences are more than evident. Of course, this can lead us to neglect actual differences (Scheffé, 1953, p. 87). In this regard, Scheffé's method is the most conservative post-hoc test. Studies have shown that multiple comparison procedures can be ranked from more conservative to less conservative as follows: (1) Scheffé, (2) Tukey HSD, (3) Tukey LSD, (4) Newman–Kuels, and (5) Ducan (Stevens, 1972). The larger the sample is and the fewer groups to be compared, the greater the likelihood that existing differences will come to light (Hair et al., 1998, p. 356).

9.5.1.3 Analysis of Covariance (ANCOVA)

Not all influencing variables possess a nominal (classifying) scale. For many questions, metric scales should be considered at the same time. Accordingly, the above example can be supplemented with the respective sales price (as an independent variable). Therefore, we are investigating the influence exerted by price—the covariable, also known as the covariate—and by the groups of different *colours* and *product names* on the dependent unit sales variable using *Analysis of Covariance* (ANCOVA).

The expansion of the two-factor model to an ANCOVA model allows every observation to be explained by the average name-specific and the colour-specific deviation (α_g, β_h) and the slope coefficient of the covariable price:

$$Y_{ghk} = \overline{Y} + \alpha_g + \beta_h + \alpha\beta_{gh} + \delta x + \varepsilon_{ghk} \tag{9.52}$$

δ represents the strength of the influence of independent metric variable (price) on the dependent variable (sales) for all groups identified by regression techniques. It assumes that the covariable influences the dependent variable in the same way for all groups.

In our example, do the name-specific and colour-specific deviations remain significant even when the price effect is taken into account beforehand? The requirements for ANCOVA are the same as those for ANOVA. Our sample satisfies the homogeneity of variance requirement ($p = 0.462$ in Levene's test).

As Fig. 9.38 shows, ANCOVA turns past findings on their head. In previous sections, where we had yet to factor in price, ANOVA led us to believe that name and colour influence sales. Now a completely different picture emerges with ANCOVA: the number of units sold is determined solely by price. A calculation of the price effect shows that the other variables are insignificant ($p = 0.924$ for *colour* and $p = 0.537$ for *name*), and thus possess no explanatory value. Compare the marginal means *with* price calculation in Fig. 9.39 and the marginal means *without* price calculation in Fig. 9.36. The insignificance of the other factors *name* and *colour* in Fig. 9.39 collapses the distance between the marginal means.

Since the number of degrees of freedom available for a significance test sinks with every additional covariable, the sample size must be increased to compensate. All explanatory covariables should be considered, but no more than necessary from a theoretical point of view. The rule of thumb for the maximum number of covariables is:

$$\text{Number of covariables} = (0.1 \cdot \text{sample size}) - (\text{number of groups} - 1) \qquad (9.53)$$

Once again, the analysis of covariance (ANCOVA) assumes that the covariable exerts the same influence on the dependent variables for all groups. Ultimately, this

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected model	54,659,399.1[a]	6	9,109,899.8	22515.4	.000
Constant term	1,435.8	1	1,435.8	3.5	.061
Price	9,416,816.4	1	9,416,816.4	23273.9	.000
Colour	3.7	1	3.7	.0	.924
Name	504.7	2	252.4	.6	.537
Colour * Name	1,082.8	2	541.4	1.3	.264
Error	94,273.6	233	404.6		
Total	23,077,156,900.0	240			
Corrected Total	54,753,672.7	239			

a. R-squared = .998 (Adjusted R-squared = .998)

Note: The influencing variables *product colour* and *product name* and their interaction effect are insignificant and thus have no influence on sales. The factor *price* is significant, however. The assumption that price influences sales is erroneous in less than 0.05% of cases.

Fig. 9.38 ANCOVA tests of between-subjects effects

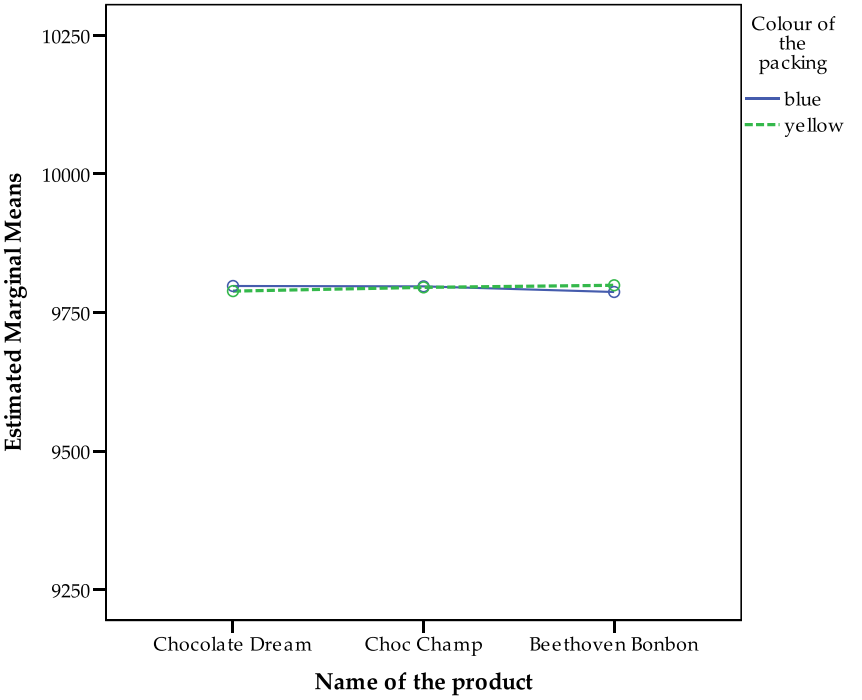


Fig. 9.39 Estimated marginal means for sales (ANCOVA)

means that ANCOVA is an ANOVA of regression residues between covariables and the dependent variable.[19] Putting it this way makes the relationship between ANOVA and regression analysis plain: the individual factors can be seen as dummy variables in a regression. But unlike ANOVA the regression can identify not only individual significances but also the marginal effects of the independent variables. Because regression extracts more information from the data than ANOVA does, there is good reason to use it right from the beginning.

As we already learned, ANOVA and ANCOVA are univariate techniques for variance analysis. Another approach to variance analysis is *Multivariate ANOVA*, also known as *MANOVA*. This can be used when investigating the influence of factors (and if needed, covariables) on multiple dependent variables at the same time in the same variance analysis. MANOVA is preferable to multiple univariate analysis only when dependent variables correlate sufficiently. For example, one could investigate whether a simultaneous influence exists between the independent

[19] Strictly speaking, all the measures of regression diagnostics (heteroscedasticity, autocorrelation, multicollinearity, etc.) should be performed when carrying out an ANCOVA (see Sect. 10.10).

factors *ad slogan* and *product name* on both dependent variables *quality perception* and *price perception*:

$$Y_1; Y_2 = f(X_1; X_2; \ldots; X_n) \tag{9.54}$$

Here we can determine the significance of an influence with the help of different key figures (Wilks' lambda, Hotelling's trace, Roy's root, Pillai's trace, etc.), but a detailed discussion of these figures and their derivations lie beyond the scope of this book.

9.5.1.4 ANOVA/ANCOVA with SPSS

This section uses the SPSS dataset *Chapter9_Chocopraline_colour_name_price.sav* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9). To calculate ANOVA/ANCOVA with SPSS, select *Analyze* → *General Linear Model* → *Univariate*…. This opens the *ANOVA/ANCOVA* dialogue box. Then assign the dependent variable (here: *sales*), the Fixed Factor(s) (here: *name* and *colour*), and the Covariate(s) (here: *price*). Then follow the steps outlined in Fig. 9.40.

9.5.1.5 ANOVA/ANCOVA with Stata

This section uses the Stata dataset *Chapter9_Chocopraline_colour_name_price.dta* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9). Select *Statistics* → *Linear models and related* → *ANOVA/MANOVA* → *Analysis of variance and covariance* to open the dialogue box. In the menu that opens, select the dependent variable and define the used independent factors (see Fig. 9.41). It assumes that all explanatory variables are categorical unless they are prefixed by "c".

9.5.1.6 ANOVA/ANCOVA with R Commander

This section uses the R dataset *Chapter9_Chocopraline_colour_name_price.RData* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9). First, we need to verify that the assumption of equal variances is fulfilled before calculating the two-sample *t*-test with R Commander. Select *Statistics* → *Variances* → *Levene's test*… In the Levene's test window, select the *Mean* as the *Center*, the *sales* variable under *Response Variable*, as well as the *colour* variable and the *name* variable simultaneously under *Factors*.

Then select *Statistics* → *Means* → *Multi-way Anova*… to open the dialogue box. In the menu that opens, select the *Response variable* (*sales*) and define the independent factors *colour* and *name* (see Fig. 9.42). Click *Graphs* → *Plot of means*… to obtain the graph of the marginal means.

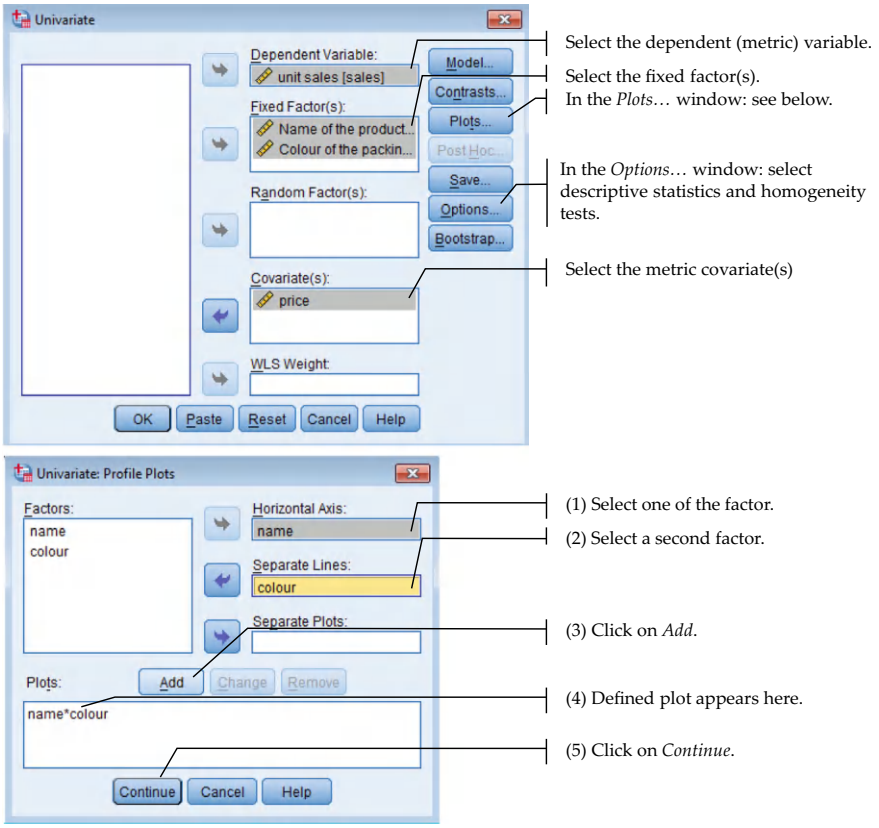


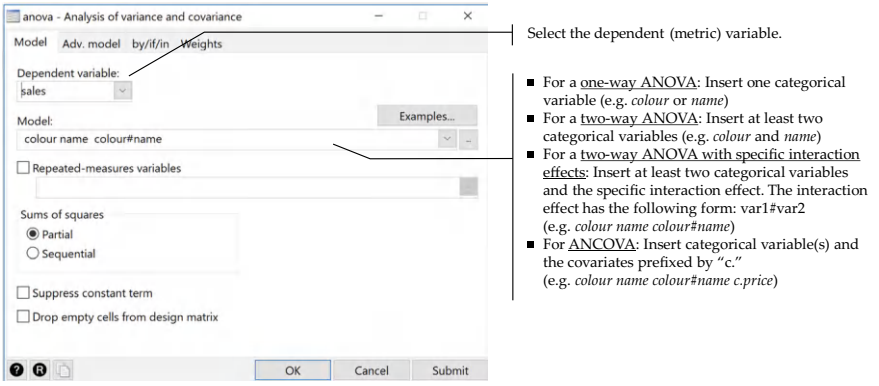
Fig. 9.40 ANOVA/ANCOVA with SPSS

9.5.1.7 ANOVA with Excel

To perform a single-factor or two-factor ANOVA in Excel, first make sure that the add-in modules Analysis-ToolPak and Analysis-ToolPak—VBA are activated.[20] If more factors are needed, then professional software packages [SPSS or Stata] should be used. Now check to see if the equality of variance assumption is satisfied using the *F*-test described in Sect. 9.3.1.4. This test compares two individual group variances. The more groups there are, the more two-sample tests must be run using the *F-test Two-Sample for Variances* tool. Excel does not have a single test (such as Levene's test) for testing the overall equality of variance.

Figure 9.43 shows the sale figures of pralines with different product names from the file *Chapter9_chocopraline_colour_name_price.xls* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9). The number of observations in each group need not be identical in the analysis. To test

[20] When using Excel, select *File* → *Options* → *Add-ins* → *GO* instead.



Levene's Test:
Statistics → Summaries, tables, and tests → Classical tests of hypotheses → Robust equal-variance test

Relevant syntax commands for analysis of variance:
anova; anova_postestimation; anovadef; hotelling; loneway; oneway; Manova; manova postestimation; sdtest; robvar

Fig. 9.41 Analysis of variance (ANOVA) with Stata

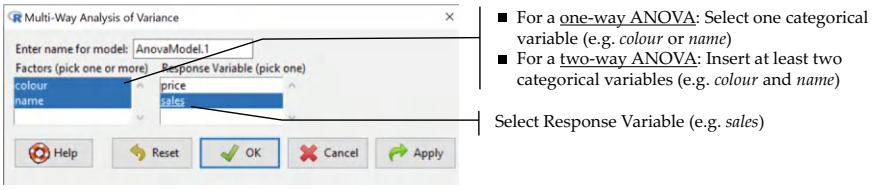
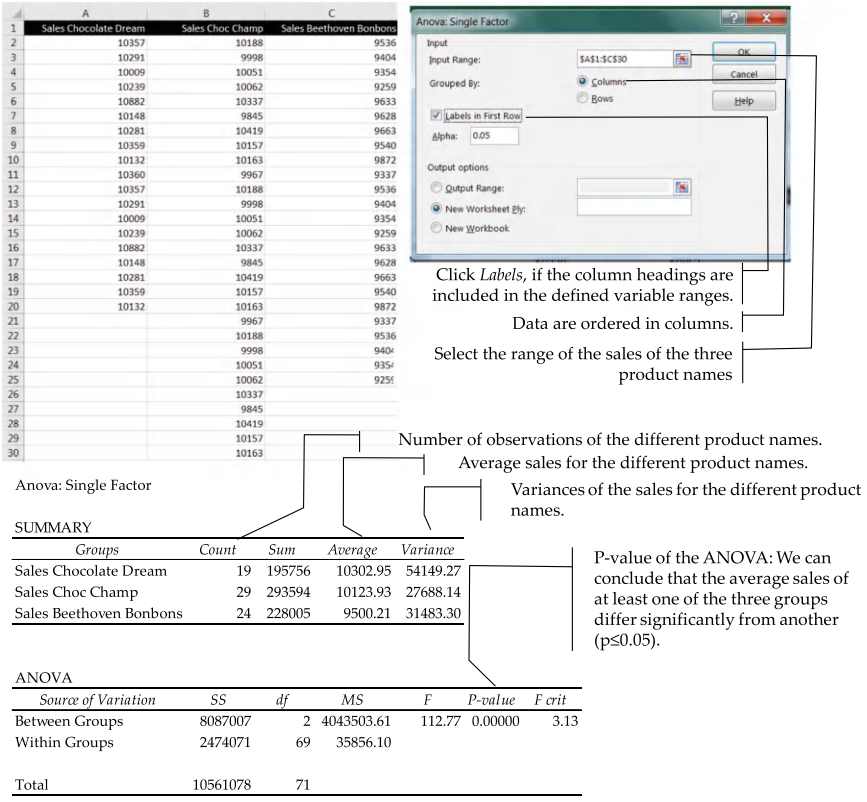


Fig. 9.42 Analysis of variance (ANOVA) with R Commander

whether the average sales figures of at least two product names differ, select *Data → Analysis → Anova: Single Factor* and enter the range of the sales of the three product names for comparison (*$A$1:$C$30*) in *Input Range*. Select the category *Columns* after *Grouped By* and check the box next to *Labels in first row*. Clicking *OK* produces the results shown in Fig. 9.43. Since the *p*-value is $0.00 \leq 0.05$, we can conclude that the average sales figures of at least one of the three groups differ significantly from another.[21]

[21] For a very good explanation of ANOVA in Excel, see https://www.youtube.com/watch?v=tPGPV_XPw-o on the YouTube channel of *Dr. James Clark* from the Kings College London (Essential Life Science Tutorials).



	A	B	C
1	Sales Chocolate Dream	Sales Choc Champ	Sales Beethoven Bonbons
2	10357	10188	9536
3	10291	9998	9404
4	10009	10051	9354
5	10239	10062	9259
6	10882	10337	9633
7	10148	9845	9628
8	10281	10419	9663
9	10359	10157	9540
10	10132	10163	9872
11	10360	9967	9337
12	10357	10188	9536
13	10291	9998	9404
14	10009	10051	9354
15	10239	10062	9259
16	10882	10337	9633
17	10148	9845	9628
18	10281	10419	9663
19	10359	10157	9540
20	10132	10163	9872
21		9967	9337
22		10188	9536
23		9998	940
24		10051	935
25		10062	925
26		10337	
27		9845	
28		10419	
29		10157	
30		10163	

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Sales Chocolate Dream	19	195756	10302.95	54149.27
Sales Choc Champ	29	293594	10123.93	27688.14
Sales Beethoven Bonbons	24	228005	9500.21	31483.30

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	8087007	2	4043503.61	112.77	0.00000	3.13
Within Groups	2474071	69	35856.10			
Total	10561078	71				

Fig. 9.43 Analysis of variance in Excel

9.5.2 Kruskal–Wallis Test (*H* Test)

In the previous section we learned about ANOVA, a procedure that, when certain assumptions are satisfied, compares more than two groups with regard to the mean of a metric variable. What do we do if the requirements for ANOVA are not met because, say, the data are ordinally scaled or are neither normally distributed nor homoscedastic?

Let us return to the study we discussed in Sect. 9.4.2 about how much men and women are willing to pay for a bottle of wine. Say we want to know whether three different groups of consumers (A, B, and C) differ in their willingness-to-pay based on an ordinal scale (see Fig. 9.44). The first option would be to perform Mann–Whitney *U* tests for all possible combinations of two independent samples. But as the number of groups increases, the number of times that Mann–Whitney *U* test must be performed multiplies. Moreover, the likelihood that one of the tests is "accidentally" significant increases as well.

Zahlungs-bereitschaft	Client
€ [0-5]	C-Client
€ [0-5]	C-Client
€ [0-5]	B-Client
€ [0-5]	C-Client
€]5-10]	C-Client
€]5-10]	B-Client
€]5-10]	B-Client
€]10-15]	C-Client
€]10-15]	C-Client
€]10-15]	B-Client
€]10-15]	B-Client
€]10-15]	C-Client
€]10-15]	C-Client
€]10-15]	B-Client
€]10-15]	A-Client
€]15-20]	B-Client
€]15-20]	C-Client
€]15-20]	B-Client
€]15-20	B-Client
€]15-20]	A-Client
€]15-20]	B-Client
€]15-20]	B-Client
€]15-20]	A-Client
€]15-20]	A-Client
€]20-25]	A-Client

Step 1: Assign rank scores

Rank willing-ness to pay
2,5
2,5
2,5
2,5
6
6
6
11,5
11,5
11,5
11,5
11,5
11,5
11,5
11,5
20
20
20
20
20
20
20
20
20
25

Step 2: Place rank scores to the specific groups

A	B	C
11,5	2,5	2,5
20	6	2,5
20	6	2,5
20	11,5	6
25	11,5	11,5
	11,5	11,5
	4	11,5
	20	11,5
	20	20
	20	
	20	
96,5	149	79,5

Step 3: Calculation

	Client	N	Mean Rank
Willingness to pay	A-Client	5	19.30
	B-Client	11	13.55
	C-Client	9	8.83
	Total	25	

Test Statistics[a,b]

	Willingness to pay
Chi-Square	7.214
df	2
Asymp. Sig.	0.027
Exact Sig.	.022
Point Probability	.000

a Kruskal Wallis Test
b Grouping Variable: Client

Fig. 9.44 Kruskal–Wallis test (H test)

When testing group differences regarding the central tendency of an ordinal or metric variable for more than two groups (i.e. for k independent samples/groups), we can use the H test developed by Kruskal–Wallis (1952, 1953). This test should especially be used whenever the requirements for ANOVA are not satisfied (see Sect. 9.5.1). The basic idea of the H test has much in common with that of the U test (see Sect. 9.4.2). It tests whether the average rank score of at least one group significantly differs from the average rank score of another group. Follow these steps to perform an H test:

Step 1: Formulate the Hypotheses

$H_0 : E(\overline{R_i}) = E(\overline{R_j})$for all i $\neq$ j
$H_1 : E(\overline{R_i}) \neq E(\overline{R_j})$, for at least onei $\neq$ j.

Step 2: Determine the Significance Level α

Next, we must define α, i.e. the maximum permissible probability that we reject an H_0 that is true. The significance level α is usually set at 1, 5, or 10%, though 5% is the most common value, which is what we will use here ($\alpha = 0.05$).

Step 3: Draw a Sample
Researchers draw a sample with a size $n = 25$ observations. Figure 9.44 shows the amount three customer groups (A, B, and C) are willing to pay for a certain wine measured on a five-level ordinal scale (1 = €0–5; 2 = €5.01–10; 3 = €10.01–15; 4 = €15.01–20; 5 = €20.01–25).

Step 4: Check the Test Requirements

• Interval or ordinal scale of measurement of the test variable	✓
• Random sampling from two or more defined populations	✓
• Samples are independent and there is no overlap between group members	✓
• All groups should have the same shape distributions of the test variable	✓

Step 5: Determine the Empirical Test Value
The *H* test checks whether at least one of the customer groups differs significantly from the others in the amount they are willing to pay.

First we assign rank scores and factor in any tied ranks (see step 1 in Fig. 9.44). Then we assign the rank scores of the respondents to the given groups (see step 2 in Fig. 9.44) and arrange them by size. The answers of four respondents (three from customer group C and one from B) fall in the lowest willingness-to-pay level. Each receives a rank score of $((1 + 2 + 3 + 4)/4) = 2.5$. We proceed in similar fashion with the remaining payment levels. Next, we add up all the rank scores of each group, yielding a value of 96.5 for A, 149 for B, and 79.5 for C. Because the number of observations varies from group to group, we calculate the average rank scores by dividing the total of each group by its number of observations. This results in 19.3 (= 96.5/5) for A, 13.6 (= 149/11) for B, and 8.8 (= 79.5/9) for C.

If all the customer groups were willing to pay the same amount, the average rank scores would be approximately the same. As with the Mann–Whitney *U* test, we must check which mean rank can be expected if the groups do not differ. A rank sum of $1 + 2 + 3 + \ldots + n$ is assigned for all observations, which is calculated using

$$\frac{n \cdot (n+1)}{2}. \tag{9.55}$$

For the 25 respondents in the study, this produces a total rank sum of

$$\frac{n \cdot (n+1)}{2} = \frac{25 \cdot (25+1)}{2} = 325. \tag{9.56}$$

Dividing the rank sum by the total number of observations, results in the expected value for the average rank score under the assumption that the average rank scores of the groups do not differ. For our example the expected mean rank is:

$$E(\overline{R}) = \frac{1+2+3+.....+n}{n} = \frac{n \cdot (n+1)}{2 \cdot n} = \frac{(n+1)}{2} = \frac{(25+1)}{2} = 13 \tag{9.57}$$

A comparison of the actual average rank scores (19.3; 13.6; 8.8) with the expected value $E(\overline{R}) = 13$ indicates that the groups differ. But are the differences between the average rank scores of the three groups statistically significant?

To answer this question, we first need to add up all the squared deviations of actual average rank scores from their expected values:

$$\sum_{j=1}^{k} \left(\overline{R}_j - E(\overline{R})\right)^2 = \sum_{j=1}^{k} \left(\overline{R}_j - \frac{N+1}{2}\right)^2 \tag{9.58}$$

For our example, this produces:

$$(19.3 - 13)^2 + (13.6 - 13)^2 + (8.8 - 13)^2 = 57.7. \tag{9.59}$$

Is 57.7 as the total deviation a lot or a little? Put differently, is it significant or not? We can only answer this question if we can find a theoretical distribution that behaves like the squared difference between the average rank score and the expected value.

Step 6: Determine the Critical Test Value

To this end, we divide the sum of squared deviations by the variance of the rank scores. The result is the H value. We can show that the sum of deviations divided by the variance of the ranks with $k - 1$ degrees of freedom follows an asymptotic chi-square distribution (see Bortz et al. (2000, p. 222)):

$$H = \sum_{j=1}^{k} \frac{\left(\overline{R}_j - E(\overline{R})\right)^2}{\frac{\sigma^2}{n_j}} \sim \chi^2_{k-1} \tag{9.60}$$

The H value assumes an infinite population. If the population is finite—which is typically the case—we have to correct the result:

$$H^{\text{finite}} = \frac{N-1}{N} \sum_{j=1}^{k} \frac{\left(\overline{R}_j - E(\overline{R})\right)^2}{\frac{\sigma^2}{n_j}} = \frac{N-1}{N} \sum_{j=1}^{k} \frac{\left(\overline{R}_j - \frac{N+1}{2}\right)^2}{\frac{N^2-1}{12 \cdot n_j}} \sim \chi^2_{k-1} \tag{9.61}$$

When factoring in tied ranks, we need to use the following (admittedly complicated) formula to obtain the corrected H value[22]:

$$H^{\text{finite}}_{\text{corr}} = \frac{H}{C} = \frac{\frac{N-1}{N} \sum_{i=1}^{k} \left(\frac{\left(\overline{R}_j - \frac{N+1}{2}\right)^2}{\frac{N^2-1}{12 \cdot n_i}} \right)}{1 - \frac{\sum_{j=1}^{m} t_j^3 - t_j}{N^3 - N}} \sim \chi^2_{k-1} \tag{9.62}$$

[22] t_i represents the respective number of rank scores for the value i. In our example, we have four rank scores of 2.5 for value 1, three rank scores of six for value two, eight rank scores of 11.5 for value three, and nine rank scores of 20 for value four.

Applying this "crazy" formula to our example yields the following corrected H value:

$$H_{\text{corr}}^{\text{finite}} = \frac{\frac{25-1}{25}\left(\frac{(19.3-13)^2}{\frac{25^2-1}{12\cdot 5}} + \frac{(13.6-13)^2}{\frac{25^2-1}{12\cdot 11}} + \frac{(8.8-13)^2}{\frac{25^2-1}{12\cdot 9}}\right)}{1 - \frac{(4^3-4)+(3^3-3)+(8^3-8)+(9^3-9)}{25^3-25}} \approx 7.2 \sim \chi^2_{3-1} = \chi^2_2 \quad (9.63)$$

Step 7: Make a Test Decision

Whether an H value of 7.2 is significant depends on the chosen significance level. A probability error of $\alpha = 0.05$ results in a theoretical value for χ^2_2 of 5.991. All values above this threshold are statistically significant and would require us to reject H_0. Now the actual empirical H value (7.2) lies far above this threshold, which means we have to reject the null hypothesis at a significance level of $\alpha = 0.05$.

The corresponding p-value in Fig. 9.44 provides information about the exact probability of the error, provided that H_0 is true. For the asymptotic significance test described above, this produces a value of $p = 0.027$. The asymptotic H test is not biased if $k = 3$ groups have at least $n_j \geq 9$ observations each, $k = 4$ groups have at least $n_j \geq 5$ observations each, and $k = 5$ groups have at least $n_j \geq 3$ each. When there are more than five groups, the number of observations no longer plays a role (Bortz et al. 2000, p. 225). In our example with three customer groups, this condition is not satisfied for the A group, which is why we have to interpret the results of the exact H test. With a p-value of $p = 0.022$, we err in 2.2% of cases when assuming that at least one group differs from another. In sum, we can conclude that the average rank scores of the customer groups differ from each other. We must perform several U tests to determine which group differs from the other, even if the average rank scores already indicate the likely post-hoc group comparisons.

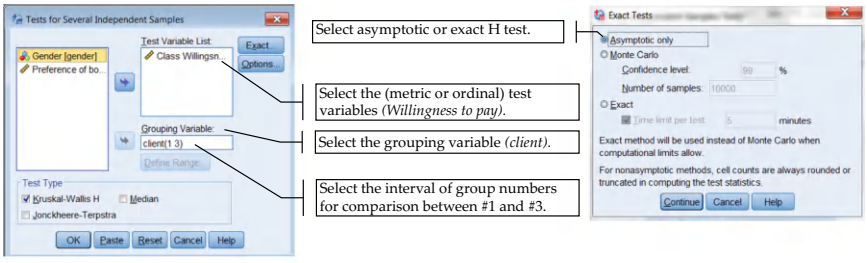
As with the U test, the H test assumes that the distributions within the groups have approximately the same shape. If this is not the case, the test will not be efficient and will tend to produce results that lead us to fail to reject H_0 despite the existence of group differences. The more the shapes of the group distributions deviate, the more a significant test result expresses differences in distribution shape as well as in central tendency. Hence, researchers interested not only in differences among average rank scores but also in the general differences between the groups—distribution shape, etc.—should not hesitate to use the H test. However, if the data are metric-scaled, the population follows a normal distribution and the variances are equal, ANOVA is the more efficient option. Here too, each technique can serve as a backup for the other when they lead to similar conclusions. When they do not, it is important to think about the reasons for this.

Now, once again, we will leave the formal world of mathematical calculations. No one today would calculate the H test by hand. Statistics software saves us from working out this complicated formula and provides a relatively simple tool for performing the Kruskal–Wallis test (H test) and interpreting its results.

9.5.2.1 Kruskal–Wallis *H* Test with SPSS

To perform the Mann–Whitney *H* test for the sample file *Chapter9_rank (Wine Bottle).sav* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9) in SPSS, select *Analyze → Nonparametric Tests → Legacy Dialogs → K Independent Samples…*. This opens the window shown on the left in Fig. 9.45. Next add the grouping variable (*client*) to the *Grouping Variable* and indicate the interval of group numbers for comparison between group #1 and group #3 (group #1: "very frequent client (A)", group #2: "frequent client (B)", and group #3: "infrequent client (C)") under *Define Range…* Enter the test variable (*Class willingness to pay*) in the *Test Variables List*. Under *Exact…* indicate whether SPSS should calculate an exact or an asymptomatic *H* test. For small samples, the results for both tests are calculated automatically.

The asymptotic test produces a *p*-value of $p = 0.027$. The asymptotic *H* test is not biased if $k = 3$ groups have at least $n_j \geq 9$ observations each, $k = 4$ groups have at least $n_j \geq 5$ observations each, and $k = 5$ groups have at least $n_j \geq 3$ each. When there are more than five groups, the number of observations no longer plays a role (Bortz et al. 2000, p. 225). In our example with three customer groups, this condition is not satisfied for the A group, which is why we have to interpret the results of the exact *H* test.



	Client	N	Mean Rank
Willingness to pay	A-Client	5	19.30
	B-Client	11	13.55
	C-Client	9	8.83
	Total	25	

Mean rank of the *Willingness to pay* for A-clients.
Mean rank of the *Willingness to pay* for B-clients.
Mean rank of the *Willingness to pay* for C-clients.

Test Statistics[a,b]

	Willingness to pay
Chi-Square	7.214
df	2
Asymp. Sig.	0.027
Exact Sig.	.022
Point Probability	.000

a Kruskal Wallis Test
b Grouping Variable: Client

Calculated χ^2-value.
Degrees of freedom of the test.
P-value of the asymptotic H test: we err in p=2.7% (≤0.05) of cases when assuming that at least one group of clients differs from the others.
P-value of the exact H test: we err in p=2.2% (≤0.05) of cases when assuming that at least one group of clients differs from the others.

Fig. 9.45 Kruskal–Wallis *H* test with SPSS

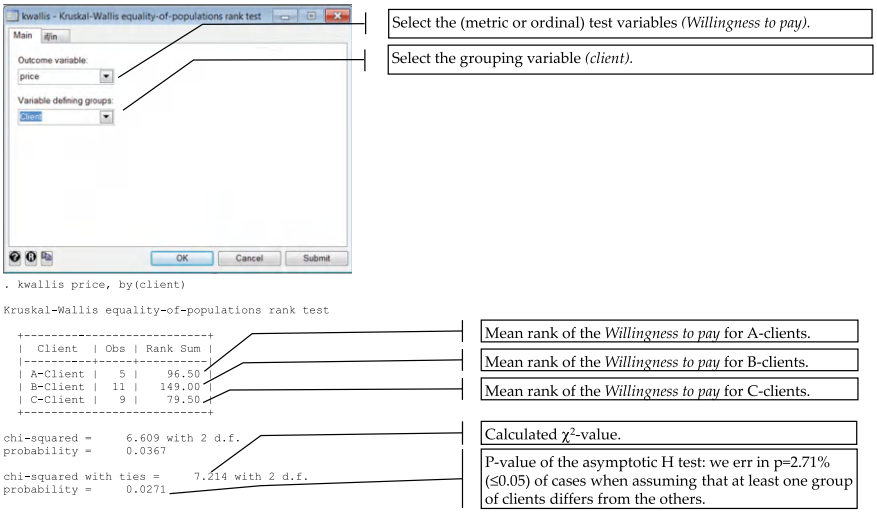


Fig. 9.46 Kruskal–Wallis *H* test with Stata

Using the results of the exact test, we err in $p = 2.2\%$ of cases when assuming that at least one group differs from another. We can conclude that the average rank scores of the customer groups differ from each other. We must perform a *U* test to determine which group differs from the other, even if the average rank scores already indicate the likely post-hoc group comparisons.

9.5.2.2 Kruskal–Wallis *H* Test with Stata

To perform the Kruskal–Wallis *H* test for the example file *Chapter9_rank (Wine Bottle).dta* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9) in Stata, select *Statistics → Nonparametric analysis → Tests of hypotheses → Kruskal–Wallis rank test*. This opens the window shown in Fig. 9.46. Next enter the grouping variable (*client*) in the *Variable defining groups* field. Under *Outcome Variable* enter the test variable (*price*).

According to the results, we err in $p = 2.271\%$ of cases when we assume that at least one group is willing to spend different amounts on the wine. In sum, we can conclude that the average rank scores of the customer groups differ from each other. We must perform a *U* test to determine which group differs from the other, even if the average rank scores already indicate the likely post-hoc group comparisons. Please note, however, that the sample size for A-clients in our example is too small to satisfy the minimum size requirement, which could bias the results.

9.5.2.3 Kruskal–Wallis *H* Test with R Commander

Select *Statistics → Nonparametric tests → Kruskal–Wallis test…* to perform the Kruskal–Wallis *H* test for the example file *Chapter9_*

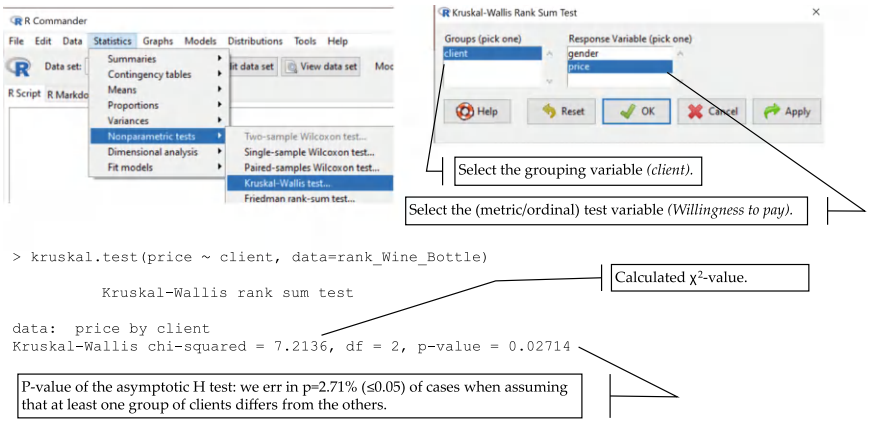


Fig. 9.47 Kruskal–Wallis *H* test with R Commander

rank (Wine Bottle).RData (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9). This opens the window shown on the right in Fig. 9.47. Next click on the grouping variable (*client*) in the *Groups (pick one)* field. Under *Response Variable (pick one)* enter the test variable (*price*).

According to the results, we are wrong in $p = 2.27\%$ of cases when we assume that at least one group is willing to spend different amounts of money on the wine. In total, we can conclude that the average rank scores of the customer groups differ from each other. We need to perform a U test to determine which group is different from the other. Please note, however, that the sample size for A-clients in our example is too small to satisfy the minimum size requirement, which could bias the results.

9.6 Other Tests

9.6.1 Chi-Square Test of Independence

The chi-square test of independence—or chi-square test for short—checks whether two nominal attributes are stochastically independent. That is to say, it tests the null hypothesis H_0 that no relationship exists between two categorical variables. The alternative hypothesis H_1, by contrast, says that a correlation exists with a certain probability error of α. Let us return to the *Titanic* survivors example from Sect. 4.2 as we consider this approach once again.

Say we want to test the frequent claim that most of the survivors were from first class and most of the victims were from third class. To start we need the information in the Titanic dataset, including the variables *gender* (child, male, female),

survival * class Crosstabulation

			class				
			Crew	First	Second	Third	Total
survival	Alive	Count	212	202	118	178	710
		Expected Count	285.5	104.8	91.9	227.7	710.0
		% within survival	29.9%	28.5%	16.6%	25.1%	100.0%
		% within class	24.0%	62.2%	41.4%	25.2%	32.3%
		% of Total	9.6%	9.2%	5.4%	8.1%	32.3%
		Residual	-73.5	97.2	26.1	-49.7	
		Standardized Residual	-4.3	9.5	2.7	-3.3	
		Adjusted Residual	-6.8	12.5	3.5	-4.9	
	Dead	Count	673	123	167	528	1491
		Expected Count	599.5	220.2	193.1	478.3	1491.0
		% within survival	45.1%	8.2%	11.2%	35.4%	100.0%
		% within class	76.0%	37.8%	58.6%	74.8%	67.7%
		% of Total	30.6%	5.6%	7.6%	24.0%	67.7%
		Residual	73.5	-97.2	-26.1	49.7	
		Standardized Residual	3.0	-6.5	-1.9	2.3	
		Adjusted Residual	6.8	-12.5	-3.5	4.9	
Total		Count	885	325	285	706	2201
		Expected Count	885.0	325.0	285.0	706.0	2201.0
		% within survival	40.2%	14.8%	12.9%	32.1%	100.0%
		% within class	100.0%	100.0%	100.0%	100.0%	100.0%
		% of Total	40.2%	14.8%	12.9%	32.1%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	187,793[a]	3	,000
Likelihood Ratio	178,414	3	,000
Linear-by-Linear Association	,000	1	,998
N of Valid Cases	2201		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 91,94.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,292	,000
	Cramer's V	,292	,000
	Contingency Coefficient	,280	,000
N of Valid Cases		2201	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Fig. 9.48 Nominal associations and chi-square test of independence

class (first, second, third, and crew), and *survival* (yes, no) for each passenger.[23] Consider the contingency table in Fig. 9.48, which categorizes survivors by class. Did all the passengers have the same chances of survival? To test the stochastic independence of two nominal variables, we proceed as follows:

Step 1: Formulate the Hypotheses

The hypotheses of the chi-square test are:

- H_0: There is no relationship between two categorical variables, i.e. the variables are statistically/stochastically independent.
- H_1: There is a relationship between two categorical variables, i.e. the variables are statistically/stochastically dependent.

Step 2: Determine the Significance Level α

Next, we must define α, i.e. the maximum permissible probability that we reject an H_0 that is true. The significance level α is usually set at 1, 5, or 10%, though 5% is the most common value, which is what we will use here ($\alpha = 0.05$).

[23] The data in *Chapter4_Titanic.sav* (SPSS), *Chapter4_Titanic.dta* (Stata), *Chapter4_Titanic.RData* (R), and *Chapter4_Titanic.xls* (Excel) contain figures on the number of persons on board and the number of victims (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9). The data are taken from the British Board of Trade Inquiry Report (1990), Report on the Loss of the Titanic (S.S.), Gloucester (reprint).

Step 3: Draw a Sample
The test requires a simple random sample.

Step 4: Check the Test Requirements
Make sure that the expected frequency in each cell is larger than five for at least 20% of the cells. If this condition is not satisfied, the Fisher–Yates test (1963) should be used instead of the chi-square test.

Step 5: Determine the Critical Test Value
Under the above assumptions, the test value has $v = (m - 1) \cdot (q - 1)$ degrees of freedom, which means that it approximates a chi-square distribution. The values for m and q correspond to the respective number of rows and columns in the contingency table. For the example in Fig. 9.48, the two rows and four columns produce $v = (4 - 1) \cdot (2 - 1) = 3$ degrees of freedom. With an α of 0.05, the critical value of a chi-square table in the column $(1 - \alpha) = 0.95$ and in the row for three degrees of freedom is $c_0 = \chi^2_{3;0.95} = 7.815$ (see chi-square table in Appendix B).

Step 6: Determine the Empirical Test Value
If the empirical value for the chi-square is greater than this critical value ($c_0 < \chi^2$), we have to reject H_0. Calculating the chi-square value by hand for our example we get:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(n_{ij} - n_{ij}^e)^2}{n_{ij}^e} = \frac{(212 - 285.5)^2}{285.5} + \frac{(673 - 599.5)^2}{599.5} + \cdots + \frac{(528 - 478.3)^2}{478.3} \approx 187.8 \tag{9.64}$$

The parameters n_{ij}^e and n_{ij} correspond to the expected and actual counts of the individual cells in the cross table.

Step 7: Make a Test Decision
The content of this step depends on whether we generate the calculations using a traditional chi-square table (see Appendix B) or using the results tables of professional statistical software. With the latter, all we need to check is whether the p-value—frequently referred to as asymptotic significance—is less than the value for α ($p \leq \alpha$). If it is, we have to reject H_0 and assume a statistical dependence with a probability error of p. When using a traditional chi-square table, by contrast, we must first determine the critical value c_0 for the given confidence level $(1 - \alpha)$ and degrees of freedom $v = (m - 1) \cdot (q - 1)$. At 187.8, the empirical value for the chi-square clearly exceeds the critical value $c_0 = 7.815$. Accordingly, we have to reject H_0.

For our example, the association is confirmed by the *chi-square test of independence* and the relatively high measure of association (see Fig. 9.48), but, as is

always the case with the chi-square test of independence, whether the association is the one supposed—in our example, a higher survival rate among first-class passengers than among third-class passengers and not the other way around—must be verified by comparing standardized residuals between actual and the expected frequencies. The positive values for the standardized residuals in Fig. 9.48 express an above-average (empirical) frequency to survive; negative values express a below-average (empirical) frequency. First-class passengers have a survival value of (+ 9.5), third-class passengers (− 3.3)—above-average and below-average rates, respectively.

Should the result of a chi-square test lead us to fail to reject H_0—i.e. if $c_0 > \chi^2$ or $p > \alpha$—we cannot prove that the size of the nominal measure of association (e.g. Cramer's V) differs significantly from null. In this event, we cannot prove a statistical association.

9.6.1.1 Chi-Square Test of Independence with SPSS

Open the dataset *Chapter9_Titanic.sav* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9). To use SPSS to generate a crosstab and calculate the chi-square test and the nominal measures of association, begin by opening the crosstab window. Select *Analyze* → *Descriptive Statistics* → *Crosstabs*… Now select the row and column variables whose association you want to analyse. For our example we must select *survival* as the row variable and *class* as the column variable. Next click on *cells*… to open a cell window. There you can select the desired contingency table calculations (see Fig. 9.49: The cell display). The association measure and the *chi-square test of independence* can be selected under *statistics*… (see Fig. 9.49: The Statistics display). Click *OK* to generate the tables in Fig. 9.48.[24] The asymptotic significance of $p = 0.000 \leq 0.05$ corroborates the association identified above. With a Cramer's V of 0.292, the association can be considered medium strong.

9.6.1.2 Chi-Square Test of Independence with Stata

To calculate the chi-square test of independence and the nominal associations with Stata, follow a similar approach. Open the dataset *Chapter9_Titanic.dta* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9). Select *Statistics* → *Summaries, tables, and tests* → *Tables* → *Two-way tables with measures of association* to open the window in Fig. 9.50. The rows, columns, and calculations must be selected for each variable. The left side displays the chi-square test of independence and the measures of association; the right side shows the cell statistics of the contingency table. Click on *OK* or *Submit* to perform the Stata calculation.[25] The results can now be interpreted as in the SPSS example.[26]

[24] For a very good explanation of how to calculate the chi-square test of Independence using SPSS, see https://www.youtube.com/watch?v=wfIfEWMJY3s on the YouTube channel of *ASK Brunel*.
[25] Syntax command: tabulate class survived, cchi2 cell chi2 clrchi2 column expected row V.
[26] For a very good explanation of how to calculate the chi-square test of independence using Stata, see: https://www.youtube.com/watch?v=GZIi9zAlzIA on the *StataCorp LLC* YouTube channel.

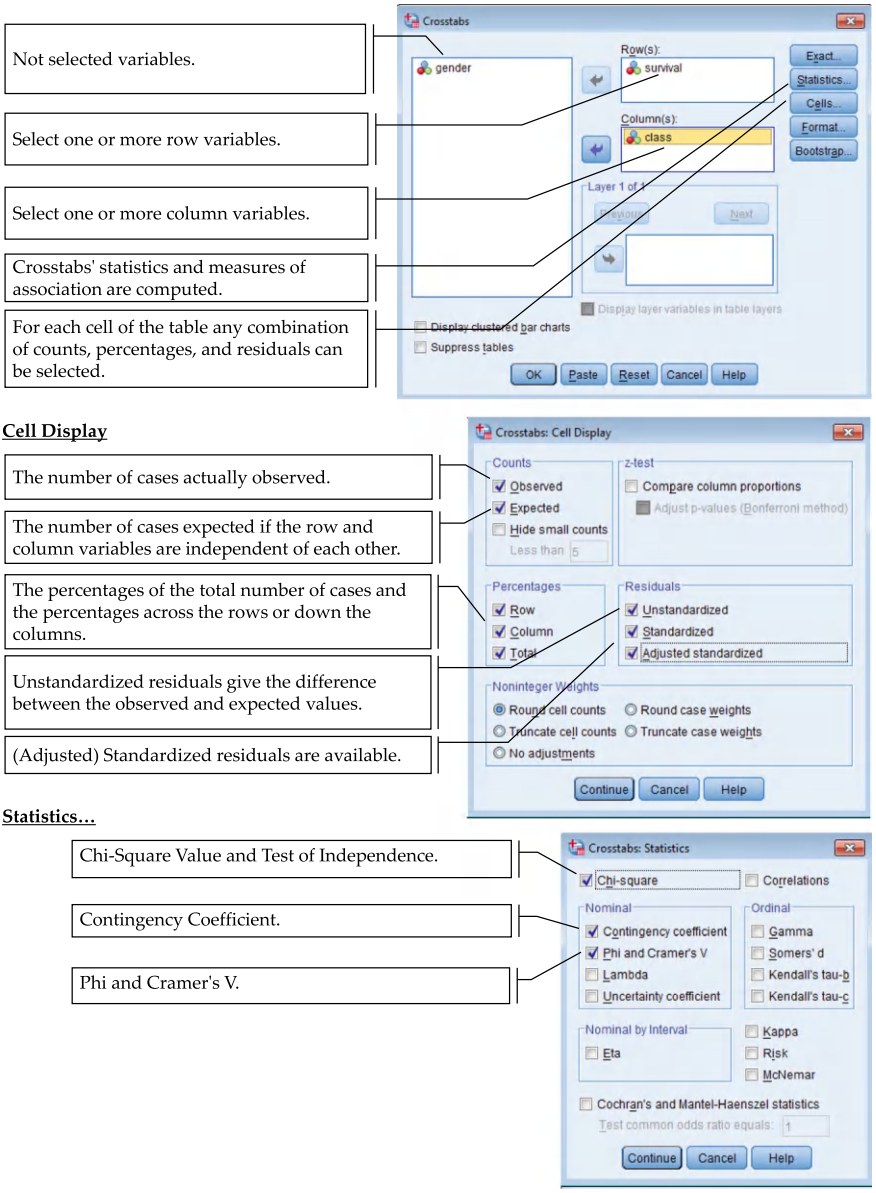


Fig. 9.49 Nominal associations and chi-square test of independence with SPSS

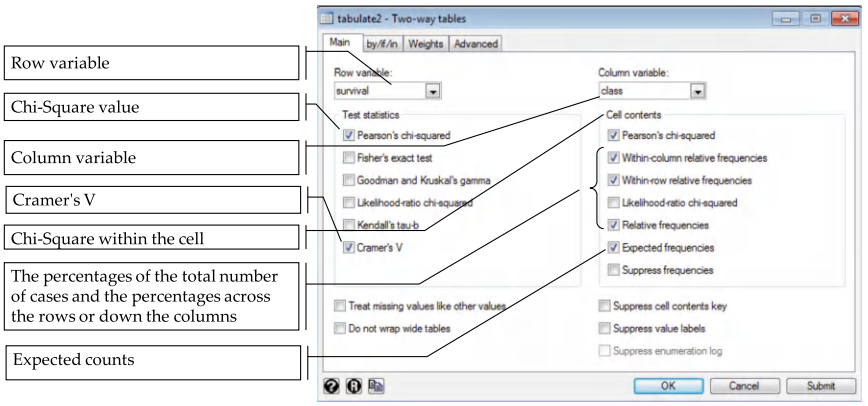


Fig. 9.50 Nominal associations and chi-square test of independence with Stata

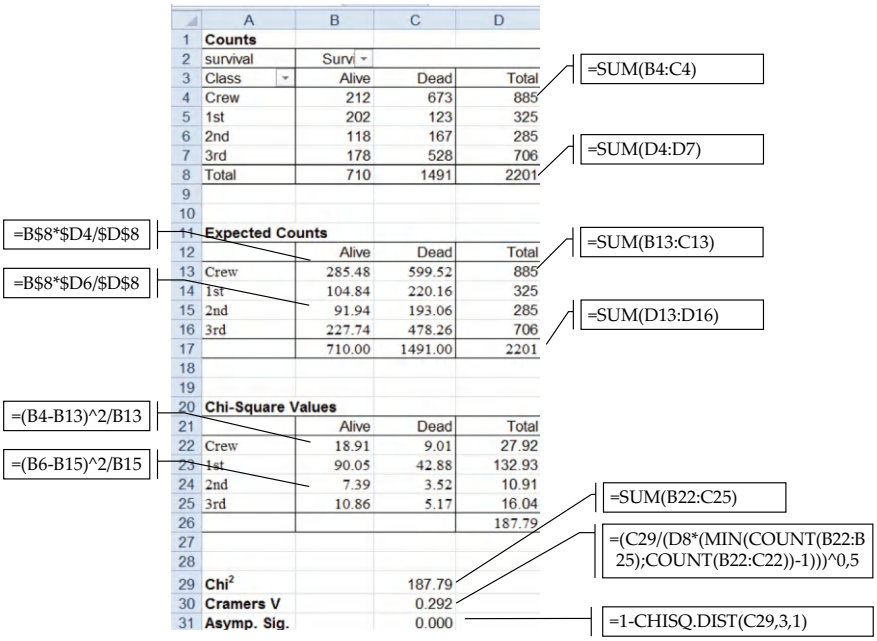
9.6.1.3 Chi-Square Test of Independence with R Commander

We have already learned how to calculate cross tabulations in Sect. 4.2.8. The chi-square test was also introduced at that time. Therefore, reference is made to that section, in particular to Fig. 4.13. With $\chi^2 = 187.8$ and three degrees of freedom, a highly significant association was found ($p < 0.01$).

9.6.1.4 Chi-Square Test of Independence with Excel

The calculation of crosstabs, related parameters (chi-square, phi, contingency coefficient, Cramer's V), and the chi-square test of independence with Excel is tricky compared with professional statistics packages. One of its main drawbacks is the shortage of pre-programmed functions for contingency tables.

Here is a brief sketch of how to perform these functions in Excel if needed. Open the dataset *Chapter9_Titanic.xls* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9). First compute the (conditional) actual counts for each cell as in Fig. 9.51. The pivot table function can be helpful when calculating these counts. Select the commands *Insert* and *Pivot Table* to open the *Create Pivot Table*. Then choose *Select a table or a range* and mark the location of the raw data. Click *OK* to store the pivot table in a *New Worksheet*. Drag the variables *survived* and *class* from the field list and drop them in the *Drop Row Fields Here* and *Drop Column Fields Here*. This generates a crosstab without conditional absolute counts. These can be added by dragging one of the variables from the field list to the field $\sum$ *values*. Then click on the variable in the field and select *Value Field Settings…* and the option *count* in the dialogue box. This generates a crosstab with the actual absolute counts. To update the crosstab when changes are made in the raw data, move the cursor over a cell and select *Options* and *Refresh* on the *PivotTable* tab. You can then calculate the expected counts using the given formula (row sum multiplied by the column sum divided by the total sum; see the second table in Fig. 9.51). In a new table we can calculate the individual chi-squares for each cell (see the third table in Fig. 9.51). The sum of



	A	B	C	D
1	**Counts**			
2	survival	Surv		
3	Class	Alive	Dead	Total
4	Crew	212	673	885
5	1st	202	123	325
6	2nd	118	167	285
7	3rd	178	528	706
8	Total	710	1491	2201
9				
10				
11	**Expected Counts**			
12		Alive	Dead	Total
13	Crew	285.48	599.52	885
14	1st	104.84	220.16	325
15	2nd	91.94	193.06	285
16	3rd	227.74	478.26	706
17		710.00	1491.00	2201
18				
19				
20	**Chi-Square Values**			
21		Alive	Dead	Total
22	Crew	18.91	9.01	27.92
23	1st	90.05	42.88	132.93
24	2nd	7.39	3.52	10.91
25	3rd	10.86	5.17	16.04
26				187.79
27				
28				
29	**Chi²**		187.79	
30	**Cramers V**		0.292	
31	**Asymp. Sig.**		0.000	

Fig. 9.51 Nominal associations and chi-square test of independence with Excel

these chi-squares equals the total chi-square value. From this, we can calculate the chi-square test of independence and Cramer's V. The formulas in Fig. 9.51 provide an example.[27]

9.6.2 Tests for Normal Distribution

The previous sections have made clear that many tests and approximation techniques assume normally distributed variables. If this assumption cannot be satisfied, then many of the parametric tests are either invalid or, provided that the samples are sufficiently large, at best asymptotically valid. For some examples, see the central limit theorem in Sect. 8.1.

The simplest way to check the normal distribution is to represent the distribution of the data in histograms. One must then decide by looking at the histograms whether the data approximate a normal distribution or not. Field (2005) criticizes the inherently subjective nature of this approach, which permits erroneous interpretations ("Well, it looks normal to me!"). As Fig. 9.52 shows, this is a justified

[27] For a very good explanation of how to calculate the chi-square test of independence using Excel, see https://www.youtube.com/watch?v=ODxEoDyF6RI on the YouTube channel of *Ken Blake*.

criticism. Two histograms describe the same distribution of a product's sales figures, but the one on the left uses five rectangles at defined intervals and the one on the right uses 25. Even the most honest reader will draw different conclusions from these graphs despite the fact that they represent the same data.

More objective approaches for checking the normal distribution are the Kolmogorov–Smirnov test (Kolmogorov, 1933; Smirnov, 1933), the Shapiro–Wilk test (Shapiro & Wilk, 1965), and the Shapiro–Francia test (Shapiro & Francia, 1972). We will now discuss them while testing the following hypotheses:

H_0: The variable being examined follows a normal distribution.
H_1: The variable being examined does not follow a normal distribution.

The Kolmogorov–Smirnov test has no distribution requirement of its own. It compares the maximum absolute difference between a cumulative distribution function of an empirical sample and a cumulative distribution function of a chosen theoretical distribution. In this way, the test checks the goodness of fit between empirical observations and a given theoretical distribution—in our case, the normal distribution.

Many statisticians do not regard the Kolmogorov–Smirnov test as the best way to check for normal distribution because its power as a statistical test is relatively low. The Kolmogorov–Smirnov test is more likely than other tests to reject the normal distribution assumption. Moreover, it does not adequately recognize deviations from the normal distribution in the outer margins. As result, statisticians tend to prefer the Shapiro–Wilk test or the Shapiro–Francia test. The former is especially suited for samples up to $n = 2000$ observations, whereas the latter works better for larger samples.

If the p-value of a test lies below the typical level of significance of $\alpha = 0.05$, we must reject the assumption of a normal distribution (H_0). Though the test can lead to different decisions in rare cases, we should always assume the absence of a normal distribution whenever one of the above tests produces significant results ($p \leq 0.05$).

9.6.2.1 Testing for Normal Distribution with SPSS

Let us use the sample file *Chapter9_chocolatepraline_colour_name_price.sav* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9) in SPSS to see whether the sales data for the two colours of praline packaging follow a normal distribution. Select *Analyze* → *Descriptive Statistics* → *Explore…* and enter the test variable (sales) under *Dependent List*. To check the groups separately, indicate a grouping variable (colour) under *Factor List* (see Fig. 9.53). Go to *Plots…* to create histograms or to perform the Kolmogorov–Smirnov test or the Shapiro–Wilk test.

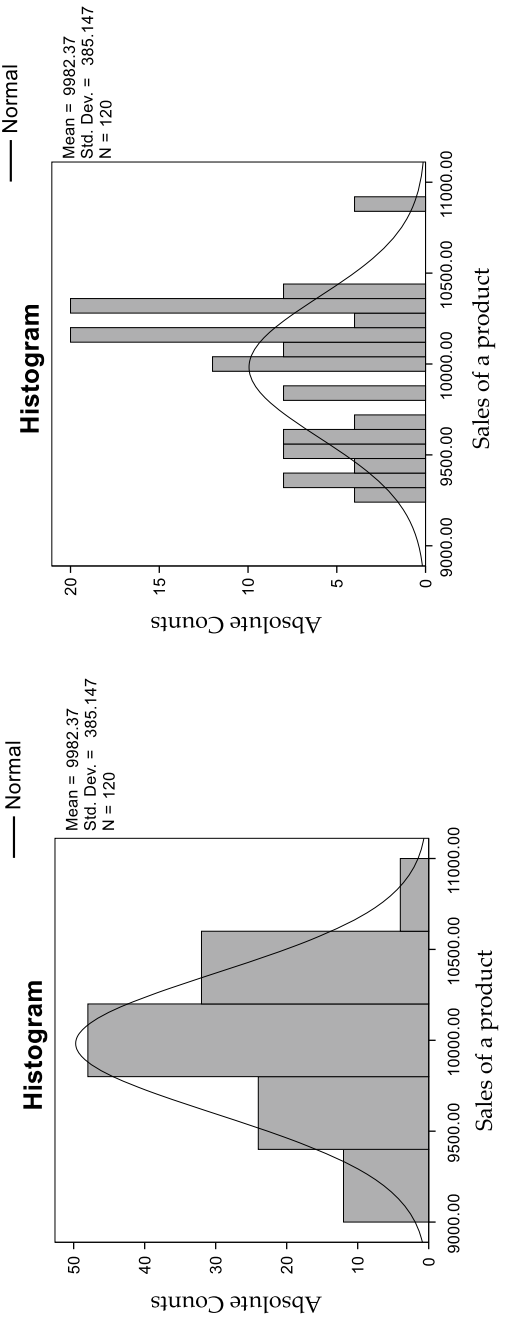
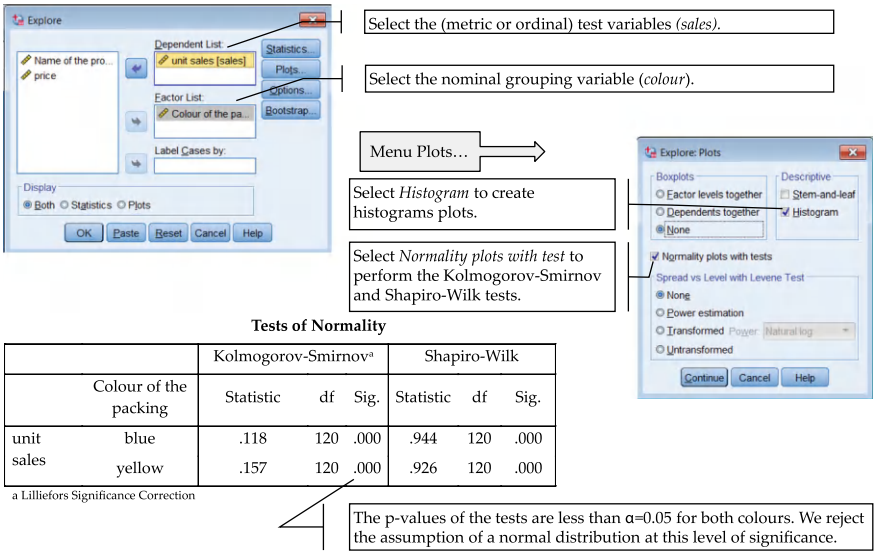


Fig. 9.52 Two histograms and their normal distribution curves



Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Colour of the packing	Statistic	df	Sig.	Statistic	df	Sig.
unit sales	blue	.118	120	.000	.944	120	.000
	yellow	.157	120	.000	.926	120	.000

a Lilliefors Significance Correction

The p-values of the tests are less than α=0.05 for both colours. We reject the assumption of a normal distribution at this level of significance.

Fig. 9.53 Testing for normal distribution with SPSS

Clicking *OK* produces the tables in Fig. 9.53. Because the *p*-values of the tests are less than $\alpha = 0.05$ for both colours, we reject the assumption of a normal distribution at this level of significance.[28]

9.6.2.2 Testing for Normal Distribution with Stata

We will now use our sample file *Chapter9_chocolatepraline_colour_name_price.dta* in Stata (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9) to see whether the sales data for the two colours of praline packaging follow a normal distribution. First create a graphic analysis with histograms of the blue (= 1) and yellow (= 2) packaging by entering the following commands in the Stata command line:

- *histogram sales if colour == 1, normal*
- *histogram sales if colour == 2, normal.*

For the Shapiro–Wilk and Shapiro–Francia tests, enter the following commands:

- swilk sales if colour == 1 or sfrancia sales if colour == 1
- swilk sales if colour == 2 or sfrancia sales if colour == 2.

[28] For a very good explanation of how to test for normal distribution using SPSS, see https://www.youtube.com/watch?v=dK-JNR3g_LU on the *Dragonfly Statistics* YouTube channel and https://www.youtube.com/watch?v=sQkB-AlJgPI on the *HowToStats.com* YouTube channel.

Performing the Kolmogorov–Smirnov tests is somewhat more complicated. First we must calculate the mean values and the standard deviations of the sales figures for each colour by entering *by colour, sort: summarize sales*. Then we must check for normal distribution using the command *ksmirnov testvar = normal((testvar − mean(testvar))/standarddev(testvar))*.

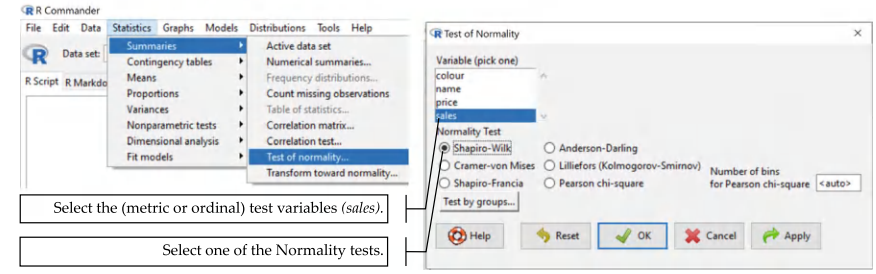
Next enter the commands for the blue and yellow packaging colours

- ksmirnov sales = normal((sales − 9982.367)/385.1468) if colour == 1
- ksmirnov sales = normal((sales − 9606.067)/490.2867) if colour == 2.

For both colours, the Shapiro–Wilk and Shapiro–Francia tests produce *p*-values that are smaller than $\alpha = 0.05$, which means we cannot assume a normal distribution.

9.6.2.3 Testing for Normal Distribution with R Commander

Let us use the sample file *Chapter9_chocolatepraline_colour_name_price.RData* in R (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9) to see whether *sales* follow a normal distribution. Click *Statistics* → *Summaries* → *Test of normality*… Select the variable (e.g. *sales*) and choose one of the tests, e.g. Shapiro–Wilk, Shapiro–Francia, or Kolmogorov–Smirnov. Clicking *OK* produces the tables in Fig. 9.54. The *p*-values of all three tests are highly significant ($p < 0.01$). We reject the assumption of a normal distribution at this level of significance.




```
> normalityTest(~sales, test="shapiro.test", data=Chocopraline_colour_name_price)
        Shapiro-Wilk normality test
data:  sales
W = 0.95273, p-value = 0.0000004654

> normalityTest(~sales, test="sf.test", data=Chocopraline_colour_name_price)
        Shapiro-Francia normality test
data:  sales

> normalityTest(~sales, test="lillie.test", data=Chocopraline_colour_name_price)
        Lilliefors (Kolmogorov-Smirnov) normality test
data:  sales
D = 0.13152, p-value = 7.779e-11
```

Fig. 9.54 Testing for normal distribution with R Commander

9.7 Chapter Exercises

Exercise 1
A defendant awaits the verdict in a courtroom. The judges must decide between pronouncing him *guilty* or *not guilty*.

(a) What are the two mistakes the judges could make in coming to their decision?
(b) Criminal trials are governed by the principle "innocent until proven guilty". Based on this principle, what would the statistical hypotheses have to be? What kind of statistical test is needed?
(c) Quickly sketch how you would proceed with such a statistical test.

Exercise 2
For the examples below, indicate the scale and suitable statistical test for the given variables. Are the samples dependent or independent? If no scale is given, the highest possible scale is presumed.

(a) Do men and women differ with regard to disposable income? The sample has a size of $n = 1000$.
(b) Twenty men and women rate how much they like a certain product on a scale from one to five. Are their preferences gender-specific?
(c) Twenty people rate how much they like a certain product on a scale from one to five both before and after hearing a presentation about its features. Did the presentation change respondent's opinion of the product?
(d) Do three programmes of study differ with regard to the proportion of male and female students enrolled?
(e) Do purchasers and non-purchasers of a product differ with regard to age? The mean age is assumed to approximate a normal distribution.
(f) Do two vehicle models differ with regard to having and not having special features?
(g) People are asked to rate a product on a scale from one to five before and after an ad campaign. Did the ad campaign change their opinion of the product?
(h) You are tasked with finding out whether parents' attitudes towards wood toys differ depending on whether they have sons or daughters. For the time being, you ignore parents with both sons and daughters. You measure their attitudes by asking them to rate wooden toys on a five-point scale (1 = "strongly dislike" to 5 = "strongly like").
(i) Which approach should you use if you include parents with sons and daughters in the survey?

Exercise 3
1000 purchasers of a new car receive the questionnaire in Fig. 9.55. For the examples below, indicate the scale and a suitable statistical test for the given variables. Are the samples dependent or independent?

(a) Do private car owners and company car owners differ with regard to the postal code they are living in?
(b) Do private car owners and company car owners differ with regard to initial impression of build quality?
(c) Does the mean number of years respondents drove their previous vehicle deviate from five?
(d) Do private car owners and company car owners differ with regard to the number of previously owned cars of the same brand?
(e) Does the overall satisfaction of customers with their new car differ from their satisfaction with the vehicle's interior?

Exercise 4
Each day, more than 10,000 customers shop at a business. A sample with a size of $n = 100$ from location A has an average customer stay of 45 min. The actual duration has a normal distribution with an unknown arithmetic mean μ and a variance of $S^2 = 15^2$ min^2. According to management, the mean stay is 55 min. Test whether a sample can rebut this claim with a probability of error of $\alpha = 0.05$. Explain your answer.

Questionnaire

1. Indicate the postal code of your city: _____

2. What is your initial impression of the new car?

	very satisfied	satisfied	dissatisfied
a) Overall	☐	☐	☐
b) Exterior	☐	☐	☐
c) Interior	☐	☐	☐
d) Build quality	☐	☐	☐

3. I drove my previous car for ____ years.

4. My car is a: ☐ private car ☐ company car

5. How many cars of the same brand have you previously owned? _____

Fig. 9.55 Questionnaire for owners of a particular car

Exercise 5
The management asks you to check whether products before shipping meet the minimum length stipulated by the customer (150 cm) with a 20% probability of error. After adjusting the machine anew, you select a sample of 120 metal strips, and find an average length of 150.1 cm with an empirical variance of $S^2_{\text{emp}} = 3^2\,\text{cm}^2$. The length of the metal strips follows a normal distribution for the total population:

(a) Using the example, formulate the statistical hypotheses so that the α mistake describes the *worst mistake.*
(b) Check whether under these conditions the assumption of a minimum length of 150 cm can be retained.
(c) Find the *p*-value from (b)!

Exercise 6
The management asks you to say with 99% certainty that the change from yellow to blue packaging will lead to better sales. The yellow packaging led to weekly sales of 10,000 packages on average. The sample of 100 stores from the 5500 Tasty Markets showed weekly sales of 10,050 packages for the blue packaging with an empirical standard deviation of $S_{\text{emp}} = 360$ packages.

(a) Use the example to formulate the test hypotheses so that the α-error is the *worst error*.
(b) Identify the corresponding β error.
(c) Test whether the demand of company management is fulfilled under these circumstances.
(d) How large must the sample size be—ceteris paribus—so that the lower limit of a one-sided confidence interface ($\alpha = 5\%$) is at least 9950 packages?

Exercise 7
You are given the task of carrying out a competition analysis for baby food. You ask 120 parents to rate on a scale from one to five how natural they find products from two manufacturers, *Unhappy Family* and *Potbelly*.

(a) Which statistical tests can be used?
(b) As part of your task, you must determine whether a product should be called *Baby Carotte* or *Baby Carrot*. You turn to the same 120 parents. You ask 60 to rate *Baby Carotte* using a 5-point scale. The other 60 must rate *Baby Carrot* using an analogue scale. Which statistical test would you use?

Exercise 8
You are investigating the effect of three different advertising strategies (circulars, local newspaper ads, and billboards) on customer preference for a certain product. You measure customer preference with a 10-point scale (1 = "I plan to buy it right away" to 10 = "I will definitely not buy it"). The results are shown in Fig. 9.56.

Ranks

	Advertising Strategy	N	Mean Rank
Perception of product quality	Circular	15	38.00
	Newspaper ads	15	10.10
	Billboards	15	20.90
	Total	45	

Test Statistics[a,b]

	Perception of product quality
Chi-Square	35.057
df	2
Asymp. Sig.	.000

a. Kruskal Wallis Test

b. Grouping Variable: Advertising Strategy

Fig. 9.56 Effect of three advertising strategies

Ranks

	Advertising Strategy	N	Mean Rank	Sum of Ranks
Perception of product quality	Newspaper ads	10	13.65	136.50
	Billboards	15	12.57	188.50
	Total	25		

Test Statistics[a]

	Perception of product quality
Mann-Whitney U	68.500
Wilcoxon W	188.500
Z	-.380
Asymp. Sig. (2-tailed)	.704
Exact Sig. [2*(1-tailed Sig.)]	.723[b]
Exact Sig. (2-tailed)	.709
Exact Sig. (1-tailed)	.379
Point Probability	.012

a. Grouping Variable: Advertising Strategy

b. Not corrected for ties.

Fig. 9.57 Effect of two advertising strategies

(a) Interpret the results. Which advertising strategy would you choose?
(b) Briefly explain how the mean rank in Fig. 9.56 was determined.
(c) Now you examine in a different study the effects of newspaper ads and billboards on perception of product quality. The results are shown in Fig. 9.57. Interpret the results. How large is the one-sided asymptotic significance?

Exercise 9

Figure 9.58 shows the results of a study about the annual revenue from different household types in a supermarket chain (see dataset *Chapter9_ttest_household_type_and_revenues*; supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9). What was studied and how are the results to be interpreted?

Group Statistics

	Household typ	N	Mean	Std. Deviation	Std. Error Mean
Annual revenues	Type 1	178	772.7131	15.87738	1.19006
	Type 2	184	778.2458	29.90867	2.20490

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Annual revenues	Equal variances assumed	81.231	.000	-2.19	360	.029	-5.53277	2.52900	-10.50623	-.55930
	Equal variances not assumed			-2.21	280.536	.028	-5.53277	2.50556	-10.46485	-.60069

Fig. 9.58 Results of a market research study

Exercise 10

For the buyers of two vehicle brands, you increase the strength of the motors in classified form. This results in the following contingency table:

(a) Indicate the expected frequencies in the case of statistical independence [in brackets under the actual number of observations].

	<51 horsepower	51–75 horsepower	76–100 horsepower	>101 horsepower	Marginal frequency (x)
Vehicle brand 1	100 (____)	55 (____)	5 (____)	0 (____)	
Vehicle brand 2	0 (____)	30 (____)	5 (____)	5 (____)	
Marginal frequency (y)					

(b) What per cent of the purchasers of brand 1 bought a vehicle with between 76 and 100 horsepower?

(c) What per cent of purchased vehicles with a motor output of between 51 and 75 horsepower belong to vehicle brand 2?

(d) Briefly sketch how you would proceed with a chi-square test for statistical independence.

(e) Calculate the chi-square value so that the chi-square test is not distorted. On the basis of this chi-square value, can we conclude with a type 1 error of $\alpha = 1\%$ that there is a statistical dependence between purchased vehicle brand and motor output?

(f) Say you increased the sample in the above example so that no rows or columns have to be aggregated. The chi-square value is now only 7.128. On the basis of this chi-square value, can we conclude with an error probability of $\alpha = 1\%$ that there is a statistical dependence between purchased vehicle brand and motor performance?

Exercise 11
In an experiment on the effectiveness of music on the willingness of customers to spend money in a supermarket, researchers selected 400 customers at random. Some of the customers shopped on days when the supermarket did not play background music. The other group of customers shopped on days in which the supermarket broadcast music and advertising. Each customer was put into one of three groups—high, medium, and low willingness-to-spend—based on the total sum of purchases.

		High expenditures ($y = 1$)	Middle expenditure ($y = 2$)	Low expenditures ($y = 3$)	Sum (X)
With music ($x = 1$)	Number (exp. frequencies)	130 (89.25)	30 (26.25)	50 (94.50)	210
Without music ($x = 2$)	Number (exp. frequency)	40 (80.75)	20 (23.75)	130 (85.50)	190
Sum (Y)	Number	170	50	180	400

(a) Determine the empirical chi-square value so that the chi-square test is not biased.
(b) Can we conclude from the data with an error probability of $\alpha = 5\%$ that there is a connection between music and the total purchase sum?

Exercise 12
After speaking with grocery store customers, you identify household size and the number of purchased bananas and place them in the following contingency table.

	1 person ($y = 1$)	2 persons ($y = 2$)	≥ 3 persons ($y = 3$)	Sum (x)
0 bananas ($x = 1$)	40 (40)	0 (4)	40 (36)	80
1 banana ($x = 2$)	103 (102.5)	15 (10.25)	87 (92.25)	205
2 bananas ($x = 3$)	5 (4)	0 (0.4)	3 (3.6)	8
≥ 3 bananas ($x = 4$)	2 (3.5)	0 (0.35)	5 (3.15)	7
Sum (y)	150	15	135	300

(a) Identify the empirical chi-square value so as not to distort the chi-square test.
(b) Can we infer from the data a correlation between household size and banana purchases?

Exercise 13
A study records the overall preference for a product ("How would you rate this product?") and attitudes towards the price of a product ("It is a fair price"). The result is the crosstab in Fig. 9.59.

(a) What per cent of respondents who rated the product as poor, strongly disagree with the statement "It is a fair price"?

How would you rate this product? * It is a fair price Crosstabulation

Count

		It is a fair price				
		strongly agree	agree	disagree	strongly disagree	Total
How would you rate this product?	excellent	25	22	14	9	70
	good	54	30	44	28	156
	fair	71	45	50	28	194
	poor	25	20	23	12	80
Total		175	117	131	77	500

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	6.214[a]	9	.718
Likelihood Ratio	6.176	9	.722
Linear-by-Linear Association	.181	1	.670
N of Valid Cases	500		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 10.78.

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate[b]T	Approximate Significance
Nominal by Nomina	Phi	.111			.718
	Cramer's V	.064			.718
	Contingency Coefficient	.111			.718
Ordinal by Ordinal	Kendall's tau-b	.012	.037	.326	.745
	Kendall's tau-c	.012	.035	.326	.745
	Spearman Correlation	.014	.044	.318	.750[c]
Interval by Interval	Pearson's R	.019	.044	.426	.671[c]
N of Valid Cases		500			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

Fig. 9.59 Product preference

(b) Is the relationship significant? To answer this question, assess the coefficients Phi, Cramer's V, Contingency Coefficient, and Kendall's Tau.

(c) After a product presentation, the same subjects are asked about their attitude to the price of the product ("After product presentation: It is a fair price") on a scale from one ("strongly agree") to four ("strongly disagree"). Their responses are then subject to a statistical test. Interpret the results from Fig. 9.60.

(d) Their responses are now tested again but this time using a different procedure. Interpret the results in Fig. 9.61. Is there a nonparametric way to analyse the change in preferences here?

Ranks

		N	Mean Rank	Sum of Ranks
After product presentation: It is a fair price - It is a fair price	Negative Ranks	159[a]	83.09	13212.00
	Positive Ranks	6[b]	80.50	483.00
	Ties	335[c]		
	Total	500		

a. After product presentation: It is a fair price < It is a fair price

b. After product presentation: It is a fair price > It is a fair price

c. After product presentation: It is a fair price = It is a fair price

Test Statistics[a]

	After product presentation: It is a fair price - It is a fair price
Z	-11.770[b]
Asymp. Sig. (2-tailed)	.000

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

Fig. 9.60 Price preference 1

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	After product presentation: It is a fair price	1.9040	500	.99638	.04456
	It is a fair price	2.2200	500	1.08718	.04862

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	After product presentation: It is a fair price - It is a fair price	-.31600	.51055	.02283	-.36086	-.27114	-13.840	499	.000

Fig. 9.61 Price preference 2

Exercise 14

What was tested in the procedure represented in Fig. 9.62? Interpret the results.

Exercise 15

You are a sales manager for a consumer product and want to maximize the number of individuals in your target customer group, so you decide to run a television commercial. For four German TV channels and three time slots (5:00–5:59 p.m.; 6:00–6:59 p.m.; 7:00–7:59 p.m.), you identify the number of households reached (n) and the average number of individuals in your target group reached. The results can be found in the files *Chapter9_advertisingcontacts.sav*, *Chapter9_advertisingcontacts.RData*, or *Chapter9_advertisingcontacts.dta* (see supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9).

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Age	500	52.8780	10.07463	.45055

One-Sample Test

Test Value = 54

	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Age	-2.490	499	.013	-1.12200	-2.0072	-.2368

Fig. 9.62 One-sample *t*-test

(a) Have the requirements for multiple-factor ANOVA been satisfied? If they have not been satisfied, what should be done?
(b) Which significant subgroups arise for the variable *television channel*?
(c) Do the time slots differ from each other? Explain.
(d) Using the marginal frequencies, explain how the interaction effect comes about.
(e) You have received offers from the television channel ARD for the time slot 7:00–7:59 p.m. and from ZDF for the time slot 6:00–6:59 p.m. ZDF's offer is somewhat more affordable than ARD's. What problem is hard to decide? What procedure should be used to make a rational decision? Explain.

Exercise 16

You are a manager of an international chain of hardware stores with locations in Germany, France, Spain, and the United Kingdom. The chain has gone by a different name in each country so far. Now you want to define a uniform corporate identity. Therefore, you decide to carry out a market study investigating how the colour of the company logo affects the average number of visitors per hour (see files *Chapter9_hardware_store.sav*, *Chapter9_hardware_store.RData*, *Chapter9_hardware_store.xls*, or *Chapter9_hardware_store.dta*; supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9).

(a) Which colour would you select?
(b) Which countries have different average numbers of visitors per hour? Base your answer on the results of Scheffé's method.
(c) Using marginal frequencies, explain why the interaction effect occurs.
(d) Are the requirements satisfied for ANOVA?

Exercise 17

Shnutella, an international manufacturer of the hazelnut chocolate spread, is looking for a new advertising ambassador and a new product slogan. Because the pan-European football tournament (UEFA Euro) is coming up, the company decides to make a pro player its advertising ambassador. Under consideration are the goalie

Oliver Kahn, the forward Luis Figo, and the midfielder Zinedine Zidane. Also being discussed are three new product slogans proposed by the ad agency *Nobrain & Partners*: (1) *Chocolate with Discipline*, (2) *Chocolate, Ohlala*, and (3) *The Chocolate of the Champs*. To help its decision-making, the manufacturer commissions a study to investigate the influence of each advertising ambassador with different combinations of slogans on sales in 500 test markets (see the files *Chapter9_Shnutella.sav*, *Chapter9_Shnutella.RData* or *Chapter9_Shnutella.dta*; supplementary material at DOI: https://doi.org/10.1007/978-3-031-78070-7_9).

(a) Which advertising ambassador would you select based on the results of single-factor ANOVA using Scheffé's method?
(b) Which slogan would you select based on the results of single-factor ANOVA using Scheffé's method?
(c) The company's management asks you to present recommendations for the new advertising ambassador and the new slogan. Interpret the results of multiple-factor ANOVA. Discuss the requirements of ANOVA, the significance of each variable for sales, and possible contradictions in your answers above. Also describe which questions cannot be clearly answered by ANOVA.

9.8 Exercise Solutions

Solution 1

(a) 1st mistake: The defendant is found guilty though he is innocent. 2nd mistake: The defendant is found innocent though he is guilty.
(b) Based on the principle of *innocent until proven guilty* we have to ensure that the probability of a false verdict is as low as possible. We have to prove that the defendant is guilty. This means:
H_0: Defendant is innocent.
H_1: Defendant is guilty.
Type 1 error: Defendant is innocent yet found guilty.
Type 2 error: Defendant is guilty yet found innocent.
The *worst mistake* is the *Type 1 error*: $\alpha = P(H_1|H_0 \text{ correct})$. Statistical tests try to keep the probability of this mistake as low as possible.
(c)
1. List hypotheses.
2. Determine the distribution of the test statistic using the null hypothesis.
3. Determine the regions of acceptance and rejection.
4. Determine the value of the test statistic in the sample.
5. Decision: If the value from the sample is in the region of rejection, the null hypothesis is rejected. Otherwise the null hypothesis is not rejected. The *worse mistake* is the first type: $\alpha = P(H_1|H_0 \text{ correct})$.

Solution 2

(a) Gender: nominal (dichotomous grouping variable); income: metric (test variable); because the sample is sufficiently large, the mean approximates a normal distribution; *t*-test for two independent samples.
(b) Gender: nominal (dichotomous grouping variable); product rating: ordinal (test variable); two independent samples; Mann–Whitney *U* test.
(c) Rating of product beforehand: ordinal; rating of product afterwards: ordinal; dependent sample; sample does not allow us to assume an approximate normal distribution of the means: Wilcoxon test.
(d) Gender: nominal; programme of study: nominal; chi-square test.
(e) Purchasers and non-purchasers: nominal (dichotomous grouping variable); age: metric (test variable); independent sample; because of the approximate normal distribution of the mean, the *t*-test can be used for two independent samples.
(f) Two vehicle models: nominal; feature: nominal; chi-square test.
(g) Rating before ad campaign: ordinal; rating after ad campaign: ordinal; dependent sample; Wilcoxon test.
(h) Parents (only sons; only daughters): nominal (dichotomous grouping variable); rating of wood toys: ordinal (test variable); two independent samples; Mann–Whitney *U* test.
(i) Parents: nominal (grouping variable); rating of wood toys: ordinal (test variable); three independent samples; Kruskal–Wallis *H* test.

Solution 3

(a) Private car owners and company car owners: nominal; postal code: nominal; chi-square test.
(b) Private care owners and company car owners: nominal (dichotomous grouping variable); impression of build quality: ordinal (test variable); two independent sample; Mann–Whitney *U* test.
(c) Mean number of years previous vehicle was driven: metric; one-sample *t*-test.
(d) Private car owners and company car owners: nominal (dichotomous grouping variable); number of vehicles of the same brand: metric (test variable); independent sample; because of the approximate normal distribution of means ($n = 1000$), the *t*-test can be used for two independent samples.
(e) Satisfaction for overall impression: ordinal; dependent sample; Wilcoxon test.

Solution 4
We have to test whether the mean stay is 55 min. H_0: $\mu = 55$; H_1: $\mu \neq 55$. From the empirical data we get: $\bar{x} = 45$. If the variance of the total population is known and the correction term is not needed, we get the following region of acceptance for H_0:

$$\left[\mu_{\text{lower}} - z_{1-\frac{\alpha}{2}}\sigma_{\bar{x}};\ \mu_{\text{upper}} + z_{1-\frac{\alpha}{2}}\sigma_{\bar{x}}\right] = [55 - 1.96 \cdot 1.5;\ 55 + 1.96 \cdot 1.5]$$
$$= [52.06;\ 57.94].$$

H_0 is not retained since the empirical arithmetic mean $\bar{x} = 45$ does not lie in the given interval. We can claim that the mean stay is not equal to 55 min with a probability of error of 5%.

Solution 5

(a) Management wants to demonstrate that the length of the metal strips is at least 150 cm on average: H_0: $\mu < 150$; H_1: $\mu \geq 150$. The *worst mistake* from the management's perspective would be the decision to send a delivery that only appears to satisfy the minimum requirement.

(b) If the variance of the total population is unknown and a correction term is not necessary ($n > 30$), a sample size of 120 results in the following interval for non-rejection of H_0:

$$[-\infty; \mu_{\text{upper}} + z_{1-\alpha}\sigma_{\bar{x}}[= \left[-\infty; \mu_{\text{upper}} + z_{1-\alpha}\frac{S_{\text{emp}}}{\sqrt{n-1}}\right[$$
$$= [-\infty; 150 + 0.842 \cdot 0.275[= [-\infty; 150.23[.$$

H_0 is retained since the empirical arithmetic mean $\bar{x} = 150.1$ lies in the interval for non-rejection of H_0. Management's assumption that the minimum length of the metal strips is on average larger than 150 cm cannot be demonstrated.

(c) The difference between the empirical mean and the hypothetical value is $150.1 - 150 = 0.1$

$$150.1 = \mu_{\text{upper}} + z_{1-\alpha}\frac{S}{\sqrt{n-1}} = 150 + z_{1-\alpha}\frac{3}{\sqrt{120-1}}$$
$$\Rightarrow \frac{0.1}{3}\sqrt{119} = z_{1-\alpha} \Rightarrow 0.36 = z_{1-\alpha} \Rightarrow \Phi(0.36) = 1 - \alpha \Rightarrow \alpha = 1 - \Phi(0.36)$$

Since the value for $\Phi(0.36)$ is not taken directly from a one-sided table of the normal distribution, it can be converted into $\Phi(0.36) = 1 - \Phi(0.64)$, which coincidentally has a value of 0.36 as well. The probability of error is thus:

$$\Rightarrow p = 1 - (1 - \Phi(0.64)) = 1 - (1 - 0.36) = 0.36$$

Solution 6

(a) Management will only approve the costs of changing to blue packaging if doing so does not lead to lower sales. We thus need to test the hypotheses: $H_0 : \mu_{\text{blue}} \leq 10{,}000$ versus $H_1 : \mu_{\text{blue}} > 10{,}000$. Management wants to show that the blue packaging achieves better results. From this perspective, the *worst error* is to decide for blue packaging if it does not lead to better sales.

(b) I do not reject the hypothesis that sales with blue packaging decline or stay the same, though in truth blue packaging leads to better sales: $P(\text{decision for } H_0 | H_1 \text{ correct}) = \beta$ error.

(c) If the variance of the total population is unknown and a correction term is not necessary, a sample size of ($n > 30$) results in the following interval of non-rejection of H_0:

$$\left[-\infty;\, \mu_{\text{upper}} + z_{1-\alpha}\sigma_{\bar{x}}\right] = \left[-\infty;\, \mu_{\text{upper}} + z_{1-\alpha}\frac{S_{\text{emp}}}{\sqrt{n-1}}\right]$$
$$= \left[-\infty;\, 10{,}000 + z_{99\%}\frac{360}{\sqrt{99}}\right] = [-\infty;\, 10{,}084.17]$$

H_0 is retained since the empirical mean $\bar{x} = 10{,}050$ lies in the given interval. The assumption of management that blue packaging does not lead to lower sales cannot be confirmed on the 1% significance level.

(d) For the sake of approximation, a normal distribution can be assumed for the calculation of a confidence interval if the calculated sample size exceeds the value of 30. The lower limit of the one-sided confidence interval with unknown variance of the total population and $(n-1) > 30$ as well as $n/N > 0.05$ is given by $\mu_{\text{lower}} = \bar{x} - z_{1-\alpha} \cdot \hat{\sigma}_{\bar{x}} \Rightarrow \mu_{\text{lower}} = \bar{x} - z_{1-\alpha}\frac{S_{\text{emp}}}{\sqrt{n-1}}$.

If the values are placed in the equation, we get:

$$9950 = 10{,}050 - 1.645\frac{360}{\sqrt{n-1}} \Rightarrow \frac{-100}{-1.645}$$
$$= \frac{360}{\sqrt{n-1}} \Rightarrow \sqrt{n-1}^2 = 5.92^2 \Rightarrow n = 36.06 \approx 37$$

Rounding can lead to minor deviations. Since the sample size is far over 30, the use of the normal distribution is permissible.

Solution 7

(a) This situation involves a paired/dependent sample. Due to the ordinal scale, the Wilcoxon test can be used. Provided the sampling distribution of the sample mean can be approximated by a normal distribution—with $n = 2 \cdot 60 = 120$ this is a realistic assumption—a t-test for paired samples (dependent) is well-suited for the sample.

(b) This situation involves two independent samples. Provided the mean approximates a normal distribution and there is a metric scale—again, a realistic assumption given the size of n—a t-test for two independent samples is well-suited for the sample. The Mann–Whitney U test is another option.

Solution 8

(a) Due to the lack of information about the distribution and the small sample, a Kruskal–Wallis test (H-test) was used. The asymptotic significance is significantly smaller than 0.05, so that at least one advertising strategy differs from the others. The H test gives no information about which of the three advertising strategies significantly differ from the others in their effects. In this example, low average values signify a high customer preference (1 = "I plan to buy it right away"). Local newspapers have the lowest average rank. A Mann–Whitney U test could determine if they differ significantly from billboards.

(b) First, sort the ordinal-scaled variable by size and assign rankings as you would with rank correlation coefficients (e.g. Spearman correlation). The average rank is then taken for each group.

(c) The Mann–Whitney U test is not significant. The exact significance is 0.723, far above the typical threshold of 0.05. Customer perception of product quality does not differ between newspaper ads and billboards. The one-sided asymptotic significance is $0.704/2 = 0.352$, which is above the threshold of significance (0.05).

Solution 9

The study investigated whether the average annual revenues of two household types differed significantly. Since the samples—178 and 184—are significantly large, a normal distribution was assumed and the t-test was used for the independent samples. The standard deviations of the two samples ($S_{\text{Typ1}} = 15.88$ monetary units; $S_{\text{Typ2}} = 29.91$ monetary units) differ significantly (Levene's Test for Equality of Variances ($F = 81.23$; $p = 0.000$)), so that the significance of the mean difference can be found in the row "Equal variances not assumed": One errs in $p = 2.8\%$ of cases when different average values are assumed. Therefore, Household type 2, with around 778 monetary units, generates a significant higher average yearly revenue than household 1, which has some 773 monetary units.

Solution 10

(a)

	<51 horsepower	51–75 horsepower	76–100 horsepower	>101 horsepower	Marginal frequency (x)
Vehicle brand 1	100 (80)	55 (68)	5 (8)	0 (4)	160
Vehicle brand 2	0 (20)	30 (17)	5 (2)	5 (1)	40
Marginal frequency (y)	100	85	10	5	200

(b) $5/160 = 3.125\%$

(c) 30/85 = 35.29%

(d)

1. List hypotheses.
2. Calculate the expected frequencies according to the null hypothesis.
3. Test whether no more than 20% of expected frequencies in the cells are less than five. If this is not the case, then the rows or columns must be aggregated until this is no longer the case.
4. Determine the empirical chi-square value.
5. Note the critical value for the given significance level from the table.
6. Decide: If the empirical chi-square value is larger than the critical value, then the two variables are dependent on each other (the null hypothesis is rejected). If the empirical chi-square value is smaller than the critical value, then the variables are independent (and the null hypothesis is not rejected).

(e) Since three of eight (= 37.5% > 20%) of the expected frequencies in the cells of the contingency table are less than five, the two last columns of the table are aggregated. This results in the following contingency table, in which only 16.6% of the expected frequencies in the cells are less than five:

	<51 horsepower	51–75 horsepower	>75 horsepower	Marginal frequency (x)
Vehicle brand 1	100 (80)	55 (68)	5 (12)	160
Vehicle brand 2	0 (20)	30 (17)	10 (3)	40
Marginal frequency (y)	100	85	15	200

In determining the assumed coefficient from the table we get for the critical value of χ^2: $c_{cr} = \chi^2\,[(1-\alpha;(m-1)\cdot(q-1)] = \chi^2\,[99\%;2] = 9.21034$. If the empirical χ^2 is bigger than the critical value ($\chi^2 > c_{cr}$), then H_0 (independence) is rejected. This yields: $\chi^2 = 57.8431 > 9.21034 \rightarrow$ One errs in less than one per cent of the cases when assuming an association between vehicle brand and motor output.

(f) In determining the assumed coefficient from the table we get: $c_{cr} = \chi^2\,[(1-\alpha;(m-1)\cdot(q-1)] = \chi^2\,[99\%;3] = 11.3449$. If $\chi^2 > c_{cr}$, then H_0 (independence) is rejected. This yields: $\chi^2 = 7.128 < 11.3449 \rightarrow$ The relationship between the variables is with a 1% error probability not statistically significant, so that the H_0 hypothesis cannot be rejected.

Solution 11

(a) $\chi^2 = \frac{(130-89.25)^2}{89.25} + \frac{(30-26.25)^2}{26.25} + ... + \frac{(130-85.5)^2}{85.5} = 84.41.$

(b) Degree of freedom: (number of rows − 1)·(number of columns − 1) = (2 − 1)·(3 − 1) = 2; determination of the critical value from the table: $c_{cr} = \chi^2\,[(1-\alpha;(m-1)\cdot(q-1)] = \chi^2\,[95\%;2] = 5.9915$; test decision: if $\chi^2 > c_{cr}$

then H_0 (independence) is rejected. This yields: $\chi^2 = 84.41 > 5.9915 \rightarrow$ One errs in less than 5% of cases when assuming a correlation between purchase expenditures and music.

Solution 12

(a) Because the expected frequency is less than five in more than 20% of the cells, the last three rows are aggregated into one (Solution 1). The chi-square value yields: $\chi^2 = 0+4+0.44+0+1.45+0.16 = 6.06$. It is also possible to aggregate the last two rows and the last two columns (Solution 2). The chi-square value yields: $\chi^2 = 0 + 0 + 0.002 + 0.002 + 0.033 + 0.033 = 0.072$.

(b) Solution 1: Degree of freedom: (number of rows − 1) (number of columns − 1) $= (2-1)(3-1) = 2$; determination of the critical value from the table: $c_{cr} = \chi^2 [(1-\alpha; (m-1)\cdot(q-1)] = \chi^2 [95\%; 2] = 5.9915$; test decision: if $\chi^2 > c_{cr}$ then H_0 (independence) is rejected. This yields: $\chi^2 = 6.06 > 5.9915 \rightarrow$ One errs in less than 5% of cases when assuming a correlation between household size and the number of purchased bananas. H_0 (independence) is rejected. Solution 2: Degree of freedom: (number of rows − 1) (now of columns − 1) $= (3-1)(2-1) = 2$; determination of the critical value from the table: $c_{cr} = \chi^2 [(1-\alpha; (m-1)\cdot(q-1)] = \chi^2 [95\%; 2] = 5.9915$; test decision: if $\chi^2 > c_{cr}$ then H_0 (independence) is rejected. This yields: $\chi^2 = 0.072 < 5.9915 \rightarrow$ One errs in more than 5% of cases when assuming a correlation between household size and the number of purchased bananas. Therefore, H_0 cannot be rejected. Depending on the aggregation, the outcomes can lead to different decisions.

Solution 13

(a) The percentage is $p = 12/80 = 15\%$.

(b) This case involves two ordinal-scaled variables, each capable of having more than two values. Hence, use of the Phi coefficient here does not make sense. Only the coefficients Tau, Spearman, and Cramer's V (the share of cells with an expected frequency of <5 is under 20%) can be applied. None of the results are significant.

(c) The Wilcoxon test tells us that the relationship is significant ($p = 0.000$). That is to say, one errs in less than 0.05% (else p would be $p = 0.001$ or even larger!) of cases when assuming that attitudes towards price before the presentation differ from attitudes towards price after the presentation. The differences between the values are negative in 159 instances and positive in only six. This means that the ranking positions are much more frequently smaller after the presentation than before. According to the given scale, smaller values mean that respondents tend to agree more with the statement "It is a fair price" after the product presentation than before.

(d) The *t*-test for paired/dependent samples indicates that the relationship is significant ($p = 0.000$). As with part (c) of the exercise, respondents think the product price is fairer after the presentation. The test results in a mean difference of (−

0.3) between ratings before the product presentation (2.2) and after (1.9). While the (non-parametric) Wilcoxon test does not require a certain distribution type, the (parametric) *t*-test for paired/dependent samples presupposes a normal distribution of the sample mean. If the assumption of a normal distribution cannot be justified, the non-parametric Wilcoxon test should be applied. In our example, the large sample size of $n = 500$ allows us to assume that the sample mean will approximate a normal distribution.

Solution 14

A one-sample *t*-test was used to see whether the average age in a population differs from $\mu = 54$ years. This test requires a large sample size, a condition satisfied by $n = 500$. With a standard deviation of $S = 10.1$ years, the sample mean is $\bar{x} = 52.9$ years. The probability of error is $p = 1.3\%$ when assuming that the average age is not $\mu = 54$. Given a significance level of 5%, the average age differs significantly from $\mu_0 = 54$ years. The 95% confidence interval for age is thus:

$$\mu_{\text{lower/upper}} = [54 - 2.01; 54 - 0.24] = [51.99; 53.76]$$

Solution 15

(a) The homogeneity of variance requirement is satisfied, as Levene's test is insignificant ($p < 0.05$). The Kolmogorov–Smirnov test yields significant results so that the assumption of the normal distribution cannot be met. Although the samples are very large, a bias of the results cannot be excluded. A Kruskal–Wallis test in addition to ANOVA should be carried out at this point (see Fig. 9.63).
(b) ANOVA shows significant differences between each television show (see Fig. 9.63).
(c) ANOVA shows significant differences between each time slot (see Fig. 9.63).
(d) An interaction effect occurs because television viewers have different preferences at different times (see Fig. 9.63).
(e) Although ZDF is more affordable, it reaches fewer viewers at a given time slot than ARD. The problem of costs per viewer cannot be solved with the available data.

Levene's Test of Equality of Error Variances[a]

Dependent Variable: Number of households reached

F	df1	df2	Sig.
1.298	11	988	.220

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + Channel + Time + Channel * Time

Tests of Between-Subjects Effects

Dependent Variable: Number of households reached

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2713194.965[a]	11	246654.088	2799.053	.000
Intercept	11877229.844	1	11877229.844	134783.866	.000
Channel	1449354.541	3	483118.180	5482.468	.000
Time	745670.711	2	372835.356	4230.969	.000
Channel * Time	438971.020	6	73161.837	830.247	.000
Error	87063.114	988	88.121		
Total	18092301.000	1000			
Corrected Total	2800258.079	999			

a. R Squared = .969 (Adjusted R Squared = .969)

Multiple Comparisons

Dependent Variable: Number of households reached

Scheffé

(I) Television Channels	(J) Television Channels	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
ARD	ZDF	-11.04*	.791	.000	-13.26	-8.82
	RTL	30.95*	.778	.000	28.77	33.13
	SAT1/Pro7	97.47*	.883	.000	94.99	99.94
ZDF	ARD	11.04*	.791	.000	8.82	13.26
	RTL	41.99*	.862	.000	39.58	44.40
	SAT1/Pro7	108.50*	.957	.000	105.83	111.18
RTL	ARD	-30.95*	.778	.000	-33.13	-28.77
	ZDF	-41.99*	.862	.000	-44.40	-39.58
	SAT1/Pro7	66.52*	.946	.000	63.87	69.17
SAT1/Pro7	ARD	-97.47*	.883	.000	-99.94	-94.99
	ZDF	-108.50*	.957	.000	-111.18	-105.83

Fig. 9.63 ANOVA for Solution 1 (SPSS)

RTL	-66.52*	.946	.000	-69.17	-63.87

Based on observed means. The error term is Mean Square(Error) = 88.121.
*. The mean difference is significant at the .05 level.

Multiple Comparisons

Dependent Variable: Number of households reached
Scheffé

(I) Time slot	(J) Time slot	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
17.00-17.59	18.00-18.59	-9.97*	.740	.000	-11.78	-8.16
	19.00-19.59	-63.19*	.723	.000	-64.96	-61.42
18.00-18.59	17.00-17.59	9.97*	.740	.000	8.16	11.78
	19.00-19.59	-53.22*	.721	.000	-54.99	-51.46
19.00-19.59	17.00-17.59	63.19*	.723	.000	61.42	64.96
	18.00-18.59	53.22*	.721	.000	51.46	54.99

Based on observed means. The error term is Mean Square(Error) = 88.121.
*. The mean difference is significant at the .05 level.

Fig. 9.63 (continued)

Solution 16

(a) Green. Marginal means are always larger than other colours; the attendant variable is significant (see Fig. 9.64).
(b) Single-factor ANOVA produces significant differences. Scheffé's method shows differences for the UK (least number of visitors), for Spain (average number of visitors), and for Germany and France (highest number of visitors). The difference in the numbers of visitors in Germany and France is not statistically significant (see Fig. 9.64).
(c) Though green is usually received best, the colours have different effects in different countries. In the multiple-factor ANOVA, there is an interaction between the country (significant by itself) and colour significant by itself (see Fig. 9.64).
(d) The homogeneity of variance requirement is satisfied, since Levene's test is insignificant (see Fig. 9.64). The Kolmogorov–Smirnov test yields significant results so that the assumption of the normal distribution cannot be met. Although the samples are very large, a bias of the results cannot be excluded. A Kruskal–Wallis test should be carried out at this point.

Solution 17

(a) There are significant differences. The best values occur with Zidane as advertising ambassador (see Fig. 9.65).

Levene's Test of Equality of Error Variances[a]

Dependent Variable: Number of visitors

F	df1	df2	Sig.
.688	11	988	.751

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + country + colour + country * colour

Tests of Between-Subjects Effects

Dependent Variable: Number of visitors

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	268649464.547[a]	11	24422678.595	2489.557	.000
Intercept	1705513916.768	1	1705513916.768	173853.769	.000
Country	96365533.102	3	32121844.367	3274.382	.000
Colour	63274857.455	2	31637428.728	3225.002	.000
Country * colour	59967125.291	6	9994520.882	1018.804	.000
Error	9692327.997	988	9810.049		
Total	2383406174.000	1000			
Corrected Total	278341792.544	999			

a. R Squared = .965 (Adjusted R Squared = .965)

Multiple Comparisons

Dependent Variable: Number of visitors

Scheffé

(I) Country	(J) Country	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
	France	23.30*	8.165	.044	.44	46.17
Germany	Spain	317.46*	8.325	.000	294.15	340.77
	UK	971.18*	9.541	.000	944.46	997.89
	Germany	-23.30*	8.165	.044	-46.17	-.44
France	Spain	294.16*	8.838	.000	269.41	318.91
	UK	947.88*	9.991	.000	919.90	975.85
	Germany	-317.46*	8.325	.000	-340.77	-294.15
Spain	France	-294.16*	8.838	.000	-318.91	-269.41
	UK	653.72*	10.122	.000	625.37	682.06
UK	Germany	-971.18*	9.541	.000	-997.89	-944.46

Fig. 9.64 ANOVA of Solution 2 (SPSS)

	France	-947.88*	9.991	.000	-975.85	-919.90
	Spain	-653.72*	10.122	.000	-682.06	-625.37

Based on observed means. The error term is Mean Square(Error) = 9810.049.
*. The mean difference is significant at the .05 level.

Multiple Comparisons

Dependent Variable: Number of visitors
Scheffé

(I) Colour	(J) Colour	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Yellow	Red	291.77*	7.626	.000	273.08	310.47
	Green	-511.40*	7.788	.000	-530.49	-492.31
Red	Yellow	-291.77*	7.626	.000	-310.47	-273.08
	Green	-803.17*	7.620	.000	-821.85	-784.49
Green	Yellow	511.40*	7.788	.000	492.31	530.49
	Red	803.17*	7.620	.000	784.49	821.85

Based on observed means. The error term is Mean Square(Error) = 9810.049.
*. The mean difference is significant at the .05 level.

Fig. 9.64 (continued)

(b) There are significant differences. The best values occur with the slogan *The Chocolate of the Champs* (see Fig. 9.65).
(c) The homogeneity of variance requirement is satisfied, since Levene's test is insignificant. A decision was made to choose Zidane as the advertising ambassador, however, because of the interaction effect of the slogan *Chocolate, Ohlala.* Due to the violation of the normal distribution assumption, the results should be checked using the Kruskal–Wallis test (see Fig. 9.65). What ANOVA cannot do is make statements about the marginal effects, as is possible with regression analysis.

Levene's Test of Equality of Error Variances[a]

Dependent Variable: Sales

F	df1	df2	Sig.
.825	8	491	.581

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + Ambassador + Slogan + Ambassador * Slogan

Tests of Between-Subjects Effects

Dependent Variable: Sales

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	63380369.7[a]	8	7922546.2	766.1	.000
Intercept	2596566347.0	1	2596566347.0	251089.3	.000
Ambassador	13369405.0	2	6684702.5	646.4	.000
Slogan	17676416.9	2	7598443.2	854.7	.000
Ambassador*Slogan	30393772.7	4	10341.2	734.8	.000
Error	5077532.0	491			
Total	2795172865.0	500			
Corrected Total	68457901.7	499			

a. R Squared = .926 (Adjusted R Squared = .925)

Multiple Comparisons

Dependent Variable: Sales

Scheffé

(I) Ambassador	(J) Ambassador	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Kahn	Zidane	-445.80*	10.871	.000	-472.49	-419.11
	Figo	-206.52*	11.330	.000	-234.34	-178.70
Zidane	Kahn	445.80*	10.871	.000	419.11	472.49
	Figo	239.28*	11.300	.000	211.54	267.03
Figo	Kahn	206.52*	11.330	.000	178.70	234.34
	Zidane	-239.28*	11.300	.000	-267.03	-211.54

Based on observed means. The error term is Mean Square(Error) = 10341.206.

*. The mean difference is significant at the .05 level.

Multiple Comparisons

Fig. 9.65 ANOVA of Solution 3 (SPSS)

Dependent Variable: Sales
Scheffé

(I) Slogan	(J) Slogan	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Chocolate with discipline	Chocolate, Ohlala	-382.00*	11.376	.000	-409.93	-354.07
	Chocolate of the Champs	-429.65*	11.391	.000	-457.61	-401.68
Chocolate, Ohlala	Chocolate with discipline	382.00*	11.376	.000	354.07	409.93
	Chocolate of the Champs	-47.65*	10.795	.000	-74.15	-21.15
Chocolate of the Champs	Chocolate with discipline	429.65*	11.391	.000	401.68	457.61
	Chocolate, Ohlala	47.65*	10.795	.000	21.15	74.15

Based on observed means. The error term is Mean Square(Error) = 10341.206.
*. The mean difference is significant at the .05 level.

Fig. 9.65 (continued)

References

Backhaus, K., Erichson, B., Gensler, S., Weiber, R., & Weiber, T. (2021). *Multivariate Analysemethoden. Eine Anwendungsorientierte Einführung* (16th ed.). Springer.

British Board of Trade (1990). *Report on the Loss of the 'Titanic' (S.S.). British Board of Trade Inquiry Report (reprint).* Gloucester, UK: Allan Sutton Publishing.

Bortz, J. (1999). *Statistik für Sozialwissenschaftler* (5th ed.). Springer.

Bortz, J., Lienert, G. A., & Boehnke, K. (2000). *Verteilungsfreie Methoden der Biostatistik* (2nd ed.). Springer.

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association, 69*, 364–367.

Conover, W. J. (1980). *Practical nonparametric statistics.* Wiley.

Dixon, W. J. (1954). Power under normality of several nonparametric tests. *The Annals of Mathematical Statistics, 25*, 610–614.

Field, A. (2005). *Discovering statistics using SPSS*. Sage.

Fisher, R. A., & Yates, F. (1963). *Statistical tables for biological, agricultural, and medical research.* Oliver and Boyed.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed.). Prentice Hall.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell' Istituto Italiano degli Attuari, 4*, 83–91.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 47*, 583–621.

Kruskal, W. H., & Wallis, W. A. (1953). Errata: Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 48*, 907–911.

Mann, H. B., & Whitney, D. R. (1947). On a test whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics, 18*, 65–78.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society, Series A*, 236.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference, part I. *Biometrika, 20A*, 175–240.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference, part II. *Biometrika, 20A*, 263–294.

Popper, K. (1934). *Logik der Forschung*. Mohr.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2*, 110–114.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika, 40*, 87–104.

Shapiro, S. S., & Francia, R. S. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association, 67*, 215–216.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*, 591–611.

Smirnov, N. V. (1933). Estimate of deviation between empirical distribution functions in two independent samples. *Bulletin Moscow University, 2*, 3–16.

Spearman, C. E. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*(1), 72–101.

Stevens, J. P. (1972). Four methods of analyzing between variations for the k-group MANOVA problem. *Multivariate Behavioral Research, 7*, 442–454.

Welch, B. L. (1947). The generalization of student's problem when several different population variances are involved. *Biometrika, 34*, 28–35.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics, 1*, 80–83.

Wilcoxon, F. (1947). Probability tables for individual comparisons by ranking methods. *Biometrics, 3*(119), 122.

Witting, H. (1960). A generalized Pitman efficiency for nonparametric tests. *The Annals of Mathematical Statistics, 31*, 405–414.



Regression Analysis

10

10.1 First Steps in Regression Analysis

Regression analysis—often referred to simply as regression—is an important tool in statistical analysis. The concept first appeared in an 1877 study on sweet-pea seeds by Sir Francis Galton (1822–1911). He used the idea of regression again in a later study on the heights of fathers and sons. He discovered that sons of tall fathers are tall but somewhat shorter than their fathers, while sons of short fathers are short but somewhat taller than their fathers. In other words, body height tends towards the mean. Galton called this process a *regression*—literally, a step back or decline. We can perform a correlation to measure the association between the heights of sons and fathers. We can also infer the *causal direction of the association*. The height of sons depends on the height of fathers and not the other way around. Galton indicated causal direction by referring to the height of sons as the *dependent variable* and the height of fathers as the *independent variable*. But take heed: regression does not necessarily prove the causality of the association. The direction of effect must be derived theoretically before it can be empirically proven with regression. Sometimes the direction of causality cannot be determined, as for example between the ages of couples getting married. Does the age of the groom determine the age of the bride, or vice versa? Or do the groom's age and the bride's age determine each other mutually? Sometimes the causality is obvious. So, for instance, blood pressure has no influence on age, but age has influence on blood pressure. Body height has an influence on weight, but the reverse association is unlikely (Swoboda, 1971, p. 308).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78070-7_10.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*,
Springer Texts in Business and Economics,
https://doi.org/10.1007/978-3-031-78070-7_10

Let us approach the topic of regression analysis with an example (see dataset *Chapter10_mail_order_business*; supplementary material at https://doi.org/10.1007/978-981-96-2524-6_10). A mail order business adds a new summer dress to its collection. The purchasing manager needs to know how many dresses to buy so that by the end of the season the total quantity purchased equals the quantity ordered by customers. To prevent stock shortages (i.e. customers going without wares) and stock surpluses (i.e. the business is left stuck with extra dresses), the purchasing managing decides to carry out a sales forecast.

What's the best way to forecast sales? The economist immediately thinks of several possible predictors or influencing variables. How high are sales of a similar dress in the previous year? How high is the price? How large is the image of the dress in the catalogue? How large is the advertising budget for the dress? But we don't only want to know which independent variables exert an influence; we want to know how large the respective influence is. To know that catalogue image size has an influence on the number of orders does not suffice. We need to find out the number of orders that can be expected on average when the image size is, say, 50 cm^2.

Let us first consider the case where future demand is estimated from the sales of a similar dress from the previous year. Figure 10.1 displays the association as a scatterplot for 100 dresses of a given price category, with the future demand plotted on the y-axis and the demand from the previous year plotted on the *x*-axis.

If all the plots lay on the angle bisector, the future demand of period (t) would equal the sold quantities of the previous year ($t - 1$). As is easy to see, this is only rarely the case. The scatterplot that results contains some large deviations, producing a correlation coefficient of only $r = 0.42$.

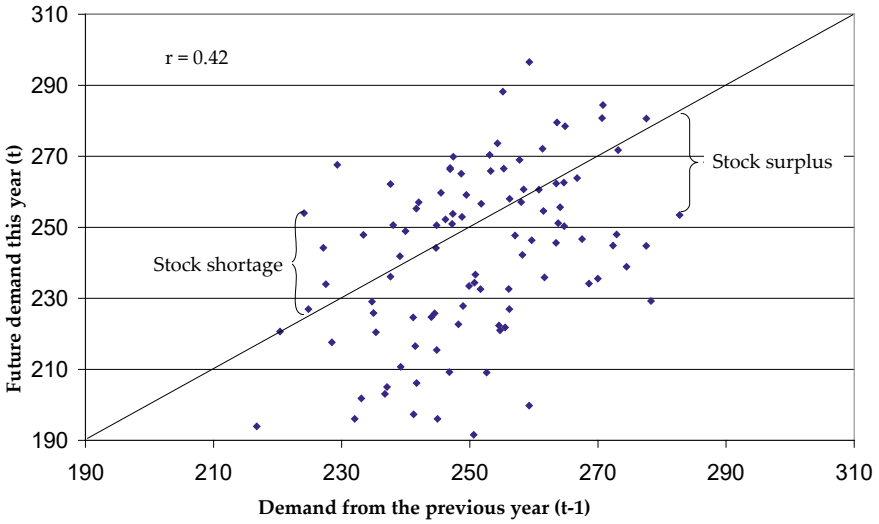


Fig. 10.1 Demand forecast using equivalence

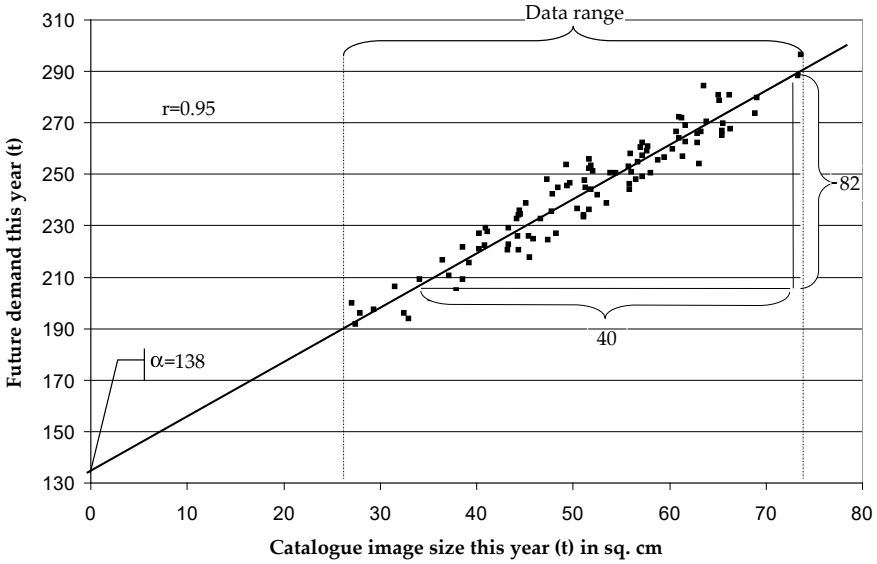


Fig. 10.2 Demand forecast using image size

Now if, instead of equivalent dresses from the previous year, we take into account the catalogue image size for the current season (t), we arrive at the scatterplot in Fig. 10.2.

We see immediately that the data points lie much closer to the line, which was drawn to best approximate the course of the data. This line is more suited for a sales forecast than a line produced using the "equivalence method" in Fig. 10.1. Of course, the proximity of the points to the line can be manipulated through axis scale. The relatively large correlation coefficient of $r = 0.95$, however, ultimately shows that the linear association between these variables is stronger. The points lie much closer to the line, which means that the sales forecast will result in fewer costs for stock shortages and stock surpluses. But, again, this applies only for products of the same quality and in a specific price category.

10.2 Coefficients of Bivariate Regression

Now we want to determine the association so we can be better predict future sales. We begin with the reasonable assumption that the relationship between catalogue image size and actual sales is *linear*. We then generate a regression line to identify an association that more or less represents the scatterplot of the data points. The linear equation consists of two components:

- The intercept is where the line crosses the y-axis. We call this point α. It determines the distance of the line along the y-axis to the origin.

- The slope coefficient (β) indicates the slope of the line. From this coefficient we can determine to what extent catalogue image size impacts demand. If the slope of the lines is two, the value on the y-axis changes by two units, while the value on the x-axis changes by one unit. In other words, the flatter the slope, the less influence x values have on the y-axis.

The line in the scatterplot in Fig. 10.2 can be represented with the algebraic linear equation

$$\hat{y} = \alpha + \beta \cdot x \tag{10.1}$$

This equation intersects the y-axis at the value 138, so that $\alpha = 138$ (see Fig. 10.2). Its slope is calculated from the slope triangle (quotient) $\beta = 82/40 \approx 2.1$. When the image size increases by ten square centimetres, the demand increases by 21 dresses. The total linear equation is

$$\hat{y} = 138 + 2.1 \cdot x. \tag{10.2}$$

For a dress with an image size of 50 cm^2, we can expect sales to be

$$\hat{y} = 138 + 2.1 \cdot 50 = 243 \tag{10.3}$$

With an image size of 70 cm^2, the equation is

$$\hat{y} = 138 + 2.1 \cdot 70 = 285 \text{ dresses.} \tag{10.4}$$

This linear estimation approximates the average influence of x variables on y variables using a mathematical function. The estimated values are indicated by $\hat{y}_i$ and the realized y values are indicated by y_i. Although the linear estimation runs through the entire quadrant, the association between the x and y variable is only calculated for the area that contains data points, referred to as the data range. If we use the regression function for estimations outside this area (as part of a forecast, for instance), we must assume that the association identified outside the data range does not differ from the associations within the data range.

To better illustrate this point, consider Fig. 10.3. The marked data point corresponds to dress model 23, which was advertised with an image size of 47.4 cm^2 and which was later sold 248 times. The linear regression estimates average sales of 238 dresses for this image size. The difference between actual sales and estimated sales is referred to as the residual or the error term. It is calculated by:

$$u_i = (y_i - \hat{y}_i) \tag{10.5}$$

For dress model 23 the residual is:

$$u_{23} = (y_{23} - \hat{y}_{23}) = 248 - 237.5 = 10.5 \tag{10.6}$$

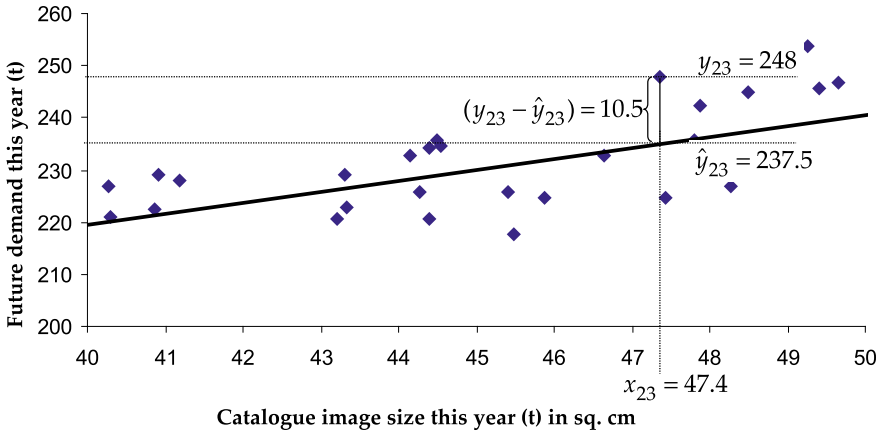


Fig. 10.3 Calculating residuals

In this way, every data point can be expressed as a combination of the result of the linear regression $\hat{y}$ and its residual:

$$y_i = \hat{y}_i + u_i \tag{10.7}$$

We have yet to explain which rule applies for determining this line and how it can be derived algebraically. Up to now we only expected that the line run as closely as possible to as many data points as possible and that deviations above and below the line be kept to a minimum and be distributed non-systematically. The deviations in Fig. 10.2 between actual demand and the regression line create stock shortages when they are located above and stock surpluses when they are located below. Since we want to prevent both, we can position the line so that the *sum of deviations* between realized points y_i and the points on line $\hat{y}_i$ is as close to zero as possible. The problem with this approach is that a variety of possible lines with different qualities of fit all fulfil this condition. A selection of possible lines is shown in Fig. 10.4.

The reason for this is simple: the deviations above and below cancel each other out, resulting in a sum of zero. All lines that run through the bivariate centroid—the value pair of the averages of x and y—fulfil the condition

$$\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0 \tag{10.8}$$

But in view of the differences in quality among the lines, the condition above makes little sense as a construction criterion. Instead, we need a line that does not allow deviations to cancel each other yet still limits the total sum of errors. Frequently, statisticians create a line that minimises *the sum of the squared deviations*

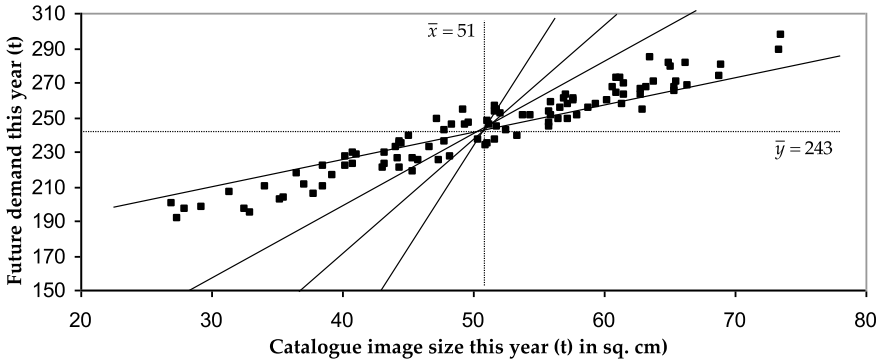


Fig. 10.4 Lines of best fit with a minimum sum of deviations

of the actual data points y_i from the points on the line $\hat{y}_i$ The minimization of the entire deviation error is

$$\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \rightarrow \min . \tag{10.9}$$

This method of generating the regression line is called the *ordinary least squares method*, or OLS. It can be shown that these lines also run through the bivariate centroid—i.e. the value pair $(\overline{x}; \overline{y})$—but this time we only have a single regression line, which fulfils the condition of the minimal squared error. If we insert equation of the regression line for $\hat{y}_i$, we get:

$$f(\alpha; \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \rightarrow \min . \tag{10.10}$$

The minimum can be achieved by using the necessary conditions for a minimum, deriving the function $f(\alpha; \beta)$ once according to α and once according to β and setting both deviations equal to zero.

$$\text{(i)} \quad \frac{\partial f(\alpha, \beta)}{\partial \alpha} = \sum_{i=1}^{n} 2 \cdot (y_i - \alpha - \beta \cdot x_i) \cdot (-1) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} y_i = n \cdot \alpha + \beta \sum_{i=1}^{n} x_i \Leftrightarrow \alpha = \overline{y} - \beta \cdot \overline{x} \tag{10.11}$$

$$\text{(ii)} \quad \frac{\partial f(\alpha, \beta)}{\partial \beta} = \sum_{i=1}^{n} 2 \cdot (y_i - \alpha - \beta \cdot x_i) \cdot (-x_i) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} (x_i y_i) = \alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 \tag{10.12}$$

The reformulation in (i) yields the formula for the constant α. We then equate (i) and (ii) to get:

$$n \cdot a + \beta \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i = \alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i, \tag{10.13}$$

so that

$$\beta = \frac{\alpha \sum_{i=1}^{n} x_i - n \cdot a - \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i}{\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^2\right)} \tag{10.14}$$

By inserting this equation in (i) we get

$$\alpha = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{10.15}$$

If we place the latter in (ii) we get:

$$\sum_{i=1}^{n} (x_i y_i) = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 \tag{10.16}$$

After several reformulations, we arrive at the formula for the slope coefficient:

$$\frac{n \sum_{i=1}^{n} (x_i y_i) - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\text{cov}(x, y)}{S_x^2} = \frac{r \cdot S_y}{S_x} = \beta \tag{10.17}$$

Of course, the regression coefficient no longer needs to be calculated by hand. Today's statistics software does it for you. Excel, for instance, has functions for determining a regression's slope and intercept. Section 10.5 discusses the use of computer applications for calculating regression.

10.3 Multiple Regression Coefficients

In the previous section, we discussed methods of regression analysis for bivariate associations. These approaches may suffice for simple models, but what do we do when there is reason to assume that a whole cluster of factors influence the dependent variable? Let's return to the mail order business example. We found that a sales forecast based on catalogue image size was better than one based on sales of an equivalent dress from the previous year. But in practice is there ever only one influencing factor at work? Realistically speaking, rarely. So why not use both variables—image size and previous sales—to estimate sales? The derivation

of association using multiple regression is analogous to that using bivariate regression. We again assume that $\alpha = \beta_0$ and that β_1 and β_2 are such that the sum of the squared residuals is minimal. In the general case of k independent variables and n observations, regression can be calculated by the following matrix equation:

$$y = X \cdot \beta + u = \begin{bmatrix} y_0 \\ \cdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1+ & x_{11} + \cdots + & x_{k1} \\ \cdots + & \cdots + \cdots + & \cdots \\ 1+ & x_{1n} + \cdots + & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \cdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ \cdots \\ u_n \end{bmatrix}$$
$$= \begin{bmatrix} \beta_0+ & \beta_1 x_{11} + \cdots + & \beta_k x_{k1} & +u_1 \\ \cdots + & \cdots & +\cdots + \cdots & +\cdots \\ \beta_0+ & \beta_1 x_{1n} + \cdots + & \beta_k x_{kn} & +u_n \end{bmatrix} \tag{10.18}$$

It can be shown that the minimum sum of squared residuals obtains exactly when the vector of the regression coefficients $\beta = (\alpha = \beta_0; \beta_1; \ldots; \beta_k)$ equals

$$\beta = \left(X'X\right)^{-1} X'y \tag{10.19}$$

Once again we could use the OLS method, though here the regression equation consists of more than two components:

- the constant $\alpha = \beta_0$;
- the first slope coefficient β_1, which describes the relationship between catalogue image size and demand; and
- the second slope coefficient β_2, which describes the relationship between previous sales and demand.

The equation to determine the multiple regression is thus:

$$\hat{y} = \alpha + \beta_1 \cdot \text{catalogue image size} + \beta_2 \cdot \text{previous sales} = \alpha + \beta_1 x_1 + \beta_2 x_2 \tag{10.20}$$

For our forecast example, the empirical calculations yield the following result:

$$\hat{y} = \alpha + 1.95 \cdot \text{catalogue image size} + 0.33 \cdot \text{previous sales} \tag{10.21}$$

The regression coefficient is determined to be $\beta_1 = 1.95$ for the catalogue image size and $\beta_2 = 0.33$ for the previous sales. At this point one can ask whether the influence of each of the two variables is statistically significant. For this, each of the two regression coefficients (β_1 and β_2) must differ significantly from zero. We have to test the following hypotheses:

$$\mathrm{H}_0 : \beta_1 = 0 \text{ versus } \mathrm{H}_1 : \beta_1 \neq 0 \tag{10.22}$$

$$\mathrm{H}_0 : \beta_2 = 0 \text{ versus } \mathrm{H}_1 : \beta_2 \neq 0 \tag{10.23}$$

As we will see later in chapter 10.5, software will provide us with the p-values for each of the two regression coefficients β_1 and β_2. If the respective p-value lies below the significance level α, we must reject the corresponding null hypothesis. In this case, the values of the dependent variable (e.g. future sales) are determined linearly by the values of the respective independent variable (e.g. size of the catalogue and/or previous sales). On the other hand, an independent variable has no effect on the dependent variable if H_0 cannot be rejected ($p > \alpha$). The significance level α is usually set at 1%, 5%, or 10%, though 5% is the most common value.

10.4 The Goodness of Fit of Regression Lines

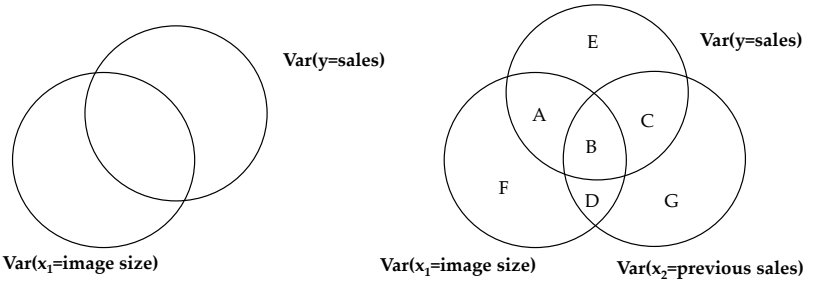
A regression seeks to describe the average association of two or more variables. In Figs. 10.1 and 10.2 we saw how regression lines can overestimate or underestimate the y values of data points. Because these kinds of errors can lead to costly surpluses and shortages, it is crucial that regression lines have a good fit. In the previous section we determined that catalogue image size (Fig. 10.2) is better suited for predicting sales than the previous sales ("equivalence") method (Fig. 10.1), as the former produced data points with greater proximity to the regression line and a greater correlation coefficient. Generally, the closer the data points are to the regression line the better the regression line is. When all the data points lie on the line, the linear regression is perfect, and the correlation coefficient is either $r = (+1)$ or $r = (-1)$. By contrast, when the data points are scattered far from the regression line, the correlation coefficient is close to zero, and the resulting forecast will be too imprecise.

Here we see that the correlation coefficient can serve to evaluate the goodness of fit with bivariate analysis. But the more common parameter is the *coefficient of determination*, symbolized by R^2. The coefficient of determination equals the square of the correlation coefficient for bivariate regressions, but it can be applied when multiple independent x variables exist. Because R^2 is squared, it only takes values between zero and one. $R^2 = 0$ when the goodness to fit is poor, and $R^2 = 1$ when the goodness to fit is perfect.

The coefficient of determination also indicates the share of y variance explained by x variance. In our example (see Fig. 10.2) the coefficient of determination is $R^2 = 0.96^2 = 0.9216 = 92.16\%$. This means that 92.16% of the variance in sales (y variable) is explained through variance in catalogue image size (x variable).

Figure 10.5 illustrates the explanation share of variance using Venn diagrams. Part 1 represents a bivariate regression, also known as a simple regression. The upper circle indicates the variance of the dependent y variables (sales); the lower circle indicates the variance of x_1 (image size). The region of intersection represents the share of y variance (sales) explained by the x_1 variance (image size). The larger the area of intersection is, the better the x_1 variable (image size) explains variance in the dependent y variable.

Part 2 takes into account the other variable as well: the previous year's sales (x_2). Here the intersection between y variance (sales) and x_1 variance (image



Bivariate Regression ("Simple Regression"):
The region of intersection represents the share of variance in y (sales) explained by variance in x_1 (image size)

Multiple Regression ("Multiple Regression"):
The regions of intersection A, B, and C represent the share of variance in y (sales) explained by variance in x_1 (image size) and x_2 (previous year's

Part 1: Part 2:

Fig. 10.5 The concept of multivariate analysis

size) and the previous year's sales (x_2) increases. With the regression lines $\hat{y}$, the variances of the independent x variables explain

$$R^2 = \left(\frac{(A + B + C)}{(A + B + C + E)} \right) \cdot 100\% \tag{10.24}$$

of y variance. The general formula for R^2 in a multiple regression can thus be expressed as follows:

$$R^2 = \frac{S_{\hat{y}}^2}{S_y^2} = \frac{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{10.25}$$

Often, statisticians subtract $\frac{1}{n}$ from the quotient of variance to calculate R^2, instead of using the quotient of variance alone. The quotient of variance consists of the *explained regression sum of squares* $\text{RSS} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$ divided by the *total sum of squares* $\text{TSS} = \sum_{i=1}^{n} (y_i - \overline{y})^2$:

$$R^2 = \frac{\text{RSS}}{\text{TSS}} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{10.26}$$

R^2 can also be calculated using the unexplained variance of y variables:

$$S_e^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{10.27}$$

In part 2 above, the unexplained variance is represented by region E. The correlation of determination can then be defined as follows:

$$R^2 = 1 - \frac{S_e}{S_y} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{10.28}$$

Expressed using the *residual or error sum of squares* $\text{ESS} = \sum_{i=1}^{n}(y_i - \overline{y})^2$, R^2 is

$$R^2 = 1 - \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{10.29}$$

Another way to evaluate goodness of fit with multiple regression is the adjusted coefficient of determination. We will learn about this approach in Sect. 10.6.

10.5 Regression Calculations with the Computer

This section uses the file *Chapter10_mail_order_business* (see supplementary material at https://doi.org/10.1007/978-981-96-2524-6_10). Each statistics package presents the results in a similar tabular form. The first table contains regression statistics such as the absolute value of the correlation coefficient and the coefficient of determination; the second table contains the sum of squares; and the third table displays the regression coefficient statistics. Figure 10.6 shows the tables of results using the example of the SPSS software.

Both regression coefficients are significantly different from zero ($p = 0.000 < 0.05$ for the catalogue image size and $p = 0.000 < 0.05$ for sales of a similar dress at $t - 1$). From these results we can determine the sales of period (t) using the following equation:

$$\hat{y} = 62.22 + 1.95 \text{ catalogue size} + 0.33 \text{ sales of a similar dress at } t - 1 \tag{10.30}$$

If a dress is advertised with an image of 50 cm^2 and a similar dress sold 150 times last year, then we can expect average sales of

$$\hat{y} = 62.22 + 1.95\ 50 + 0.33\ 150 \approx 209 \text{ dresses} \tag{10.31}$$

The sum of squares explained by the regression is 52,733.837. The total sum of squares to be explained is 55,999.390, so that the sum of squares unexplained by the regression is $55{,}999.390 - 52{,}733.837 = 3265.553$. From this we can also calculate the coefficient of determination, were it not already indicated above:

$$R^2 = \frac{52{,}733.873}{55{,}999.390} = 94.2\% \tag{10.32}$$

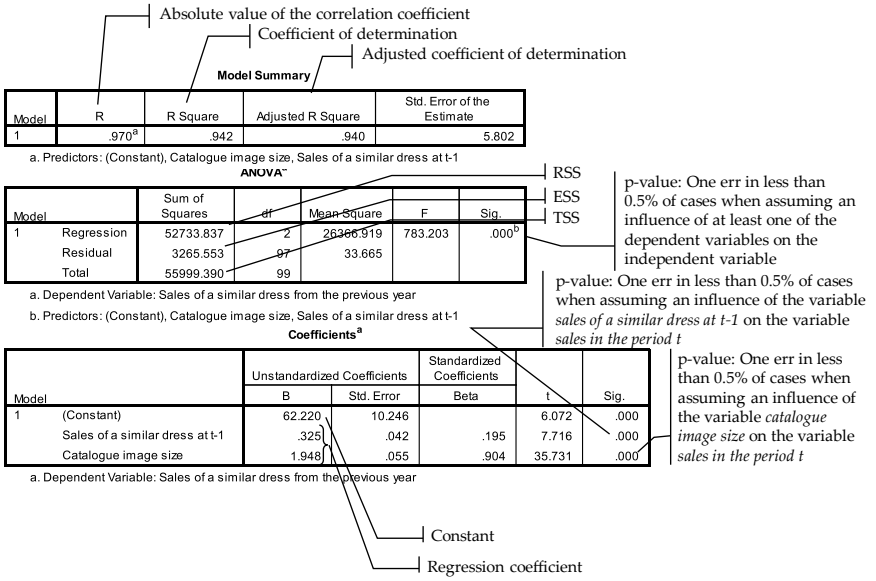


Fig. 10.6 Output from the regression function (for SPSS)

The variance of the independent x variables (the demand of a similar dress in the season $t - 1$; the catalogue image size) explains the variance of the dependent variable (the sales of a dress in the season t) for $R^2 = 94.2\%$.

The following sections briefly describe how to generate the above results using SPSS, Stata, R and Excel.

10.5.1 Regression Calculations with SPSS

In SPSS, open the *Linear Regression* window shown in Fig. 10.7 by selecting *Analyze* → *Regression* → *Linear*. Then assign the dependent and independent variables and confirm the selection by clicking *OK*.

10.5.2 Regression Calculations with Stata

With Stata access the regression menu by choosing *Statistics* → *Linear models and related* → *Linear regression*. Then enter the dependent variable in the *Dependent variable* field and the independent variables in the *Independent variables* field (see Fig. 10.8) and click *OK* or *Submit*.

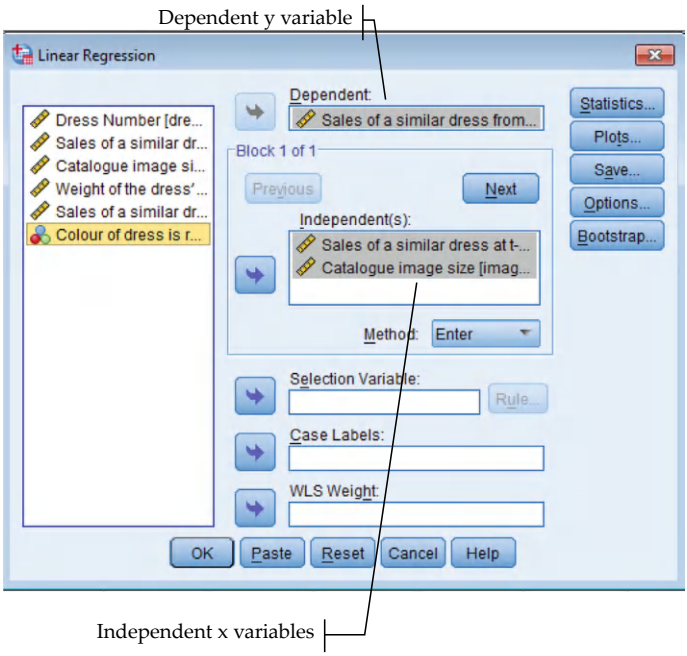


Fig. 10.7 Regression with SPSS

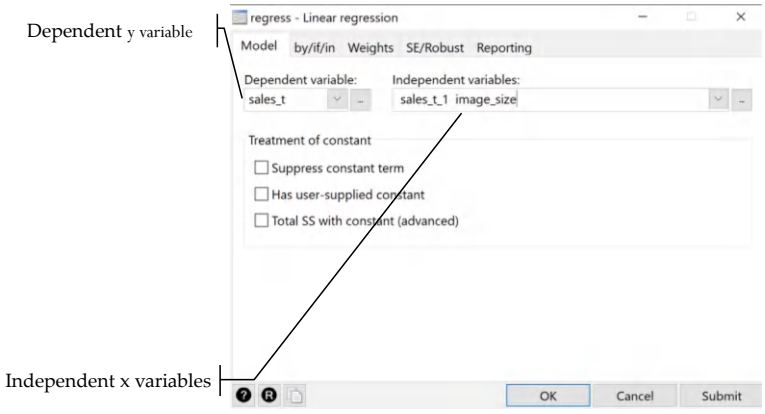


Fig. 10.8 Regression with Stata

10.5.3 Regression Calculations with R Commander

With R Commander access the regression menu by choosing *Statistics* → *Fit Models* → *Linear regression…* Then enter the dependent variable in the *Response*

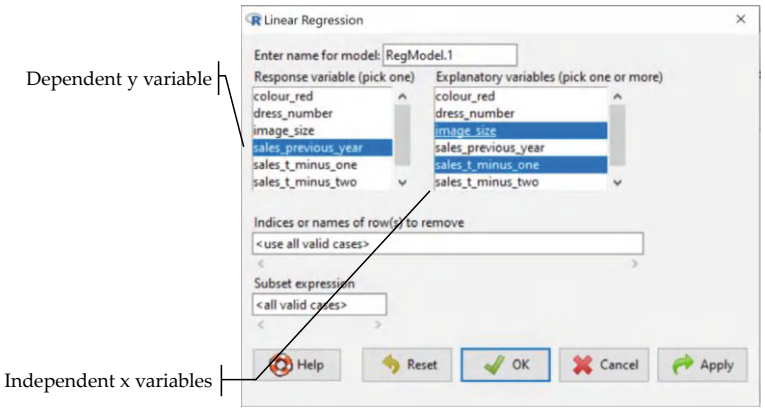


Fig. 10.9 Regression with R commander

variable field and the independent variables in the *Exploratory variables* field (see Fig. 10.9) and click *OK* or *Apply*.

10.5.4 Regression Calculations with Excel

Excel's *Linest function* calculates the most important regression parameters. But this function is relatively inflexible and complicated to use.[1] A more flexible approach is Excel's *regression* function. To use it, first activate the analysis function via the Add-Ins Manager.[2] Now select the *regression function* under *Data* → *Data Analysis* so that the window presented in Fig. 10.10 appears. Next assign the fields for dependent and independent variables. Keep in mind that the

[1] To use the *Linest function* for the dataset *Chapter10_mail_order_business.xls* (see supplementary material at https://doi.org/10.1007/978-981-96-2524-6_10), mark a field in the Excel sheet in which the regression results are to appear. With k regressors—in our case $k = 2$—this field must have five lines and $k + 1$ rows. Next choose the Linest command under *Formulas* → *Insert Function* → *Statistical*. Insert the dependent y variables (B2:B101) into the field *Known_y's*, and the *x* variables (C2:D101) into the field *Known_x's*. If the regression contains a constant, the value one must be entered into the *const field* and the *stats field*. The command will NOT be activated by the enter button, but by the simultaneous activation of the buttons STRING + SHIFT + ENTER. In the first line, the coefficients β_1 to β_k are displayed. The last row of the first line contains the value of the constant β. The other lines display the remaining parameters, some of which we have yet to discuss. The second line shows the standard errors of the coefficients; the third line, the coefficient of determination (R^2) and the standard error of the residuals; the fourth line, the f value and the degree of freedom. The last line contains the sum of squares of the regression (RSS) and residuals (ESS).

[2] The Add-Ins Manager can be accessed via *File* → *Options* → *Add-ins* → Manage: Excel Add-ins → Go…

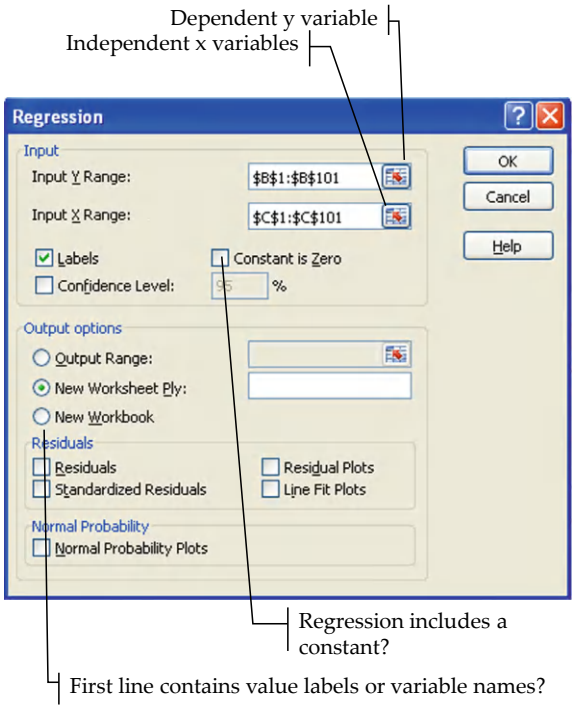


Fig. 10.10 Regression with excel

independent variables must be arranged next to each other in the Excel tables and may contain no missing values.

10.6 Goodness of Fit of Multiple Regressions

The inclusion of an additional predictor variable x improved our model, as the coefficient of determination could be increased from $R^2 = 0.90$ for a regression only considering image size to $R^2 = 0.94$.

Which value would the coefficient of determination have assumed had we substituted for *previous year*'s *sales of a similar dress* a completely crazy variable such as the body weight of the dress's seamstress? By definition, the coefficient of determination remains constant at $R^2 = 0.90$, as the catalogue image size retains its explanatory power under all conditions. Even in the worst case, the sum of squares of the regression remains constant, and this is generally true whenever another variable is added.

Inexperienced users of regression analysis may seek to integrate as many explaining variables as possible into the model to push up the coefficient of determination. But this contradicts a model's basic goal, which is to explain a phenomenon with as few influencing variables as possible. Moreover, the random

inclusion of additional variables increases the danger that some of them have little to no explanatory power. This is referred to *overparametrization.*

In practice, statisticians frequently calculate what is called the adjusted R^2, which penalizes overparametrization. With every additional variable, the penalization increases. The adjusted coefficient of determination can be calculated by the following equation, where n is the number of observations and k the number of variables in the model (including constants):

$$R^2_{\text{adj}} = R^2 - \frac{\left(1 - R^2\right)(k-1)}{(n-k)} = 1 - \left(1 - R^2\right)\frac{n-1}{n-k} \tag{10.33}$$

It's only worth putting an additional variable in the model if the explanatory power it contributes is larger than the penalization to the adjusted coefficient of determination. When building models, the addition of new variables should stop when the adjusted coefficient of determination can no longer be increased. The adjusted coefficient of determination is suited for comparing regression models with a differing number of regressors and observations.

The penalization invalidates the original interpretation of R^2—the share of y variance that can be explained through the share of x variance. In unfavourable circumstances the adjusted coefficient of determination can even take negative values. For $R^2 = 0$ and $k > 1$ the following equation applies:

$$R^2_{\text{adj}} = 0 - \frac{(1-0)(k-1)}{(n-k)} = \left(-\frac{(k-1)}{(n-k)}\right) < 0 \tag{10.34}$$

10.7 Regression with an Independent Dummy Variable

In our previous discussions of regression both the (dependent) y variables and the (independent) x variables had a metric scale. Even with a least squares regression, the use of other scales is problematic. Indeed, ordinal and nominal variables are, with one small exception, impermissible in a least squares regression. We will now consider this exception.

In the chapter on calculating correlation we found out that so-called dummy variables—nominal variables that possess the values zero and one only—can be understood as *quasi-metric* under certain conditions (see 4.5.1). Their effects on the regression calculation can also be interpreted in the same way. Consider our mail order business example. You guess that red dresses sell better than other dress colours, so you decide for a regression with the independent variables *catalogue image size* (in cm^2) and *red as dress colour* (1: yes; 0: no). The second variable represents the two-value dummy variable: either *red dress* or *no red dress.* Figure 10.11 shows the results of the regression.

All regression coefficients are significantly different from zero (p-values < 0.05). One err in 1.6% of cases when assuming an influence of the colour *red*

	dress_no	sales_t	image_size	red
1	1	245	48.49	0
2	2	197	29.32	0
3	3	202	35.27	0
4	4	194	32.99	0
5	5	288	73.37	1
6	6	239	45.10	0
7	7	263	61.63	0
8	8	264	61.00	0
9	9	247	49.66	0
10	10	220	44.39	0
11	11	225	47.43	0

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.955[a]	.911	.910	7.154

a. Predictors: (Constant), Colour of dress is red (0:not red; 1: red), Catalogue image size

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	51035.513	2	25517.756	498.65	.000[b]
	Residual	4963.877	97	51.174		
	Total	55999.390	99			

a. Dependent Variable: Sales of a similar dress from the previous year

b. Predictors: (Constant), Colour of dress is red (0:not red; 1: red), Catalogue image size

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	142.942	3.874		36.896	.000
	Catalogue image size	1.945	.078	.902	24.820	.000
	Colour of dress is red (0: not red; 1: red)	6.061	2.480	.089	2.444	.016

Dependent Variable: Sales of a similar dress from the previous year

Fig. 10.11 Regression output with dummy variables

on the *sales in the period t and one err in less than 0.5%* of cases when assuming an influence of the *catalogue image size* on the *sales in the period t*.

The regression can be expressed with the following algebraic equation:

$$\hat{y} = 142.9 + 1.95 \cdot \text{catalogue image size} + 6.1 \cdot \text{red} \tag{10.35}$$

On average, sales increase by 1.95 dresses for every additional square centimetre in catalogue image size ($\beta_1 = 1.95$). The sales of red dresses are around six units higher on average than other dress colours ($\beta_2 = 6.1$). Ultimately, the dummy variable shifts parallel to the regression line by the regression coefficient ($\beta_2 = 6.1$) for the observations coded with one (red dress). The slope of the regression line remains unchanged for every dress colour (red or not) with regard to the metric variable (catalogue image size). The only aspect that changes is the location of the regression line. For dummy variables coded with one, the line shifts parallel upward for positive regression coefficients and downward for negative regression coefficients (see Fig. 10.12).

The dummy variables coded with zero serve the benchmark group. It is also conceivable that there is more than one dummy variable. For instance, we could have three variables: red ("dress colour red" [1: yes; 0: no]), green ("dress colour green" [1: yes; 0: no]), and ("dress colour blue" [1: yes; 0: no]). Each of the coefficients yields the deviation for each of the three colours in relation to the remaining dress colours (neither red nor green nor blue). Say we obtain the following regression:

$$\hat{y} = 140 + 1.9 \cdot \text{catalogue image size} + 6 \cdot \text{red} + 5 \cdot \text{green} + 4 \cdot \text{blue} \tag{10.36}$$

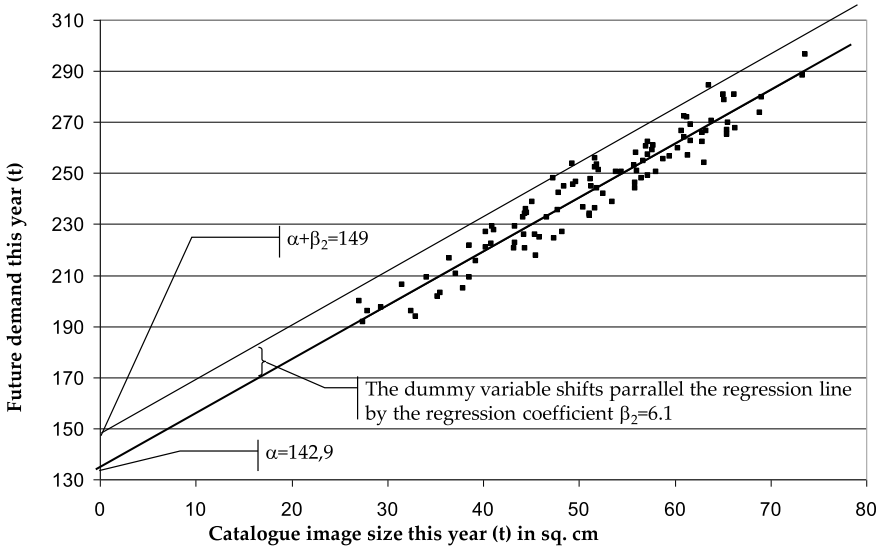


Fig. 10.12 The effects of dummy variables shown graphically

The number of red dresses (6 units) is higher than that of other dress colours that are neither red nor green nor blue. The number of green dresses (5 units) and the number of blue dresses (4 units) also lie above the benchmark.

10.8 Leverage Effects of Data Points

Let's look at the data points for the mail order business shown in Fig. 10.13. Consider the graph's first data point, which represents a dress advertised with 27.1 cm^2 of catalogue space and sold 200 times. Say we keep the catalogue image size the same but reduce the amount sold by 150 units, from 200 to 50. In Fig. 10.13 the new data point is indicated by the left arrow. The change results in a new regression, represented by the dotted line (regression 2). The new slope is 2.4 (versus 2.1) and the value of the constant is 118 (versus 135). The decrease in sales on the left side of the scatterplot creates a corresponding downward shift on the left side of the regression line. We can describe this phenomenon with the beam balance metaphor used previously. The pointer in the middle of the scale—the scatterplot's bivariate centroid—remains fixed, while the "*beam*" tips to the left, as it would under the pressure of a weight. Now let's see what happens when we apply the same change (150 fewer units) to a data point in the centre of the scatterplot. This resulting line—represented by regression 3—has the same slope as that of the original regression, while the value of the constants has dropped slightly, from 135 to 133. Here the reduction has no influence on the marginal effects of the x

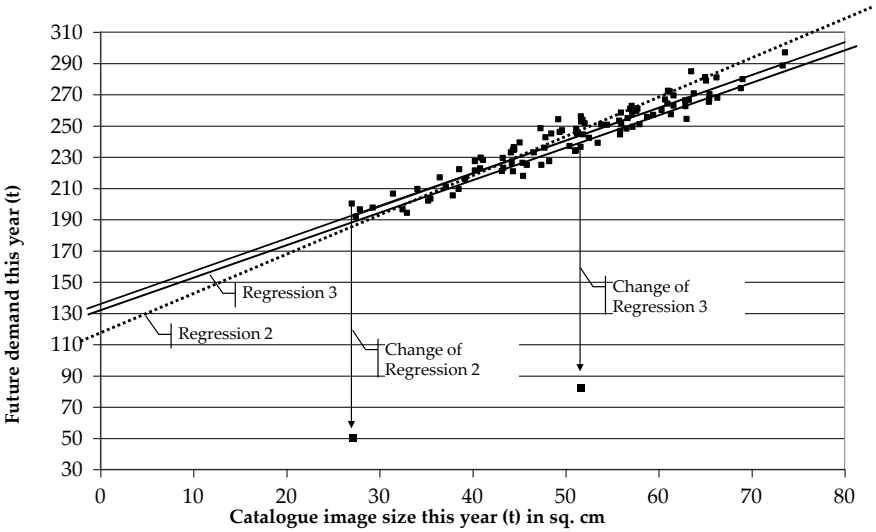


Fig. 10.13 Leverage effect

variables (slope coefficient). It expresses itself only in a parallel downward shift in the regression line.

This graph clearly shows that data points at the outer edges of the scatterplot have greater influence on the slope of the regression line than data points in the centre. This phenomenon is called *leverage.* But since the undesired outliers occupy the outer edges, special attention must be paid to them when creating the regression. It is a good idea to calculate the regression with and without outliers and, from the difference between them, determine the influence of outliers on slope. Should the influence be important, the outliers should be removed or the use of a non-linear function considered (see Sect. 10.9).

10.9 Non-linear Regressions

As we have seen, linear bivariate regressions are just that: straight lines that best fit a set of data points. But can straight lines really capture real-life associations? This is a justified question. Let us consider the meaning of linearity more closely. Linear associations come in two types:

- The first type contains regression coefficients (α, β_1, β_2, …, β_k) that are linear or non-linear. If the regression coefficients for x values remain constant, one speaks of a regression that is linear in its parameters. In this case, we can get by with a single least squares regression. If the regression coefficients change depending on x values, one speaks of a non-linear regression in the parameters. Here separate least squares regressions can be calculated for different segments

of the x-axis. In the example in Fig. 10.6 we have a linear regression in the parameters, as both the constants ($\alpha = 62.22$) and the regression coefficients ($\beta_1 = 1.95$ and $\beta_2 = 0.33$) remain the same over the course of the entire x-axis.

- The second type contains independent x variables that exert a linear or non-linear influence on the dependent y variable while the value of the regression coefficients (α, β_1, β_2, …, β_k) remain constant. In part 4 of Fig. 10.14, for instance, we see a logarithmic association. This regression is non-linear in the variables, also known as a non-linear regression. If the regression coefficients remain constant in Fig. 10.14, a least squares regression can be carried out, although the regression is non-linear.

Using the least squares regression we can also represent non-linear associations: a regression need not be limited to the form of a straight line. Let's look at an example to understand how to approach regressions with variables that have a non-linear association. Figure 10.15 displays monthly sales figures [in €10,000 s] and the number of consultants in 27 store locations. A linear regression calculated from these data produces the following regression line:

$$\hat{y} = 0.0324 \cdot x + 55.945;\ R^2 = 0.66 \tag{10.37}$$

If the number of consultants in a district increases by one, sales increases on average by

$$\Delta\hat{y} = 0.0324 \cdot 1 \cdot [€10{,}000] = €324 \tag{10.38}$$

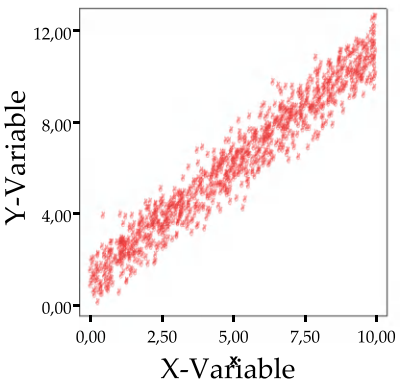
Yet a closer examination reveals that this regression line contains systematic errors. In a district containing between 20 and 100 consultants, the regression line underestimates sales throughout, while in a district with 140 consultants or more, the regression line overestimates sales throughout. The reason: a non-linear association exists between the x and y values, leading to a non-linear regression line.

If we convert the x variable to a logarithmic function—the form of the scatterplot suggests a logarithmic regression—we get the upper scatterplot in Fig. 10.16. Here the x-axis does not track the number of consultants but the logarithmic function of the number of consultants. Now the regression line

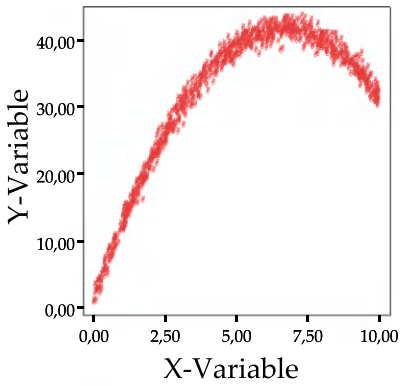
$$\hat{y} = 1.7436 \cdot \ln(x) + 51.61 \tag{10.39}$$

contains *no systematic errors*, as the positive and negative deviations alternate over the course of the regression. What is more, the calculated coefficient of determination increases to $R^2 = 0.97$.

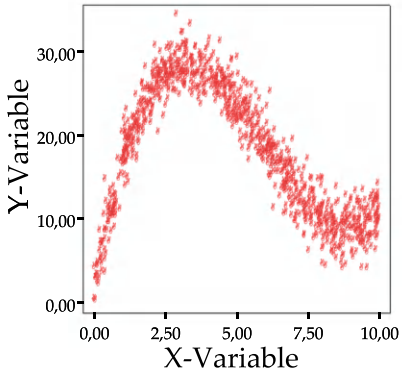
Of course, we can also choose *not* to convert the x-axis scale to a logarithmic function (see the lower scatterplot in Fig. 10.16) and nevertheless enter the logarithmic regression into the scatterplot. This makes the non-linear association between the variables apparent. The algebraic form of the regression function remains the same, as we've changed only the way we represent the functional relationship, not the functional relationship itself ($\hat{y} = 1.7436 \cdot \ln(x) + 51.61$).



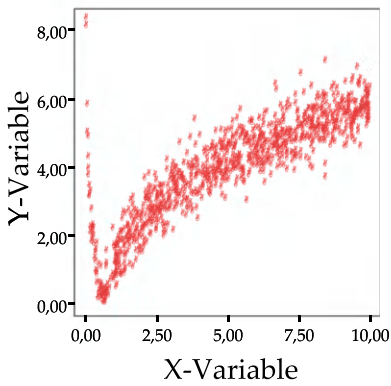
1) Linear in variables and parameters: $y=1+1\cdot x$



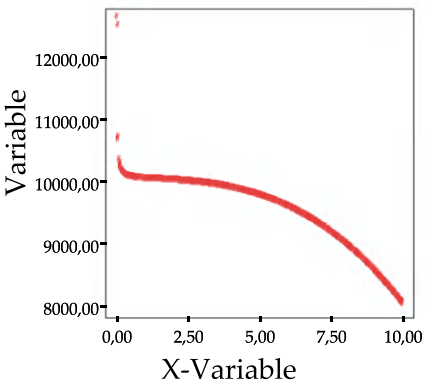
2) Linear in parameters: $y=1+12\cdot x-0.9\cdot x^2$



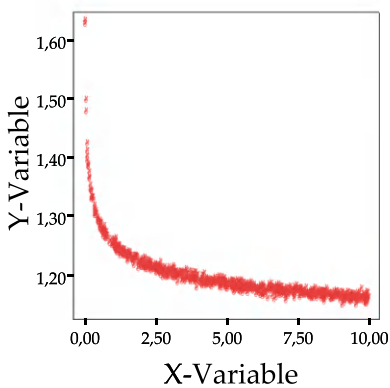
3) Linear in parameters: $y=1+19\cdot x+4\cdot x^2+1\cdot x^3$



4) Linear in parameters: $y=1+2\cdot \ln(x)$

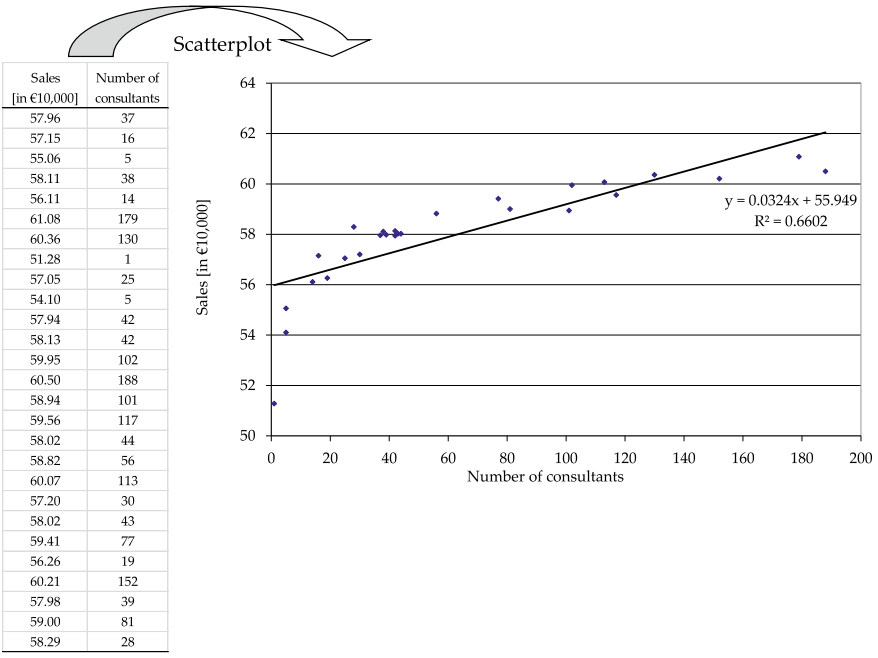


5) Linear in parameters: $y=10{,}000+ (25/x) -2\cdot x^3$



6) Linear in parameters: $y=1 + (0.25/x^2)$

Fig. 10.14 Variables with non-linear distributions



Sales [in €10,000]	Number of consultants
57.96	37
57.15	16
55.06	5
58.11	38
56.11	14
61.08	179
60.36	130
51.28	1
57.05	25
54.10	5
57.94	42
58.13	42
59.95	102
60.50	188
58.94	101
59.56	117
58.02	44
58.82	56
60.07	113
57.20	30
58.02	43
59.41	77
56.26	19
60.21	152
57.98	39
59.00	81
58.29	28

Fig. 10.15 Regression with non-linear variables (1)

10.9.1 Approaches to Regression Diagnostics

In the preceding section we learned how to determine the association between multiple independent variables and a single dependent variable using a regression function. For instance, we discovered that sales for a certain dress could be estimated by the equation

$$\hat{y} = 62.22 + 1.95 \cdot \text{catalogue image size} + 0.33 \cdot \text{previous sales}. \tag{10.40}$$

In addition, we used the adjusted coefficient of determination to find out more about the regression line's goodness of fit and were thus able to say something about the quality of the regression. Proceeding in this vein, we could, for instance, compare the quality of two potential regressions. But how can we identify systematic errors in a regression? To do this, we must once again consider the individual data points using a bivariate regression. Every actual y_i value can be expressed as a combination of the value estimated by the regression $\hat{y}_i$ and the accompanying error term u_i. Since $\hat{y}_i$ represents the outcome of the regression equation from x_i, we get:

$$y_i = \hat{y}_i + u_i = \alpha + \beta \cdot x_i + u_i \tag{10.41}$$

Sales [in €10,000]	Number of consultants
57.96	37
57.15	16
55.06	5
58.11	38
56.11	14
61.08	179
60.36	130
51.28	1
57.05	25
54.10	5
57.94	42
58.13	42
59.95	102
60.50	188
58.94	101
59.56	117
58.02	44
58.82	56
60.07	113
57.20	30
58.02	43
59.41	77
56.26	19
60.21	152
57.98	39
59.00	81
58.29	28

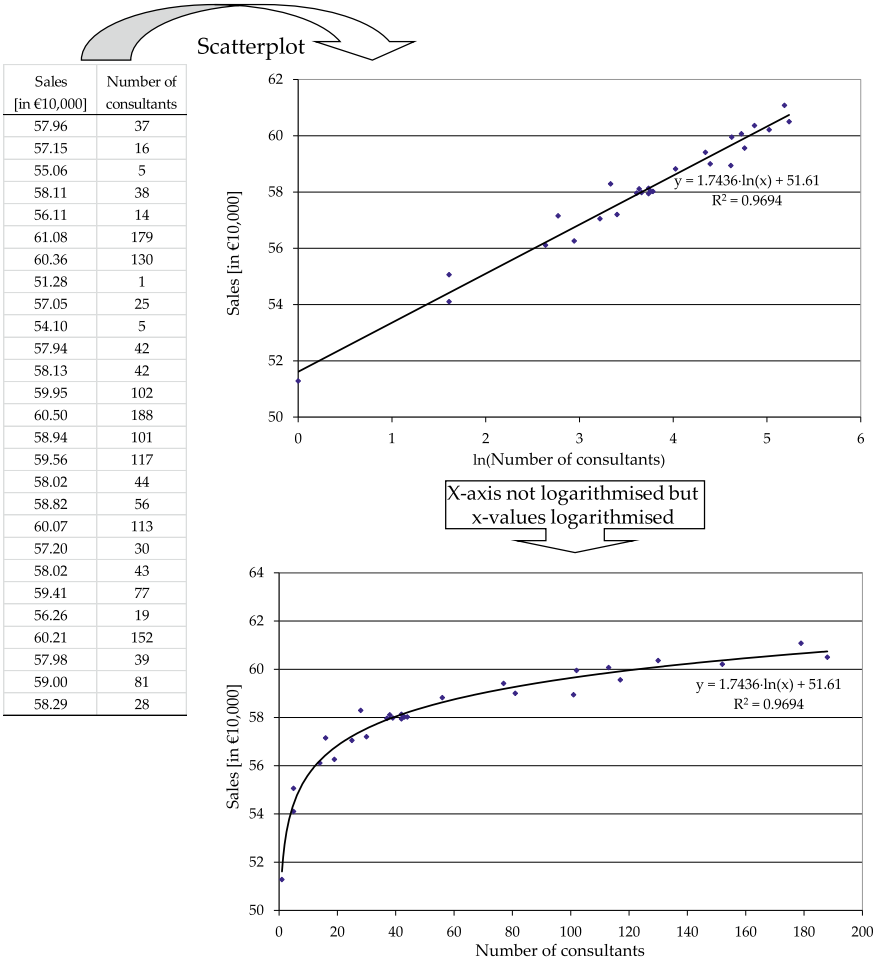


Fig. 10.16 Regression with non-linear variables (2)

To avoid systematic errors in a regression and to estimate its quality we must identify certain conditions for the error term u:

1. Positive and negative values should cancel each other out. The condition is automatically fulfilled in the regression calculation.
2. The regression's independent variables (x variables) should not correlate with the error term. The case described in Fig. 10.15—where deviations u only appear in a certain direction along the x-axis (e.g. above the line)—should not occur. This would mean that y values are being systematically over- or underestimated. A solution to this problem is proposed below.

3. The demand that error terms should not correlate is a similar criterion:

$$\mathrm{Cov}(u_i; u_j) = 0\, i \neq j \tag{10.42}$$

This is called the condition of *missing autocorrelation.* It says there should be no systematic association between the error terms. In our mail order business, an autocorrelation occurs when mostly positive deviations obtain with image sizes of 40 cm^2 or smaller and with image sizes 60 cm^2 or larger, and mostly negative deviations obtain with image sizes between 40 and 60 cm^2. Figure 10.17 displays three possible correlations with autocorrelated error terms. For obvious reasons, systematic errors are undesirable in terms of methods as well as outcomes. Generally, the autocorrelation can be traced back to an error in the model specification and thus requires us to reconsider our choice of models. We can do this by transforming into non-linear functional regressions (as with non-proportional increases) or by adding a *missing variable* (i.e. considering a neglected influence).

4. The variance for every u_i should be constant: $\mathrm{Var}(u_i) = \sigma^2$. This condition is referred to as variance homogeneity, or *homoscedasticity* (*homo* means *the same*; *scedasticity* means *variance*). If this condition is not fulfilled, one speaks of variance heterogeneity or heteroscedasticity. In this case, the coefficients of an OLS estimation remain unbiased and consistent. However, they are no longer efficient. This means that other estimators may have lower variance and therefore provide better estimates. The standard errors of the coefficients are no longer reliable. As a result, hypothesis testing can be misleading and may result in incorrect conclusions about the significance of the predictors. This occurs when the data points are distributed at different concentrations over the x-axis. Frequently, these "eruptions" of data points are caused by a missing variable in the model. Figure 10.18 provides examples of the undesirable effect. Here

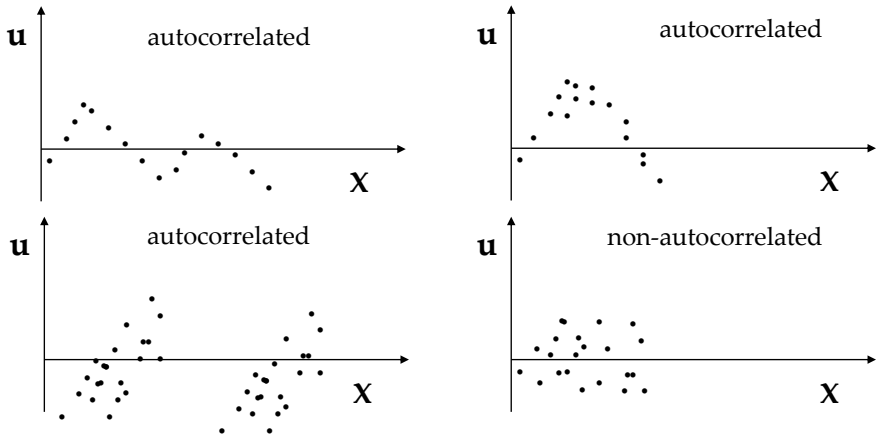


Fig. 10.17 Autocorrelated and non-autocorrelated distributions of error terms

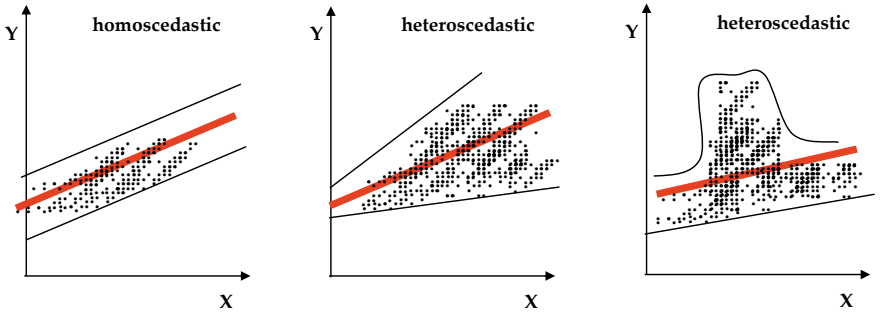


Fig. 10.18 Homoscedasticity and heteroscedasticity

too, the model must be checked for error specification (missing variables or an erroneous selection of the functional distribution). Another way to correct for bias in standard errors is to use robust standard errors (e.g. White's (1980) heteroscedasticity-consistent standard errors), which provide heteroscedasticity-consistent standard errors.

We can examine the quality conditions for the error term u with a graphical analysis (see for instance Figs. 10.17 and 10.18). But this approach does not always suffice. In practice, statisticians use test methods from inductive statistics, but a discussion of these methods lies beyond the scope of this chapter.

5. With regressions that have more than one independent x variable, the independent x variables should not have an association. If the association between two or more x variables is too large, so-called *multicollinearity* occurs, which falsifies the regression outcome.

Ultimately, this condition entails nothing more than choosing two variables for the predictor x variables whose meaning is different or at least dissimilar. If we estimate the market share for petrol using gross and net prices from the SPSS file *Chapter10_multicollinearity_petrol_example.sav* (see supplementary material at https://doi.org/10.1007/978-981-96-2524-6_10), we get the output displayed in Fig. 10.19.

SPSS is unable to calculate the influence of the gross and net price at the same time. The reason is that gross price can be derived directly from the net price plus value added tax. The variables are thus linearly dependent. With a value added tax of 19%, we arrive at the following association:

$$\text{net price} = \text{gross price} \big/ 1.19 \tag{10.43}$$

The regression

$$\hat{y} = \beta_o + \beta_1 \cdot \text{net price} + \beta_2 \cdot \text{gross price} \tag{10.44}$$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.442	.201		7.171	.000
	Net price of own product (SPARAL high-octane petrol)	-.871	.167	-.723	-5.229	.000

a. Dependent Variable: Market share for high-octane petrol

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	Gross price of own product (SPARAL high-octane petrol)	.[b]	.	.	.	.000

a. Dependent Variable: Market share for high-octane petrol
b. Predictors in the model: (Constant), Net price of own product (SPARAL high-octane petrol)

Fig. 10.19 Solution for perfect multicollinearity

can be converted into:

$$\hat{y} = \beta_o + \left(\frac{\beta_1}{1.19} + \beta_2\right) \cdot \text{gross price} \Leftrightarrow \hat{y} = \alpha + \beta \cdot \text{gross price} \qquad (10.45)$$

It would have been necessary to calculate the two regression coefficients β_1 and β_2, although there is only one linearly independent variable (gross or net). If perfect multicollinearity exists, it is impossible to determine certain regression coefficients.[3] For this reason, most computer programmes remove one of the variables from the model. This makes senses both from the perspective of methods and that of outcomes. What additional explanatory value could we expect from a net price if the model already contains the gross price?

But perfect multicollinearity rarely occurs in practice; it is almost always *high but not perfect*. So when we speak of multicollinearity we really mean *imperfect multicollinearity*. It is not a question of whether multicollinearity exists or not. It is question of the strength of the association of independent x variables. Why is imperfect multicollinearity a problem for determining the regression?

Consider the case where we use the company's price and a competitor's price for estimating petrol market share. From Sect. 4.7.1 we know that while the correlation between the prices is not perfect it is still quite high: $r = 0.902$. Imperfect multicollinearity often causes the following effects:

[3] In Sect. 10.3 we calculated the regression coefficients $= \beta(\alpha = \beta_0; \beta_1; \ldots; \beta_k)$ as follows: $\beta = (X'X)^{-1}X'y$. The invertibility of $(X'X)$ assumes that matrix X displays a full rank. In the case of perfect multicollinearity, at least two rows of the matrix are linearly dependent so $(X'X)$ can no longer be inverted.

Coefficients[a]

Model	Unstandardized Coefficients		Stand. Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	1.446	.206		7.02	.000		
Net price of own product (SPARAL high-octane petrol)	-.799	.393	-.663	-2.04	.053	.187	5.348
Competitor price (JETY high-octane petrol)	-.065	.319	-.066	-.20	.841	.187	5.348

a. Dependent Variable: Market share for high-octane petrol

Fig. 10.20 Solution for imperfect multicollinearity

- If the competitor's price is omitted in the regression, the coefficient of determination drops 0.001 to $R^2 = 0.522$. The additional influence of the competitor's price appears to have only a slight effect. But if we use only the competitor's price as the predictor variable for sales in the regression, the explanatory power turns out to be $R^2 = 0.44$, which is quite high. This is a sign of multicollinearity, as the company's price and the competitor's price appear to behave similarly when explaining market share trends.
- The algebraic sign of the regressor is unusual. The competitor's price appears to have the same direction of effect on market share as the company's own price, i.e. the higher the competitor's price, the lower the market share (see Fig. 10.20).
- Removing or adding an observation from the dataset leads to large changes in the regression coefficients. In the case of multicollinearity, the regression coefficients strongly react to the smallest changes in the dataset. For instance, if we remove observation 27 from the dataset *Chapter10_multicollinearity_petrol_example* (see supplementary material at https://doi.org/10.1007/978-981-96-2524-6_10) and calculate the regression anew, the influence of the company's price sinks from $\beta_1 = -0.799$ to $\beta_1 = -0.559$, or by more than 30%.
- So-called *Variance Inflation Factors (VIF)* can indicate yet another sign of multicollinearity. For every independent x variable, we must check the association with the other independent x variables of the regression. To do this we perform a so-called *auxiliary regression* for every independent variable. If there are five independent x variables in a regression, we must carry out five auxiliary regressions. With the first auxiliary regression, the initial independent x variable (x_1) is defined as dependent and the rest (x_2 to x_5) as independent. The creates the following regression:

$$x_1 = \alpha_o + \alpha_1 \cdot x_2 + \alpha_2 \cdot x_3 + \alpha_3 \cdot x_4 + \alpha_4 \cdot x_5 \tag{10.46}$$

The larger the coefficient of determination $R^2_{\text{Aux}(1)}$ for this auxiliary regression, the stronger the undesired association between the independent variable x_1 and the

other independent variables of the regression equation. Remember: multicollinearity exists when two or more independent x variables correlate. Accordingly, the degree of multicollinearity can also be expressed by the $R^2_{Aux(i)}$ of the auxiliary regression of the ith independent variable. VIF builds on the idea of auxiliary regression. Every independent x variable receives the quotient

$$\text{VIF}_i = \frac{1}{1 - R^2_{\text{Aux}(i)}} \tag{10.47}$$

If the $R^2_{\text{Aux}(i)}$ value of the auxiliary regression of an independent variable is (close to) zero, no multicollinearity exists and VIF = 1. If, by contrast, the $R^2_{\text{Aux}(i)}$ of an auxiliary regression is very large, multicollinearity exists and the value of VIF is high. Hair et al. (2006, p. 230) note that VIF = 10 is a frequently used upper limit but recommend a more restrictive value for smaller samples. Ultimately, every researcher must make his or her own decision about the acceptable degree of multicollinearity and, when the VIF is conspicuously high, check the robustness of the results. Keep in mind, though, that a VIF as low as 5.3 already has a very high multiple correlation, namely, $r = 0.9$. For this reason, whenever the VIF is 2.0 or higher—VIF = 2.0 translates into a multiple correlation of $r = 0.71$—you should test your results, checking to see how they respond to minor changes in the sample.

- Some statistic software programmes indicate *Tolerance* as well as VIF, with $\text{Tolerance}(i) = \left(1 - R^2_{\text{Aux}(i)}\right)$. When the value of Tolerance is (close to) one, then multicollinearity does not exist. The more the value of Tolerance approaches zero, the larger the multicollinearity. In Fig. 10.20 the VIFs and the Tolerances of the dataset *Chapter10_multicollinearity_petrol_example.sav* (see supplementary material at https://doi.org/10.1007/978-981-96-2524-6_10) are indicated on the right edge of the table. Both metrics clearly indicate multicollinearity.

As we have seen, multicollinearity has undesirable effects. Influences should not only have the correct algebraic sign. They must remain stable when there are small changes in the dataset. The following measures can be taken to eliminate multicollinearity:

- Remove one of the correlating variables from the regression. The best variable to remove is the one with the highest VIF. From there proceed in steps. Every variable you remove lowers the VIF values of the regression's remaining variables.
- Check the sample size. A small sample might produce multicollinearity even if the variables are not multicollinear throughout the population. If you suspect this could be the case, include additional observations in the sample.
- Reconsider the theoretical assumptions of the model. In particular, ask whether your regression model is overparameterized.

- Not infrequently, correlating variables can be combined into a single variable with the aid of factor analysis (see Chap. 14).

10.9.2 Chapter Exercises

Exercise 73
You're an employee in the market research department of a coffee roasting company who is given the job of identifying the euro price of the company's coffee in various markets and the associated market share. You discover that market share ranges between 0.20 and 0.55. Based on these findings you try to estimate the influence of price on market share using the regression indicated below.

Regression function: market share $\hat{y} = 1.26 - 0.298 \cdot \text{price}$

(a) What average market share can be expected when the coffee price is three euros?
(b) You want to increase the market share to 40%. At what average price do you need to set your coffee to achieve this aim?
(c) The regression yields an R^2 of 0.42. What does this parameter tell us?
(d) How large is the total sum of squares when the error sum of squares of the regression is 0.08?

Exercise 74
You have a hunch that the product sales mentioned in Exercise 8 are not determined by price alone. So you perform a multiple regression using Excel (or statistics software like SPSS). The results of the regression are listed in Fig. 10.21 (see also the dataset *Chapter10_ Sales and Dealership and Price*; supplementary material at https://doi.org/10.1007/978-981-96-2524-6_10).

(a) Derive the regression function in algebraic form from the data in the table.
(b) Does the model serve to explain sales? Which metric plays a role in the explanation and what is its value?
(c) Assume you lower the price in every country by 1000 monetary units. How many more products would you sell?
(d) What is the effect of increasing advertising costs by 100,000 monetary units? Explain the result and propose measures for improving the estimating equation.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.975[a]	.951	.927	.510

a. Predictors: (Constant), Advertising budget [in 100,000s MUs], Number of dealerships, Unit price [in 1,000s of MUs]

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	30.439	3	10.146	39.008	.000[b]
	Residual	1.561	6	0.260		
	Total	32.000	9			

a. Dependent Variable: Sales [in 1,000s of units]

b. Predictors: (Constant), Advertising budget [in 100,000s MUs], Number of dealerships, Unit price [in 1,000s of MUs]

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	(Constant)	24.346	3.107	7.84	.000
	Number of dealerships	.253	.101	2.50	.047
	Unit price [1,000s of MUs]	-.647	.080	-8.05	.000
	Advertising budget [in 100,000s MUs]	-.005	.023	-0.24	.817

a) Derive the regression function in algebraic form from the data in the table.

b) Does the model serve to explain sales? Which metric plays a role in the explanation and what is its value?

c) Assume you lower the price in every country by 1,000 monetary units. How many more products would you sell?

d) What is the effect of increasing advertising costs by 100,000 monetary units? Explain the result and propose measures for improving the estimating equation.

Fig. 10.21 Regression results (1)

Exercise 75

You're given the job of identifying the market share of a product in various markets. You determine the market share ranges between 51.28 and 61.08%. You try to estimate the factors influencing market share using the regression in Fig. 10.22:

(a) Derive the regression function in algebraic form from the above table.
(b) Determine R^2 and the adjusted R^2.
(c) How large is the residual sum of squares?
(d) Does the model have an explanatory value for determining market share?
(e) What's a reasonable way to improve the model?
(f) What happens when the product price is raised by one monetary unit?

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	???	???	???	.652

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	124.265	2	???	145.971	.000
	Residual	???	24	???		
	Total	134.481	26			

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	(Constant)	38.172	1.222	31.24	.000
	price	.-7.171	.571	-12.56	.000
	ln(price)	.141	.670	-0.21	.835

a) Derive the regression function in algebraic form from the above table.

b) Determine R^2 and the adjusted R^2.

c) How large is the residual sum of squares?

d) Does the model have an explanatory value for determining market share?

e) What's a reasonable way to improve the model?

f) What happens when the product price is raised by one monetary unit?

Fig. 10.22 Regression results (2)

Exercise 76

You're an employee in the market research department of a company that manufactures oral hygiene products. You're given the task of determining weekly sales of the toothpaste *Senso White* at a specific drugstore chain over the past three years. You attempt to estimate the factors influencing weekly market share using the regression in Fig. 10.23. The potential factors include:

- the price of Senso White (in €),
- Senso White advertised with leaflets by the drugstore chain (0 = no; 1 = yes),
- other toothpaste brands advertised with leaflets by the drugstore chain (0 = no; 1 = yes),
- other toothpaste brands advertised in daily newspapers by the drugstore chain (0 = no; 1 = yes),
- Senso White advertised in daily newspapers by the drugstore chain (0 = no; 1 = yes),

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.883	.780	.771	187.632

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	18627504.189	6	???	84.000	.000
	Residual	5245649.061	149	???		
	Total	13423873153.250	155			

Model		Unstandardized Coefficients		Standardized Coefficients BETA	Sig.
		B	Std. Error		
1	(Constant)	9897.875	146.52		.000
	Price of Senso White	-949.518	59.094	-.64	.000
	Senso White advertised with leaflets	338.607	188.776	.19	.075
	Other toothpaste brands advertised with leaflets	-501.432	74.345	-.27	.000
	Other toothpaste brands advertised in daily newspapers	-404.053	87.042	-.18	.000
	Senso White advertised in daily newspapers	245.758	73.186	.13	.001
	Senso White advertised with leaflets that contain images	286.195	202.491	.15	.160

a) Derive the regression equation in algebraic form from

b) What sales can be expected with a toothpaste price of €2.50 when the drugstore chain uses no advertising for Senso White and uses leaflets for a competing toothpaste?

c) Interpret R, R^2, and adjusted R^2. Explain the purpose of the adjusted R^2.

d) What is the beta needed for?

e) Assume you want to improve the model by introducing a price threshold effect to account for sales starting with €2.50. What is the scale of the price threshold effect? Which values should be used to code this variable in the regression?

Fig. 10.23 Regression toothpaste

- Senso White advertised with leaflets that contain images by the drugstore chain (0 = no; 1 = yes).

(a) Derive the regression equation in algebraic form from Fig. 10.23.
(b) What sales can be expected with a toothpaste price of €2.50 when the drugstore chain uses no advertising for Senso White and uses leaflets for a competing toothpaste?
(c) Interpret R, R^2, and adjusted R^2. Explain the purpose of the adjusted R^2.

(d) What is the beta needed for?
(e) Assume you want to improve the model by introducing a price threshold effect to account for sales starting with €2.50. What is the scale of the price threshold effect? Which values should be used to code this variable in the regression?

Exercise 77

The fast food chain Burger Slim wants to introduce a new children's meal. The company decides to test different meals at its 2261 franchises for their effect on total revenue. Each meal variation contains a slim burger and, depending on the franchise, some combination of soft drink (between 0.1 and 1.0 L), salad, ice cream, and a toy. These are the variables:

- Revenue: Revenue through meal sales in the franchise [in MUs]
- Salad: Salad = 1 (salad); Salad = 0 (no salad)
- Ice Cream: Ice Cream = 1 (ice cream); Ice Cream = 0 (no ice cream)
- Toy: Toy = 1 (toy); Toy = 0 (no toy)
- Sz_Drink: Size of soft drink
- Price: Price of meal.

You perform two regressions with the results in Fig. 10.24 (see *Chapter10_burger* dataset; see supplementary material at https://doi.org/10.1007/978-981-96-2524-6_10):

(a) Calculate R^2.
(b) What is the adjusted R^2 needed for?
(c) Using regression, determine the average revenue generated by a meal that costs five euros and contains a slim burger, a 0.5 L soft drink, a salad, and a toy.
(d) Using regression two determine which variable has the largest influence. Explain your answer.
(e) Compare the results of regressions one and two. Which of the solutions would you consider in a presentation for the client?
(f) Consider the scatterplot in Fig. 10.25. What's the problem? Describe the effects of the results from regressions one and two on interpretability. How can the problem be eliminated?

10.9.3 Exercise Solutions

Solution 73

(a) Market share $= 1.26 - 0.298 \cdot \text{Price} = 1.26 - 0.298 \cdot 3 = 36.6\%$.
(b) $0.40 = 1.26 - 0.298 \cdot \text{price} \Leftrightarrow = \frac{0.40-1.26}{-0.298} = €2.89$

Regression 1:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	???	???	.747	3911.430

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	???	4	???	1668.726	.000
	Residual	34515190843.303	2256	???		
	Total	136636463021.389	2260			

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	BETA		
1	(Constant)	25949.520	265.745		97.648	.000
	Price	4032.796	73.255	.58	55.051	.000
	Salad	-7611.182	164.631	-.49	-46.232	.000
	Ice Cream	3708.259	214.788	.18	17.265	.000
	Toy	6079.439	168.553	.38	36.068	.000

Regression 2:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.866	.750	.750	3891.403

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	102488948863.420	5	???	1353.613	.000
	Residual	34147514157.969	2255	???		
	Total	136636463021.389	2260			

Model		Unstandardized Coefficients		Standardized Coefficients	Sig.	Tolerance	VIF
		B	Std. Error	BETA			
1	(Constant)	25850.762	265.143		.000		
	Price	-30.079	827.745	-.004	.971	.008	129.174
	Sz_Drink	24583.927	4989.129	.590	.000	.008	129.174
	Salad	7619.569	163.797	-.490	.000	.999	1.001
	Ice Cream	3679.932	213.765	.182	.000	.997	1.003
	Toy	6073.666	167.694	.382	.000	.999	1.001

a) Calculate R^2.

b) What is the adjusted R^2 needed for?

c) Using regression, determine the average revenue generated by a meal that costs five euros and contains a slim burger, a 0.5 litre soft drink, a salad, and a toy.

d) Using regression two determine which variable has the largest influence. Explain your answer.

e) Compare the results of regressions one and two. Which of the solutions would you consider in a presentation for the client?

f) Consider the scatterplot in Figure 10-25. What's the problem? Describe the effects of the results from regressions one and two on interpretability. How can the problem be eliminated?

Fig. 10.24 Regression results burger slim

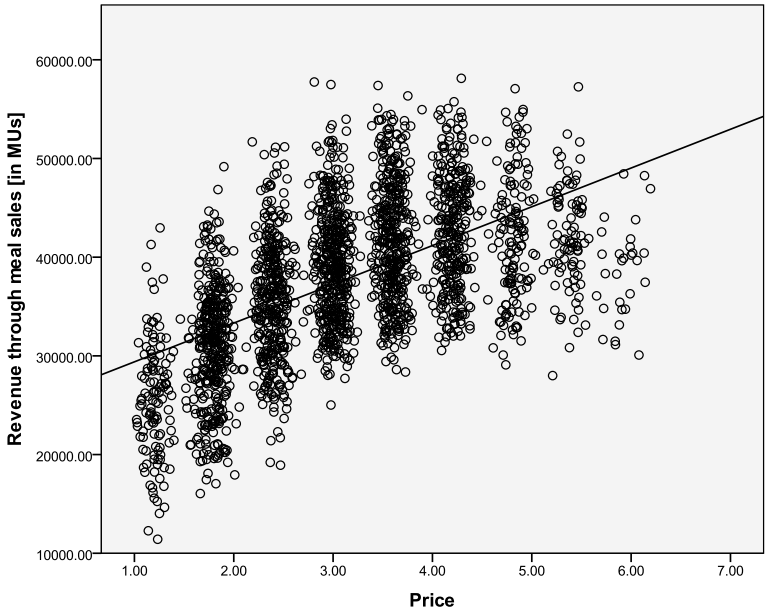


Fig. 10.25 Scatterplot

(c) 42% of the variance in market share is explained by variance in the independent variable price.
d) $R^2 = 1 - \frac{\text{ESS}}{\text{TSS}} \Leftrightarrow \text{TSS} = \frac{\text{ESS}}{1-R^2} = \frac{0.08}{0.58} = 0.14$.

Solution 74

(a) $\hat{y} = 24.346 + 0.253 \cdot x_1 - 0.647 \cdot x_2 - 0.005 \cdot x_3$, where:

x_1 number of locations;
x_2 item price [in 1000s of MUs];
x_3 advertising budget [in 100,000s of MUs].

The p-value of the F-Test is $p = 0.000$. There is at least one variable with a significant influence. These are the variables *number of dealerships* ($p = 0.047 < 0.05$) and unit price ($p = 0.000 < 0.05$). The insignificant influence of advertising budget ($p = 0.817 > 0.05$) would, in practice, eliminate the variable x_3 from the regression (see part d) of the exercise, yielding the following result: $\hat{y} = 24.346 + 0.253 \cdot x_1 - 0.647 \cdot x_2$.

(b) The p-value of the F-Test is $p = 0.000$. There is at least one variable with a significant influence. We already know the coefficient of determination: $R^2 = 0.951$.
(c) The regression coefficient for the item price is $\alpha_2 = -0.647$. Since the item price is measured by 1000 s of units, a price decrease of 1000 MUs affects sales as follows: $\Delta\text{sales} = (-1) \cdot (-0.647) = 0.647$. Sales is also measured by 1000 s of units, which means that total sales increase by $1000 \cdot 0.647 = 647$ units.
(d) The regression coefficient for advertising expenses is $\alpha_3 = -0.005$. Since the advertising expenses are measured by 100,000 s of MUs, an increase of advertising expenses by 100,000 MUs affects sales as follows: $\Delta\text{sales} = (+1) \cdot (-0.005) = (-0.005)$. Sales are measured in 1000 s of units, which means they will sink by $1000 \cdot (-0.005) = (-5)$. The result arises because the variable *advertising budget* is an insignificant influence (close to zero); advertising appears to play no role in determining sales.

Solution 75

(a) $\hat{y} = 38.172 - 7.171 \cdot x_1 + 0.141 \cdot x_2$, where:
x_1 price of the company's product;
x_2 price of the competition's product put through the logarithmic function.
(b) $R^2 = \frac{\text{Explained Sum of Squares (RSS)}}{\text{Total Sum of Squares (TSS)}} = \frac{124.265}{134.481} = 0.924$; $R^2_{\text{adj}} = 1 - \left(1 - R^2\right)\frac{n-1}{n-k} = 1 - (1 - 0.924)\frac{27-1}{27-3} = 0.918$
(c) $\text{RSS} + \text{ESS} = \text{TSS} \Leftrightarrow \text{ESS} = \text{TSS} - \text{RSS} = 10.216$
(d) Yes, because R^2 has a very high value.
(e) By eliminating the price subjected to the logarithmic function (see exercise section a)).
(f) The regression coefficient for the price is $\alpha_1 = -7.171$. This means sales would sink by $(+1) \cdot (-7.171) = -7.171$ percentage points.

Solution 76

(a) $\hat{y} = 9898 - 949.5 \cdot \text{price} + 338.6 \cdot \text{LL}_{\text{sw}} - 501.4 \cdot \text{LL}_{\text{OT}} - 404.1 \cdot \text{NP}_{\text{OT}} + 245.8 \cdot \text{NP}_{\text{sw}} + 286.2 \cdot \text{LL}_{\text{SW_image}}$
(b) $\hat{y} = 9898 - 949.5 \cdot 2.5 + 338.6 \cdot 0 - 501.4 - 1 - 404.1 - 0 + 245.8 \cdot 0 + 286.2 - 0 \approx 7023$
(c) R equals the correlation coefficient; R^2 is the model's coefficient of determination and expresses the percentage of variance in sales explained by variance in the independent variables (right side of the regression function). When creating the model, a high variance explanation needs to be secured with as few variables as possible. The value for R^2 will stay the same even if more independent variables are added. The adjusted R^2 is used to prevent an excessive

number of independent variables. It is a coefficient of determination corrected by the number of regressors.

(d) Beta indicates the influence of standardized variables. Standardization is used to make the independent variables independent from the applied unit of measure, and thus commensurable. The standardized beta coefficients that arise in the regression thus have commensurable sizes. Accordingly, the variable with the largest coefficient has the largest influence.

(e) Create a new metric variable with the name *Price_low*. The following conditions apply: Price_low = 0 (when the price is smaller than €2.50); otherwise Price_low = Price. Another possibility: create a new variable with the name *Price_low*. The following conditions apply here: Price_low = 0 (when the price is less than €2.50); otherwise Price_low = 1.

Solution 77

(a) $R^2 = \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{34{,}515{,}190{,}843.303}{136{,}636{,}463{,}021.389} = 0.7474$

(b) In order to compare regressions with varying numbers of independent variables.

(c) Average proceeds $= 25{,}949.5 + 5 \cdot 4032.79 - 7611.182 + 6079.44 =$ 44581.752 MU

(d) Lettuce, because the standardized beta value has the second largest value.

(e) The price and size of the beverage in regression two show a high VIF value, i.e. a low tolerance. In addition, the R^2 of regression one to regression two has barely increased. The independent variables in regression two are multicollinear, distorting significances and coefficients. The decision impinges on regression one.

(f) Non-linear association exists. As a result, systematic errors occur in certain areas of the *x*-axis in the linear regression. The residuals are autocorrelated. The systematic distortion can be eliminated by using a logarithmic function or by inserting a quadratic term.

References

Hair, J., et al. (2006). *Multivariate data analysis* (6th ed.). Prentice Hall International.

Swoboda, H. (1971). *Exakte Geheimnisse: Knaurs Buch der modernen Statistik*. Knaur.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48*(4), 817–838.

Logistic Regression

11

After the Ordinary Least Squares (OLS) regression has been discussed in detail in the previous Chap. 10, this section is devoted to the topic of the so-called logit, binary, or logistic regression. Logistic regression and OLS regression are both used to analyse the relationship between a dependent variable and one or more independent variables. However, they differ in one important aspect:

Whereas OLS regression is used when the dependent variable is continuously scaled and a linear relationship between the dependent and independent variables needs to be analysed, logistic regression is applied when the relationship is non-linear and the dependent variable is binary, i.e. has only two possible outcomes (e.g. yes/no, 0/1). And indeed, there is a wide range of questions in business and economic research that involve analysing or predicting events with only two possible outcomes, e.g.: which influencing factors significantly affect the probability…

- that a customer will or will not buy a product,
- that a customer will or will not react to an advertising campaign,
- that a bank customer becomes financially insolvent,
- that an employee will leave the company, or
- that a company will go through a merger or acquisition.

It is therefore necessary to be able to predict the probability of a dichotomous dependent variable using the concept of logistic regression. However, it is not only

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78070-7_11.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*,
Springer Texts in Business and Economics,
https://doi.org/10.1007/978-3-031-78070-7_11

a matter of estimating an a priori probability, but also of identifying the factors that significantly influence that probability.

The parameters of a logistic regression are usually estimated using the so-called maximum likelihood method. For binary outcomes, this method gives a much better fit of the model to the data than OLS regression. It is therefore worth learning and understanding this concept. Subsequently, the theoretical foundations of logistic regression are illustrated and explained step by step using the example of the space shuttle Challenger disaster in 1986. We will answer the question of the a priori probability of a successful space shuttle mission given the local weather conditions on the day of the shuttle launch. All the required data can be found in the *Chapter11_Challenger* dataset (see supplementary material at https://doi.org/10.1007/978-3-031-78070-7_11).

The Challenger disaster was a tragic event in the history of space exploration that occurred on 28th January 1986. The space shuttle Challenger was part of the National Aeronautics and Space Administration's (NASA) space programme and was supposed to carry out a routine mission, with the notable inclusion of school teacher Christa McAuliffe as the first civilian in space. On that day, the Challenger exploded shortly after lift-off, resulting in the loss of seven astronauts.

The cause of the disaster was a failure of the so-called O-rings at very cold temperatures, which connected the shuttle's booster rockets to each other. As the Challenger ascended, the O-rings, due to the cold weather, were not flexible enough to provide adequate sealing. Hot gases leaked through the compromised seals, ultimately causing the shuttle's fuel tank to explode.

On the eve of the disaster, a three-hour conference call with representatives from Morton Thiokol (manufacturer of the solid rocket motor), the Marshall Space Flight Center (NASA centre for motor design control), and the Kennedy Space Center ignited a lively debate about the risk of a possible failure of the O-rings at low temperatures. The weather forecast predicted 31 °F. Significant concerns about safety were raised by some participants pointing to anomalies of the O-rings that had occurred during previous space shuttle flights. They expressed their concern that the rubber material of the O-rings would lose its elasticity due to the cold weather conditions and argued for a flight deferral. Unfortunately, only for the seven flights where anomalies occurred during previous shuttle flights, the relationship between the number of defective O-rings (y-axis) and air temperature (x-axis) was plotted as shown in Fig. 11.1. The relationship with the incomplete set of data suggests that damage to O-rings occurred not only on cold days, but also on days with higher temperatures. The Pearson correlation showed a non-significant linear relationship of $r = 0.024$ ($p = 0.96$) and both Kendall's tau-B and the Spearman correlation showed no monotonic relationship at all ($r < 0.000$). For this reason, the management decided to grant the launch permit. From this point of view, the decision to launch did not appear to be wrong.

However, the 16 flights with no anomalies were not considered at all. This may have been the first crucial error, as the linear and non-linear relationships between the number of anomalies and the outside temperature could have been observed in a much better way (see Fig. 11.2). A complete consideration of all data would

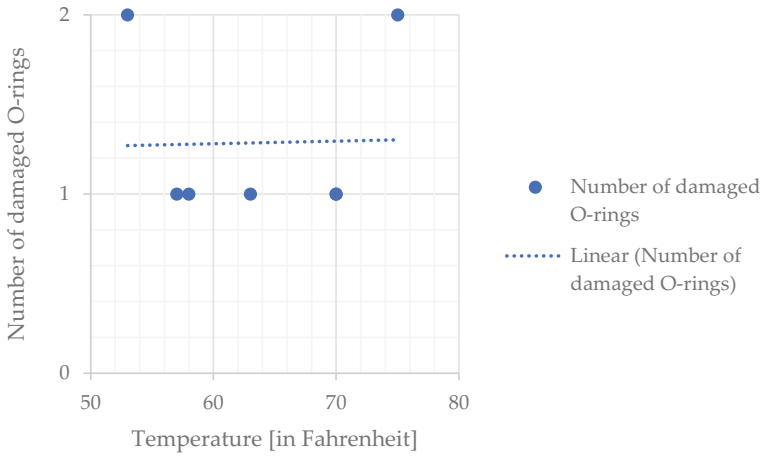


Fig. 11.1 Number of damaged O-rings for flights where anomalies occurred during previous shuttle flights. *Source* Own graphic, data from Dalal et al. (1989, p. 949)

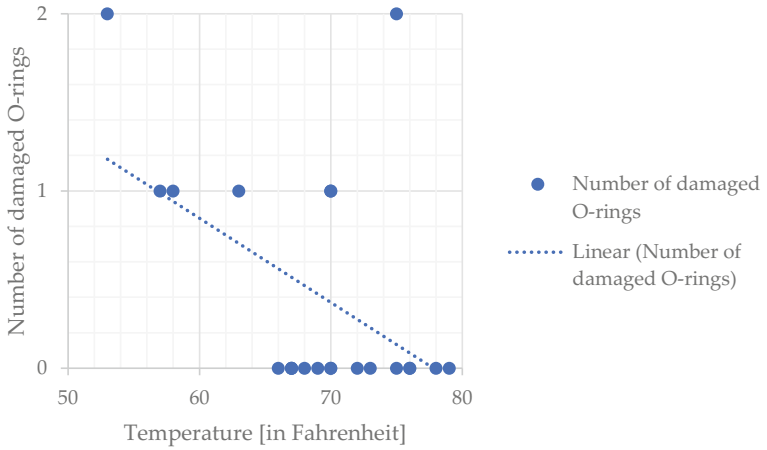


Fig. 11.2 Number of damaged O-rings during previous shuttle flights. *Source* Own graphic, data from Dalal et al. (1989, p. 949)

have shown a significant linear correlation ($r = (-0.511)$; $p = 0.013$), a strong monotonic correlation ($\tau - B = 0.367$; $p = 0.037$) as well as that the O-rings did not work properly during any launch below 66 °F. The opponents of a launch had correctly identified the risk of a launch. However, they provided incomplete data.

Following the accident, the Rogers Commission appointed by President Ronald Reagan concluded that a "careful analysis of the flight history of O-ring performance would have revealed the correlation of O-ring damage in low temperature. Neither NASA nor Thiokol carried out such an analysis; consequently, they were

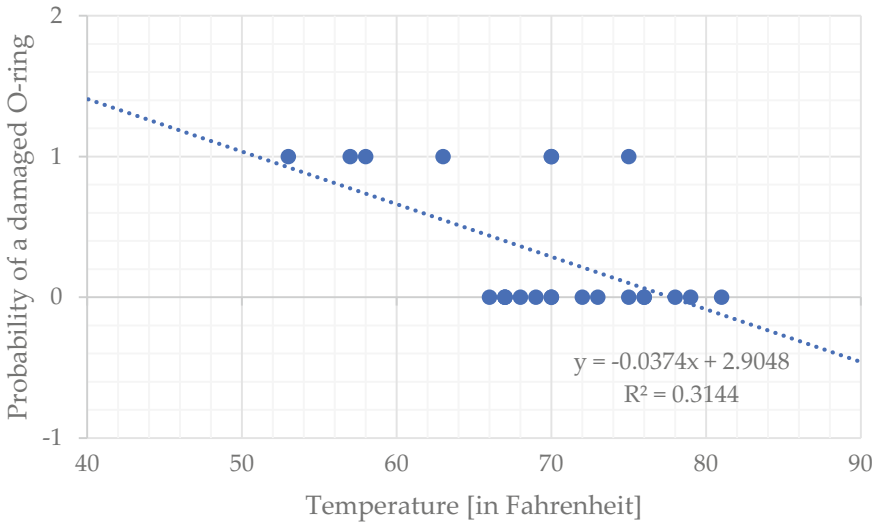


Fig. 11.3 Probability of a damaged O-ring during previous shuttle flights. *Source* Own graphic, data from Dalal et al. (1989, p. 949)

unprepared to properly evaluate the risks of launching the […] mission in conditions more extreme than they had encountered before" (Presidential Commission on the Space Shuttle Challenger Accident (1986, p. 148).

But how could the logistic regression have been used to get an even better analysis of the problem? Looking more closely, correlation analyses are not necessarily appropriate in this case either. The general question is how the external temperature affects the probability of an O-ring failure. Since the probability is defined as a number between zero (no occurrence) and one (certain event), a regression should not predict values for y that are less than zero or greater than one. This is not guaranteed if a simple OLS regression is applied.

Figure 11.3 shows graphically that when an OLS regression is used, temperatures below about 50 °F will result in an estimated probability of O-ring failure of greater than 100%, and temperatures above about 78 °F will result in a negative probability.

This is only the first problem that arises when OLS regression is used. In addition, the error terms are not normally distributed, nor do they satisfy the condition of homoscedasticity. This results in erroneous and inefficient estimates. In other words, the estimated results are wrong. It is for this reason that OLS regressions are generally not used in empirical research when the dependent variable is dichotomous.

A different approach is therefore needed in order to obtain consistent and unbiased estimates for probabilities with restricted values between zero and one. In practice, various approaches have been suggested, with logistic regression being

one of the most commonly used method. Let's apply this method to our example of potential O-ring failures described above.

First, we face the problem that the a priori probability of failure of an O-ring cannot be observed directly. The only thing that we are able to observe in a direct way is the outcome of whether or not a ring has failed ($y = 1$ failure; $y = 0$ no failure) depending on a set of independent variables (x_i) such as the outside temperature.

$$P(y = 1|x) = f(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k). \tag{11.1}$$

In theory, the unobservable a priori probabilities are referred to as latent variable. However, these unobservable a priori probabilities do have an impact on the observed data (in our case, the actual failure or non-failure of an O-ring). And this is exactly the concept we will use in order to determine the "hidden" a priori probabilities.

First, we need to find a way of defining a latent variable on the left-hand side of the equation in such a way that its deduced value for the probability variable is constrained between zero and one. In order to do this, we need to take a short detour and look at the definitions of odds and log-odds. This will allow us later to predict the a priori probabilities.

11.1 Odds, Log-Odds, and Maximum Likelihood Estimation

Odds are a measure of the likelihood of an event occurring compared to its non-occurrence. In statistics, the odds are the ratio of the probability that an event will occur (p) to the probability that it will not occur ($1 - p$).

$$\text{odds} = \frac{p}{(1-p)}. \tag{11.2}$$

This concept is widely used in the world of sports betting. Odds indicate how much a better can win if the bet is successful. Odds reflect the probability that a particular outcome of an event will occur. Assuming 80% of the bets are on Team A to win and 20% are on Team B to win (assuming a draw of the match is not possible), the odds are calculated as follows:

$$\text{odds} = \frac{0.8}{(1-0.8)} = 4. \tag{11.3}$$

The odds for Team B to win are four to one, meaning you would have to wager one unit to win four units. Odds can never be negative, as this would imply that the probability of one team winning would either be negative or greater than 100%.

However, odds have no upper limit and may converge towards infinity: the higher the probability of one team winning, the higher the odds if the other team

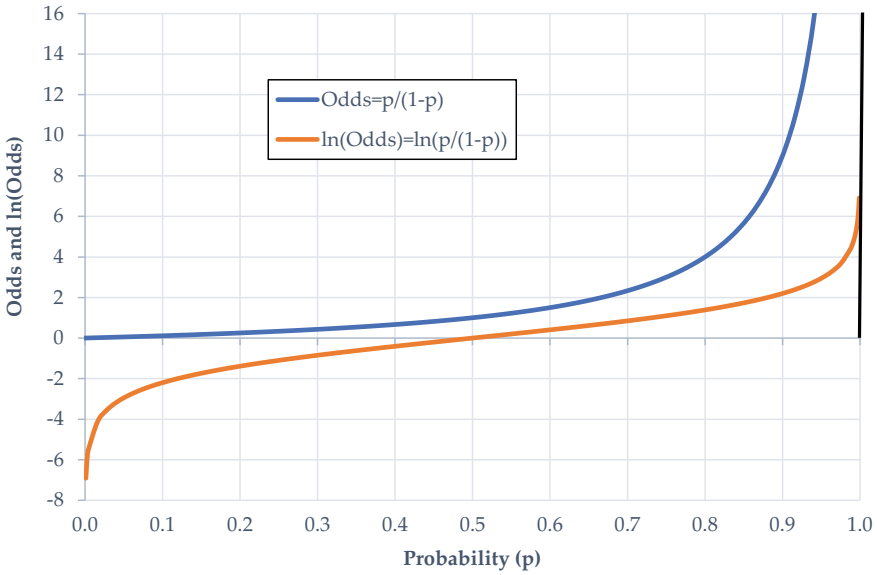


Fig. 11.4 Odds and log-odds. *Source* Own graphic, data from Dalal et al. (1989, p. 949)

wins. In Fig. 11.4, this is expressed by the blue curve, which tends to infinity for the odds on the *y*-axis, with a probability close to 100% on the *x*-axis. Whereas probabilities have a range of values between zero and one ($0 \leq p \leq 1$), odds have a range of values between zero and infinity ($0 \leq \text{odds} \leq \infty$).

Returning to our example of O-ring failures, the 23 shuttle flights including seven flights with O-ring failures and 16 flights with no O-ring failures result in odds of one O-ring failure flight to 2.29 without O-ring failures:

$$\text{odds} = \frac{\frac{16}{23}}{\left(1 - \frac{16}{23}\right)} = 2.29. \tag{11.4}$$

The next step is to take the logarithm of these odds, which results in the so-called log-odds:

$$\ln(\text{odds}) = \ln\left(\frac{p}{(1-p)}\right). \tag{11.5}$$

The course of such a log-odds is illustrated by the curve shown in orange in Fig. 11.4. A probability that tends to 100% on the *x*-axis will tend to infinity on the *y*-axis for the log-odds. With a probability of zero, the values tend towards negative infinity. The possible range of log-odds is between negative and positive infinity ($-\infty \leq \text{log-odds} \leq \infty$).

We have now solved our first problem: we can use log-odds on the left-hand side of the regression equation, for which estimates are between negative and positive infinity:

$$\ln\left(\frac{p}{(1-p)}\right) = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k. \tag{11.6}$$

The log-odds can be transformed into the latent variable of probability values p, constrained between zero and one, as follows[1]:

$$\mathrm{p} = \frac{1}{\left(1 + \mathrm{e}^{-(\beta_0+\beta_1 \cdot x_1+\cdots+\beta_k \cdot x_k)}\right)}. \tag{11.7}$$

The requirement of the probability values being between zero and one is met. Logistic functions show a variety of different behaviours (see Fig. 11.5): if β_1 is negative (see the blue S-curve in the graph at the top of Fig. 11.5), the probability decreases as the x-value increases. This is exactly the case in our example of the O-ring failure: the higher the temperature, the lower the probability of an O-ring failure. In the case of a positive coefficient β_1, the probability increases as the value of x increases (see the other S-curves in the graph at the top of Fig. 11.5). Thereby the slope in the middle section of the S-curve increases as the size of the coefficient increases. Changes in the constant term will shift the S-curve to the right for a negative β_0 and to the left for a positive β_0 (see the graph at the bottom of Fig. 11.5).

Given these types of S-curves, how should we proceed to estimate a logistic regression? Due to its non-linear nature, OLS regression is not applicable. In empirical research, logistic regression is therefore estimated by means of the so-called maximum likelihood estimation (MLE). This method relies on the concept to estimate the parameters of a probability distribution by maximizing a so-called likelihood function. This function measures how likely it is that the observed data would occur if the values of the parameters were changed. The aim of MLE is to provide the best fit to the data by setting the regression coefficients (β_0, …, β_k)

[1]
$$\begin{aligned}
&\left(\frac{p}{(1-p)}\right) = \mathrm{e}^{(\beta_0+\beta_1 \cdot x_1+\cdots+\beta_k \cdot x_k)} \\
&\Leftrightarrow p = (1-p) \cdot \mathrm{e}^{(\beta_0+\beta_1 \cdot x_1+\cdots+\beta_k \cdot x_k)} \\
&\Leftrightarrow p = \mathrm{e}^{(\beta_0+\beta_1 \cdot x_1+\cdots+\beta_k \cdot x_k)} - p \cdot \mathrm{e}^{(\beta_0+\beta_1 \cdot x_1+\cdots+\beta_k \cdot x_k)} \\
&\Leftrightarrow p + p \cdot \mathrm{e}^{(\beta_0+\beta_1 \cdot x_1+\cdots+\beta_k \cdot x_k)} = \mathrm{e}^{(\beta_0+\beta_1 \cdot x_1+\cdots+\beta_k \cdot x_k)} \\
&\Leftrightarrow p \cdot \left(1 + \mathrm{e}^{(\beta_0+\beta_1 \cdot x_1+\cdots+\beta_k \cdot x_k)}\right) = \mathrm{e}^{(\beta_0+\beta_1 \cdot x_1+\cdots+\beta_k \cdot x_k)} \\
&\Leftrightarrow p = \frac{\mathrm{e}^{(\beta_0+\beta_1 \cdot x_1+\cdots+\beta_k \cdot x_k)}}{\left(1 + \mathrm{e}^{(\beta_0+\beta_1 \cdot x_1+\cdots+\beta_k \cdot x_k)}\right)} \\
&\Leftrightarrow p = \frac{1}{\left(1 + \mathrm{e}^{-(\beta_0+\beta_1 \cdot x_1+\cdots+\beta_k \cdot x_k)}\right)}
\end{aligned}$$

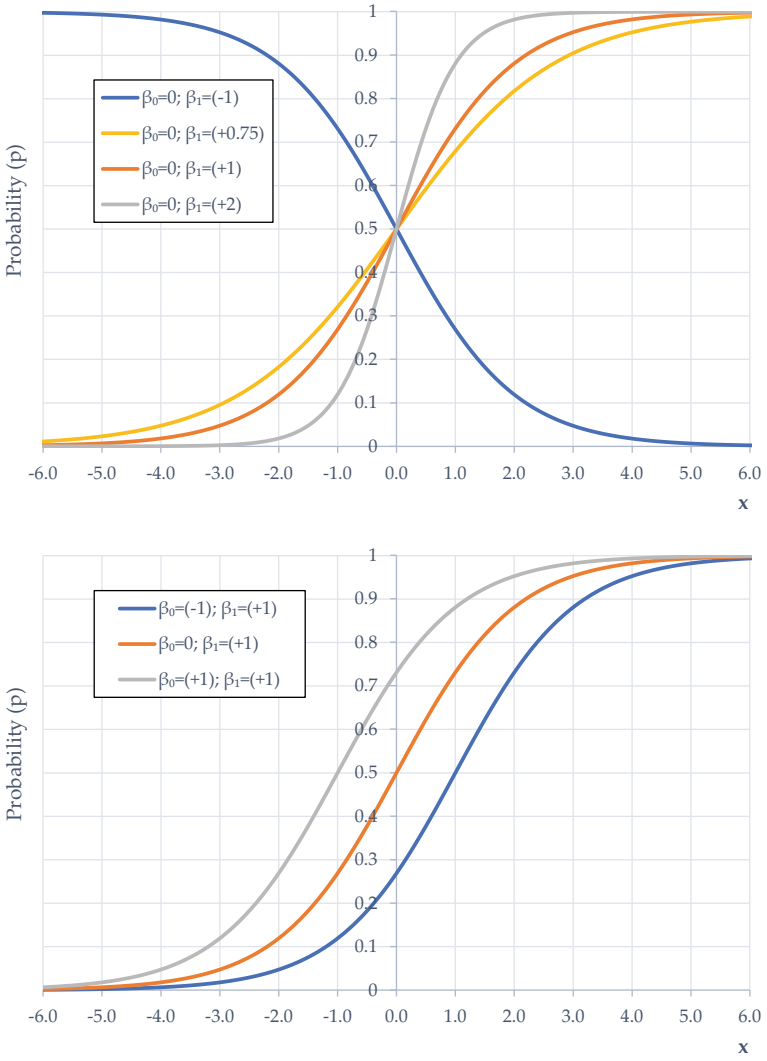


Fig. 11.5 Different behaviour of logistic functions. *Source* Own graphic

that make the observed data most likely. This is usually done by means of iterative optimisation methods such as the Newton–Raphson algorithm or gradient descent methods. In practice, however, we leave the complex calculations to the computer. "The analyst needs not fully understand the magic of this process. However, when conducting analyses that invoke iteration, it is important to confirm that the iteration "converges" (i.e. stabilizes). There is no guarantee that for a given dataset an iterative process will converge. So, to avoid "infinite loops", software has a default maximum number of iterations that is allowed. If the solution does not

converge, override the default iteration limits by allowing more iterations. If you set this number ridiculously high, and the solution still will not converge, the likely culprit is insufficient sample size" (Thompson, 2006, p. 409).

11.2 Estimating a Logistic Regression

Let us now apply the findings we have just gained to the example of the Challenger disaster. Figure 11.6 shows the binary regression curve between temperature (independent) and O-ring failure (dependent). The negative and non-linear relationship between the two variables is evident: the higher the temperature, the lower the probability of an O-ring failure.

In a next step, we leave the exact calculation of the logistic regression to the statistical software. In addition to the metric scaled temperature, we want to consider another independent variable, namely the fact that a given flight was a Challenger flight or the flight of another space shuttle (Columbia, Discovery, or Atlantis). In other words, it is a dummy variable that assumes the value one if it was a Challenger flight and the value zero if it was not (Challenger yes = 1; no = 0). As with OLS regressions, both metric variables and dummy variables are allowed to be used as independent variables in binary-logistic regressions. At the time of the accident, nine of the 23 flights of the space shuttle had been carried out with the Challenger.

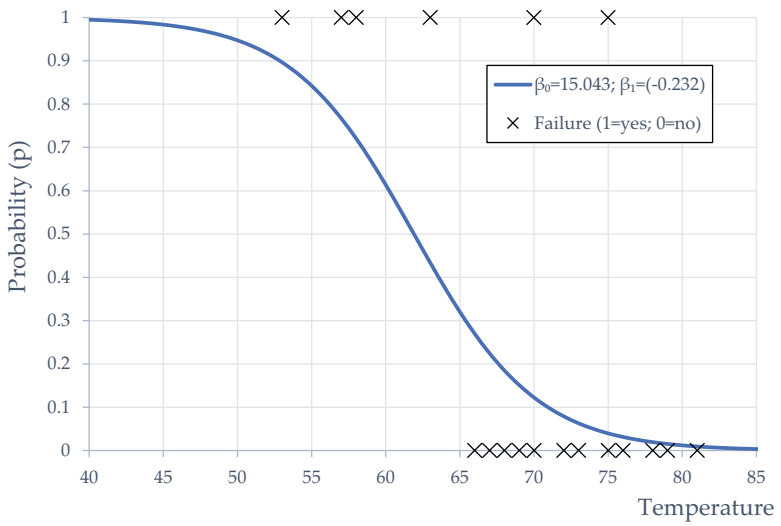


Fig. 11.6 Logistic regression and the shape of the S-curve regarding O-ring failures. *Source* Own graphic, data from Dalal et al. (1989, p. 949)

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients Constant	Temperature in Fahrenheit	Space Shuttle Challenger (yes=1/no=0)
Step 1	1	20.631	9.964	-.158	.621
	2	19.791	14.086	-.222	.840
	3	19.751	15.350	-.241	.887
	4	19.751	15.446	-.243	.889
	5	19.751	15.446	-.243	.889

a. Method: Enter

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 28.267

d. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	8.516	2	.014
	Block	8.516	2	.014
	Model	8.516	2	.014

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	8.917	8	.349

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	19.751[a]	.309	.437

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Temperature in Fahrenheit	-.243	.110	4.842	1	.028	.785	.632	.974
	Space Shuttle Challenger (yes=1/no=0)	.889	1.198	.551	1	.458	2.433	.232	25.466
	Constant	15.446	7.468	4.278	1	.039	5107830.286		

a. Variable(s) entered on step 1: Temperature in Fahrenheit, Space Shuttle Challenger (yes=1/no=0).

Classification Table[a]

Observed		Predicted Damaged O-rings (yes=1/no=0) 0	1	Percentage Correct
Step 1 Damaged O-rings (yes=1/no=0)	0	15	1	93.8
	1	3	4	57.1
Overall Percentage				82.6

a. The cut value is .500

Fig. 11.7 Logistic regression results. *Source* Own graphic

The results of the logistic regression based on SPSS are shown in Fig. 11.7. At this point, we are not able to make sense of all these results. Therefore, for the time being, we are going to focus only on the table "Variables in the Equation" (see in Fig. 11.7).

The negative coefficient $B = (-\ 0.243)$ for the temperature in Fahrenheit indicates that as the temperature increases, the probability of an O-ring failure decreases. The positive coefficient $B = (+\ 0.889)$ for the dummy variable suggests that when the space shuttle Challenger is involved (i.e. yes = 1), the likelihood of an O-ring failure increases compared to all other shuttle missions. However, as we will see later, this effect is not statistically significant. Unlike OLS regression, the estimated coefficients of the independent variables cannot be interpreted in a direct way as marginal effects for the probability p of an event. Instead, the coefficients express the linear relationship between the independent variables and the log-odds. As shown in Eq. 11.6, they indicate how the log-odds of the dependent variable change with a change of one unit in the predictor variable.

$$\ln\left(\frac{p}{(1-p)}\right) = 15.446 - 0.243 \cdot \text{temperature} + 0.889 \cdot \text{challenger}[Y = 1; N = 0]. \tag{11.8}$$

However, as we have shown in Eq. 11.7, this can be transformed into estimates of the probability (latent variable) of an O-ring failure as follows:

$$\begin{aligned} p &= \frac{1}{\left(1 + \mathrm{e}^{-(\beta_0 + \beta_1 \cdot \text{temperature} + \beta_2 \cdot \text{challenger})}\right)} \\ &= \frac{1}{\left(1 + \mathrm{e}^{-(15.446 - 0.243 \cdot \text{temperature} + 0.889 \cdot \text{challenger})}\right)}. \end{aligned} \tag{11.9}$$

On the day of the Challenger disaster, the temperature was 31 °F. So, the probability of O-ring failure during the Challenger mission was 99.985%:

$$\mathrm{p}_{31\,^\circ\mathrm{F}} = \frac{1}{\left(1 + \mathrm{e}^{-(15.446 - 0.243 \cdot 31 + 0.889 \cdot 1)}\right)} = 99.985\%. \tag{11.10}$$

But if the marginal effect of a one-unit change in temperature cannot be obtained directly from a regressor, how can it be derived indirectly? A little trick is used to do this by calculating the so-called odds ratio. The first thing we need to do is to transform Eq. 11.6 so that we get the formula for calculating the odds for a given value of x_1:

$$\text{odds}(x_1) = \left(\frac{p}{(1-p)}\right) = e^{(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k)}. \tag{11.11}$$

If x_1 increases by one unit, the value for odds is:

$$\text{odds}(x_1 + 1) = \left(\frac{p}{(1-p)}\right) = \mathrm{e}^{(\beta_0 + \beta_1 \cdot (x_1 + 1) + \cdots + \beta_k \cdot x_k)} = \mathrm{e}^{(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k)} \cdot \mathrm{e}^{\beta_1}. \tag{11.12}$$

In the next step, the two odds are simply set in relation to each other, resulting in the odds ratio (Exp(B)):

$$\text{Exp}(B) = \text{Odds Ratio} = \frac{\text{odds}(x_1 + 1)}{\text{odds}(x_1)} = \frac{\mathrm{e}^{(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k)} \cdot \mathrm{e}^{\beta_1}}{\mathrm{e}^{(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k)}} = \mathrm{e}^{\beta_1}. \tag{11.13}$$

The odds ratio tells us how the outcome changes when the predictor variable increases by one unit. In interpreting Exp(B), the following generally applies:

- Exp(B) > 1: a one-unit increase in the predictor variable increases the odds of the outcome.
- Exp(B) < 1: a one-unit increase in the predictor variable decreases the odds of the outcome.
- Exp(B) = 1: a one-unit increase in the predictor variable has no effect on the odds of the outcome.

The standard software provides the value for each individual regressor labelled in the column Exp(B) in order to save the user from having to make a manual calculation. In our example, an odds ratio of Exp(B) = 0.785 for the independent variable temperature (see Fig. 11.7) means that for every unit increase in temperature, the odds of the event occurring decrease by about 21.5% (because 1 − 0.785 = 0.215). An odds ratio of 2.433 for the Challenger dummy variable (see Fig. 11.7) suggests that the odds of an O-ring failure during a Challenger mission are more than twice as high (2.433 − 1 = 143.3%) as for other space shuttles. However, the 95% confidence interval for Exp(B) is very wide [0.232, 25.466], ranging from values far below 1 (decrease) to well above 1 (increase). This suggests a high level of uncertainty in the estimated effect. As we will see in the next section, this makes the effect of this dummy variable statistically insignificant (p = 0.458).

11.3 Testing for Significance

11.3.1 Overall Model Significance (Likelihood Ratio Test)

Similar to the F-test in linear regression (OLS), the likelihood ratio test (LRT) assesses whether (at least one of) the used predictors contribute significantly to the explanation of the model. The hypothesis is therefore tested that none of the independent variables exert an influence on the dependent variable (H_0):

$$H_0 : \beta_0 = \beta_1 = \cdots = \beta_k = 0$$
$$\text{versus } H_1 : \text{ at least one of the parameters is not equal to zero.} \qquad (11.14)$$

The concept of this test is quite simple: as mentioned above, logistic regression is estimated using the maximum likelihood method by including all independent variables in the model. Based on this, the so-called log-likelihood of the "full model" (LL_{full}) can be calculated. In the modelling process, it is important that this full model with its set of independent variables has as much explanatory power as possible. In order to assess the level of the explanatory power, we need to define a basis for comparison. What could be more plausible than using the log-likelihood of the so-called null model, which is nothing else than the log-likelihood of the model with just the constant term and no other independent variable (LL_0)? The values of the two log-likelihoods in our example would be LL_{full} = (− 9.876) and LL_0 = (− 14.156). From these numbers, we can see that the log-likelihoods are negative, because the logarithm of probabilities (which are between zero and one) is always negative. We are also able to determine that the value for the full model is larger than the one for the null model.

The likelihood ratio test then compares this difference between the full model (including all predictors x_1,…, x_k) and the null model (without predictors) to assess whether the predictors significantly improve the model's explanatory power. The

test simply multiplies each of the log-likelihoods by (− 2):

$$\text{LRT} = (-2)(\text{LL}_0 - \text{LL}_{\text{full}}) \tag{11.15}$$

$$\text{LRT} = -2\text{LL}_0 + 2\text{LL}_{\text{full}} \tag{11.16}$$

Why is the difference now multiplied by (− 2)? Well, the LRT statistic then follows asymptotically a χ^2 distribution with k degrees of freedom, where k is the number of additional parameters in the full model. It applies as follows:

$$\chi^2 = (-2\text{LL}_0) + (2\text{LL}_{\text{full}}). \tag{11.17}$$

In our example in Fig. 11.7, the results of the LRT can be found in the "Omnibus Tests of Model Coefficients" table and the "Model Summary" table: the score for − 2LL for the full model is $(-2\text{LL}_{\text{full}}) = 19.751$. The increase over the null model is equal to the value of the reported chi-squared value of $\chi^2 = 8.561$. For the null model, this results in a value of $(-2\text{LL}_0) = 28.312 = 19.751 + 8.561$. This results in a corresponding result for the likelihood ratio test:

$$\chi^2 = -2\text{LL}_0 + 2\text{LL}_{\text{full}} = 28.312 - 19.751 = 8.561. \tag{11.18}$$

With $k = 2$ degrees of freedom (two independent variables), this results in a p-value of $p = 0.014$. Assuming a significance level of 0.05, this means not only that the full model is significantly better than the null model in its exploratory power, but above all that at least one of the two independent variables exerts an influence on the dependent variable (O-ring failure) and therefore at least one of the parameters of β is not equal to zero.

11.3.2 Significance of Individual Coefficients (Wald Test)

The question that remains to be answered is which of the independent variables has a significant impact on the dependent variable. For each independent variable x_i, it is therefore necessary to test whether the regression coefficient is significantly different from zero in terms of its impact on the log-odds of the outcome:

$$H_o \beta_i = 0 \text{ versus } H_1 : \beta_i \neq 0. \tag{11.19}$$

In the case of an OLS regression, this is expressed in terms of the t-value and the associated p-value (see Chap. 10). The Wald statistic is a generalization of the t-statistic concept, so it can be used in the case of non-linear logistic regression. It is defined as the squared ratio of the estimated coefficient to the standard error:

$$W = \left(\frac{\widehat{\beta}}{\text{S.E.}\left(\widehat{\beta}\right)} \right)^2. \tag{11.20}$$

The only difference from the t-statistic is that the Wald statistic is squared. Therefore, the Wald statistic asymptotically follows a χ^2 distribution with one degree of freedom instead of a t-distribution. A larger Wald statistic leads to a lower p-value. If the p-value is lower than the significance level, we reject the null hypothesis.

In our example in Fig. 11.7, the Wald statistic and the associated p-value can be found in the "Variables in the Equation" table. With a Wald statistic of $W = 4.842$ and a p-value of $p = 0.028$, the odds ratio of $\text{Exp}(B) = 0.785$ for the temperature in Fahrenheit is statistically significant at the five per cent level of significance. This means that the probability of O-ring failure decreases significantly by approximately 22% for each unit increase in temperature. For the dummy variable of Challenger flights, compared to flights with other shuttles, this is completely different. In this case, the Wald statistic of $W = 0.551$ results in a p-value of $p = 0.458$, which means that the effect of this dummy variable is not statistically significant ($p = 0.458$) at the five per cent level of significance. Controlled for external temperature, Challenger flights did not have significantly higher O-ring failure probabilities than other shuttle missions.

11.4 Pseudo-R^2

In Chap. 10, we have seen the traditional R^2 and the adjusted R^2 as widely used parameter to measure the proportion of the variance explained by the variance of the independent variables in a linear model. This intuitive concept is emphasized in many empirical studies when assessing the quality of linear OLS regression. Not surprising, because the interpretation sounds so catchy: so and so many per cent of the variance of the dependent variable is explained by variance in the independent variable(s). Unfortunately, the classic coefficient of determination R^2 cannot be used as a quality measure in logistic regression. However, the good news is that the so-called pseudo-R^2 can be used. These parameters also lie in an interval between zero and one with hardly any exceptions. They do not result from a quotient of the explained variance to the total variance, but from the quotient of the (log) likelihood of the full model (including all independent variables) and the null model (including no independent variables). The pseudo-R^2 is therefore a measure of the gain in information provided by the independent variables of the full model compared to the null model. In practice, McFadden's R^2, Cox-Snell's R^2, and Nagelkerke's R^2 are often used.

McFadden's R^2 (1978) is calculated as the ratio of the log-likelihood of the full (fitted) model to the log-likelihood of a null model (a model with only the intercept):

$$R^2_{\text{McF}} = 1 - \left(\frac{\text{LL}_{\text{full}}}{\text{LL}_0} \right) \tag{11.21}$$

McFadden R^2 is the most conservative pseudo-R^2. In theory, values range between 0 and 1, but are often much lower than would be expected from a classical R^2. McFadden (1978, 307) personally states that "values tend to be considerably lower than those of the R^2 index and should not be judged by the standards for a 'good fit' in ordinary regression analysis. For example, values of 0.2–0.4 […] represent an excellent fit". In our example in Fig. 11.7, result for McFadden R^2 is not directly provided, but can be calculated from the result of the "Iteration History" table[2]:

$$R^2_{\text{McF}} = 1 - \left(\frac{-9.876}{-14.134}\right) = (1 - 0.699) = 0.301. \tag{11.22}$$

This means that the greater the improvement of the full model, the closer the ratio $\text{LL}_{\text{full}}/\text{LL}_0$ will be to zero and thus the Mc Fadden R^2 will approach one. In our example, the model with predictors improves the log-likelihood by 30% compared to the null model.

Cox-Snell R^2 (Cox & Snell, 1989) is based on the ratio of the likelihood of the null model divided by the likelihood of the full model (not the log-likelihoods!) and takes into account the size of the sample:

$$\text{R}^2_{\text{CS}} = 1 - \left(\frac{\text{L}_0}{\text{L}_{\text{full}}}\right)^{\frac{2}{n}}. \tag{11.23}$$

It is defined in such a way that it can theoretically range between zero (if the L_{full} equals L_0) and an upper bound that is typically less than one, even if the model is a perfect fit to the data. Despite this limitation, it is often used in empirical studies. In our example in Fig. 11.7, the Cox-Snell R^2 result given in the "Model Summary" table is:

$$R^2_{\text{CS}} = 0.309. \tag{11.24}$$

This does not represent a percentage of variance explained, as it would for linear regression, but indicates that the logistic regression model has improved the prediction of the outcome relative to the null model. Instead, it reflects a relative improvement in the predictive power of the model.

The Nagelkerke R^2 (1991) is an adapted version that modifies the Cox-Snell R^2 to ensure that it ranges between 0 and 1, making it easier to interpret:

$$R^2_N = 1 - \left(\frac{R^2_{\text{CS}}}{1 - L_0^{(2/k)}}\right). \tag{11.25}$$

[2] $(-2\text{LL}_0) = 28.267$ results for the null model (see Fig. 11.7), so resulting in $\text{LL}_0 = (-14.134)$ for the log likelihood. $(-2\text{LL}_{\text{full}}) = 19.751$ results for the full model (see Fig. 11.7), so resulting in $\text{LL}_{\text{full}} = (-9.876)$ for the log likelihood.

Again, as it is the case with the other pseudo-R^2, Nagelkerke R^2 is not a measure of the explained variance. However, it reflects the improvement in the likelihood of the observed outcomes under the fitted model compared to the null model. Research in many disciplines often favours Nagelkerke's R^2 because it provides a pseudo-R^2 that is easy to understand and communicate, especially in situations where the audience may not be familiar with more complex statistical concepts. In our example in Fig. 11.7, the Nagelkerke R^2 result given in the "Model Summary" table is:

$$R_N^2 = 0.437. \tag{11.26}$$

In summary, all types of pseudo-R^2 are conceptually different from the linear regression R^2. They are not comparable, neither in terms of their magnitude nor in terms of their interpretation. Low pseudo-R^2 "in a logistic regression are the norm and this presents a problem when reporting their values to an audience accustomed to seeing linear regression values" (Homer and Lemeshow 2000, 167). Therefore, researchers should not be discouraged by low pseudo-R^2 values. On the contrary, large values of pseudo-R^2 may be a first indication of an incorrect specification of the empirical model.

11.5 Goodness of Fit of a Logistic Regression

The various pseudo-R^2s have provided us with measures that reflect the degree of explanation of a model in addition to the null model. However, they do not tell us how well the logistic regression as a whole fits the observed data. In empirical practice, the Hosmer–Lemeshow test is often used to test if a logistic regression model is good at predicting the observed outcomes. Without going into too much detail, the basic concept of the test will be briefly described using the data from our example of the Challenger disaster.

As we have seen, based on logistic regression, each space shuttle flight has its own individual probability of O-ring failure, e.g. given the outside temperature of 78 °F for Challenger Flight 41-B, Eq. 11.7 yields an O-ring failure probability of around seven per cent:

$$p = \frac{1}{\left(1 + \mathrm{e}^{-(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k)}\right)} = \frac{1}{\left(1 + \mathrm{e}^{-(15.446 - 0.243 \cdot 78 + 0.889 \cdot 1)}\right)} = 7\%. \tag{11.27}$$

The Hosmer–Lemeshow test takes all these flights and groups them according to their predicted probabilities. These predicted probabilities are then sorted in ascending order. One group could be flights with a 0–10% chance of an O-ring failure, the second group 10–20%, and so on and so forth. It is common to have a number of ten groups. The resulting contingency table is shown in Fig. 11.8. The χ^2 statistic in the Hosmer–Lemeshow test is calculated by summing the squared

differences between observed and expected values for both events and non-events across the (10) different groups. These differences are normalized by the expected values. This is the usual procedure for a chi-square test, which we learned about in Sect. 9.6.1. In the example, this gives a $\chi^2 = 8.917$ with eight degrees of freedom (10 groups-2), resulting in a p-value of $p = 0.349$. The resulting p-value indicates how well the logistic regression model fits the observed data. A significant result would indicate a poor fit, suggesting that the predictions of the model do not fit well with the observed outcomes. If the model fits well, the observed and expected frequencies should be close and the p-value will be high and insignificant. The p-value of $p = 0.349$ therefore indicates that the logistic regression estimates are not significantly different from the observations, meaning that the logistic regression model fits the observed data.

Valuable information about the quality of the selected model can be obtained by combining the results of the Hosmer–Lemeshow test and the likelihood ratio test. A significant Hosmer–Lemeshow test combined with a significant likelihood

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	8.917	8	.349

Contingency Table for Hosmer and Lemeshow Test

		Damaged O-rings (yes=1/no=0) = 0		Damaged O-rings (yes=1/no=0) = 1		
		Observed	Expected	Observed	Expected	Total
Step 1	1	2	1.942	0	.058	2
	2	2	1.905	0	.095	2
	3	2	2.662	1	.338	3
	4	2	3.294	2	.706	4
	5	2	1.585	0	.415	2
	6	2	1.499	0	.501	2
	7	2	1.385	0	.615	2
	8	2	1.120	0	.880	2
	9	0	.462	2	1.538	2
	10	0	.146	2	1.854	2

Classification Table[a]

			Predicted		
			Damaged O-rings (yes=1/no=0)		Percentage Correct
	Observed		0	1	
Step 1	Damaged O-rings (yes=1/no=0)	0	15	1	93.8
		1	3	4	57.1
	Overall Percentage				82.6

a. The cut value is .500

Fig. 11.8 Hosmer–Lemeshow test and classification table. *Source* Own graphic

ratio test suggests that although the full model is significantly better than the null model, the fit of the full model is still inadequate. This scenario calls for further model refinement or reconsideration of the modelling approach.

11.6 Classification Performance

The logistic regression classification table—also known as the confusion table—is a useful way of assessing the predictive classification performance of a binary model.

The individual observations are divided into two classes with respect to their two actual outcomes, actual positive ($Y = 1$) and actual negative ($Y = 0$). In our example, the seven shuttle flights with O-ring failure are the actual positive cases, all the other 16 flights are the actual negative cases. These two classes are then separated again for cases where the individual probability estimated by logistic regression exceeds (prediction: positive = 1) or falls below (prediction: negative = 0) a defined threshold. Typically, a threshold of 0.5 is used: if the predicted probability is greater than or equal to 0.5, the outcome is classified as positive, otherwise, it is classified as negative (see Table 11.1).

Therefore, the confusion table consists of four classes: cases where the model correctly predicted either the positive class (true positives (TP)) or the negative class (true negatives (TN)), but also the two incorrectly predicted classes false positives (FP), where the model incorrectly predicted a positive class when the actual class was negative, and false negatives (FN), where the model incorrectly predicted a negative class when the actual class was positive.

The prediction of the logistic regression model tends to be accurate when the percentage of correctly classified instances is high; however, it can be misleading in cases of imbalanced class distributions (see below). The accuracy indicator shows the percentage of correctly classified outcomes:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{11.28}$$

In the example of the shuttle flights in Fig. 11.7, the accuracy is 82.6% of correctly classified cases (= 19/23). But more interesting is the question of how sensitive the model is to O-ring failure. In other words, which percentage of O-ring

Table 11.1 Classification table of a binary regression model

	Prediction: negative ($\widehat{Y} = 0$)	Prediction: positive ($\widehat{Y} = 1$)
Actual: negative ($Y = 0$)	True negatives (TN)	False positives (FP)
Actual: positive ($Y = 1$)	False negatives (FN)	True positives (TP)

failures would the model have correctly predicted?:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{11.29}$$

This occurs in four out of seven cases, giving a sensitivity of only about 57%. O-ring failures are therefore not well identified. Precision, on the other hand, is the proportion of positive predictions that are actually classified correctly:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{11.30}$$

This is the case in four out of five cases, or 80%.

The last key indicator is known as specificity. It is the proportion of actual negative cases that were correctly identified by the model:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{11.31}$$

This is the case in 15 out of 16 cases, or 94%, in our space shuttle example. The low value for the sensitivity shows that the unreflective use of confusion tables to assess the predictive power of a logistic model can be problematic.

Logistic regression models are always best at predicting the most frequent outcomes. If the class distribution between the two outcomes is highly unbalanced (e.g. in case of rare events), the accuracy may be misleading. A binary regression model that always predicts the majority class may still have high overall accuracy. However, it may perform poorly at identifying the minority class. Therefore, each of the two outcome categories should account for at least five per cent of the observations. Otherwise, so-called rare event models should be used (King & Zeng, 2001).

Another problem is the threshold dependence of the prediction. In our example, all observations with an estimate probability greater than 50% are assigned to the O-ring failure prediction group. It makes no difference whether the probability is 51% or 99%, whereas two observations with 49% and 51% probabilities would be assigned to different prediction groups. Selecting an appropriate threshold is often challenging. A model with an inappropriate threshold may have poor predictive power, even if the underlying model is robust.

Using the shuttle example, a lower threshold would have increased sensitivity and identified more positive cases than the 57% of O-ring failures detected. However, this comes at the cost of an increasing number of false positives, which brings us to the final issue: the confusion table does not take into account the cost of false positive and false negative classifications. The consequences and associated costs of misclassification can vary, so the choice of an appropriate threshold is again particularly important.

11.7 Impact of Heteroskedasticity

In logistic regression, the variance of the error terms depends on the predicted probability p. This implies that some sort of heteroscedasticity is inherent in logistic regression and is accounted for by the model. However, "additional" heteroscedasticity can cause problems that ultimately lead to biased significance tests and incorrect conclusions:

- Biased standard errors can lead to unreliable Wald tests and likelihood ratio tests, affecting p-values and the decisions about the significance of predictors.
- Measures of model fit, such as pseudo-R^2 or the Hosmer–Lemeshow test, may be biased.
- While heteroskedasticity does not affect the consistency of the regression coefficients in the OLS model, the coefficients are inconsistent in a logistic regression. In other words, as the sample size increases, the estimators do not converge to the true parameter values.

In most cases, heteroskedasticity arises when the model is misspecified or when there are systematic errors not addressed by the model. Therefore, the model must be checked for error specification, such as for missing variables or an erroneous choice of the functional distribution.

11.8 Impact of Multicollinearity

Multicollinearity in logistic regression occurs when two or more independent variables have a high correlation with one another. The effects of multicollinearity in linear OLS models have already been discussed in detail in Chap. 10 and are little different from those in logistic regression. In general, multicollinearity in a logistic regression leads to instability of coefficient estimates (sometimes even algebraic signs of the regressor may be unusual) and inflated standard errors leading to unreliable hypothesis testing. Therefore, it can also affect the classification accuracy in the confusion table.

The first indications of the presence of multicollinearity are large changes in the regression coefficients when an observation is removed or added to the dataset, and removing a variable from the model results in only a small decrease in pseudo-R^2. A reliable indication of the presence of multicollinearity is provided by the *Variance Inflation Factors (VIF)*. See Chap. 10 for more detailed information. The larger the VIF, the stronger the undesired correlation between two or more independent variables of the logistic regression. If, by contrast, VIF equals one, no multicollinearity exists. Hair et al. (2006, p. 230) note that VIF = 10 is a frequently used upper limit but recommend a more restrictive value for smaller samples. Ultimately, every researcher must make his or her own decision about the acceptable degree of multicollinearity and, when the VIF is conspicuously high, check the robustness of the results. Keep in mind, though, that a VIF as low as 5.3 already

has a very high multiple correlation, namely, $r = 0.9$. For this reason, whenever the VIF is 2.0 or higher—VIF = 2.0 translates into a multiple correlation of $r = 0.71$—you should test your results, checking to see how they respond to minor changes in the sample. Some statistic software programmes indicate *Tolerance* as well as VIF. When the value of Tolerance is (close to) one, then multicollinearity does not exist. The more the value of Tolerance approaches zero, the larger the multicollinearity.[3]

The following measures can be taken to eliminate multicollinearity:

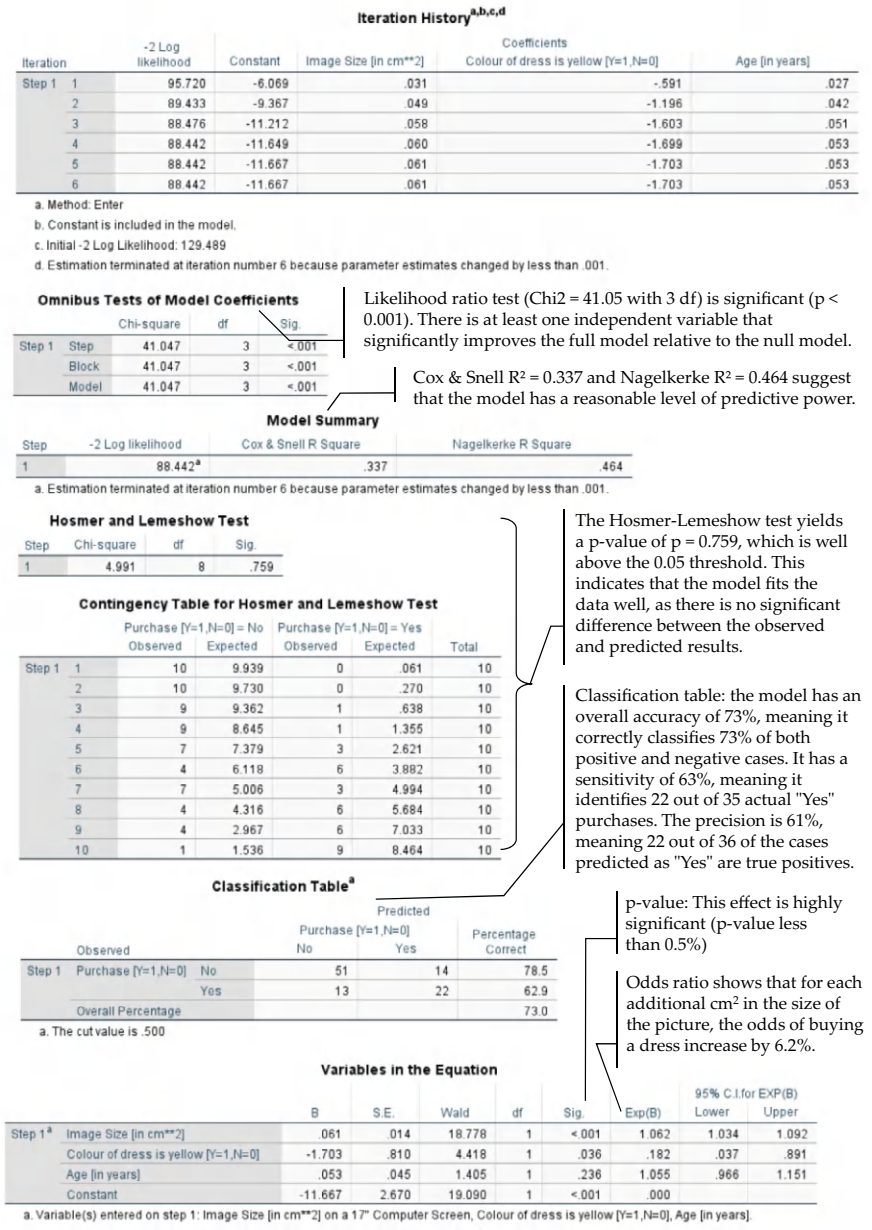
- Remove one of the correlating variables from the logistic regression. The best variable to remove is the one with the highest VIF. From there proceed in steps. Every variable you remove lowers the VIF values of the logistic regression's remaining variables.
- Check the sample size. A small sample might produce multicollinearity even if the variables are not multicollinear throughout the population. If you suspect this could be the case, include additional observations in the sample.
- Reconsider the theoretical assumptions of the model. In particular, ask whether your regression model is overparameterized.
- Not infrequently, correlating variables can be combined into a single variable with the aid of factor analysis (see Chap. 14).

11.9 Another Example: Logistic Regression with the Computer

A fashion brand operates an online platform offering various clothing items, including dresses. The retailer wants to find out how certain visual aspects of the online presentation of the products influence the purchasing behaviour of customers. Therefore, he collects the following information for 100 users of the website: has the person ordered a dress (yes = 1; no = 0) and how old is the person (in years)? How large (in cm^2) was the image of the dress displayed on the screen when ordering? As yellow is seen as a bold and less universal colour, the retailer believes that yellow dresses might be less popular. In this way, he also notes whether the ordered dress was yellow (yellow = 1; another colour = 0). All data are available in the *Chapter11_mail_order_business_2* dataset (see supplementary material at https://doi.org/10.1007/978-3-031-78070-7_11). He then subjected the collected data to a logistic regression analysis with purchase or non-purchase as the dependent variable (see Fig. 11.9).

The results of the analysis show a good model fit of the data. The − 2 log-likelihood statistic improved from $LL_0 = 129.489$ for the null model to $LL_{full} =$

[3] In many statistical packages, the Variance Inflation Factor (VIF) and Tolerance are not directly available from the standard logistic regression menus, as it is primarily associated with linear regression. For this reason, the calculation of VIF and tolerance for predictors in a logistic regression can be done by linear regression using the same predictors.



Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients: Constant	Image Size [in cm**2]	Colour of dress is yellow [Y=1,N=0]	Age [in years]
Step 1	1	95.720	-6.069	.031	-.591	.027
	2	89.433	-9.367	.049	-1.196	.042
	3	88.476	-11.212	.058	-1.603	.051
	4	88.442	-11.649	.060	-1.699	.053
	5	88.442	-11.667	.061	-1.703	.053
	6	88.442	-11.667	.061	-1.703	.053

a. Method: Enter
b. Constant is included in the model.
c. Initial -2 Log Likelihood: 129.489
d. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	41.047	3	<.001
	Block	41.047	3	<.001
	Model	41.047	3	<.001

Likelihood ratio test (Chi2 = 41.05 with 3 df) is significant ($p < 0.001$). There is at least one independent variable that significantly improves the full model relative to the null model.

Cox & Snell $R^2 = 0.337$ and Nagelkerke $R^2 = 0.464$ suggest that the model has a reasonable level of predictive power.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	88.442[a]	.337	.464

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4.991	8	.759

The Hosmer-Lemeshow test yields a p-value of $p = 0.759$, which is well above the 0.05 threshold. This indicates that the model fits the data well, as there is no significant difference between the observed and predicted results.

Contingency Table for Hosmer and Lemeshow Test

		Purchase [Y=1,N=0] = No: Observed	Purchase [Y=1,N=0] = No: Expected	Purchase [Y=1,N=0] = Yes: Observed	Purchase [Y=1,N=0] = Yes: Expected	Total
Step 1	1	10	9.939	0	.061	10
	2	10	9.730	0	.270	10
	3	9	9.362	1	.638	10
	4	9	8.645	1	1.355	10
	5	7	7.379	3	2.621	10
	6	4	6.118	6	3.882	10
	7	7	5.006	3	4.994	10
	8	4	4.316	6	5.684	10
	9	4	2.967	6	7.033	10
	10	1	1.536	9	8.464	10

Classification table: the model has an overall accuracy of 73%, meaning it correctly classifies 73% of both positive and negative cases. It has a sensitivity of 63%, meaning it identifies 22 out of 35 actual "Yes" purchases. The precision is 61%, meaning 22 out of 36 of the cases predicted as "Yes" are true positives.

Classification Table[a]

			Predicted: Purchase [Y=1,N=0] = No	Predicted: Purchase [Y=1,N=0] = Yes	Percentage Correct
Observed					
Step 1	Purchase [Y=1,N=0]	No	51	14	78.5
		Yes	13	22	62.9
	Overall Percentage				73.0

a. The cut value is .500

p-value: This effect is highly significant (p-value less than 0.5%)

Odds ratio shows that for each additional cm^2 in the size of the picture, the odds of buying a dress increase by 6.2%.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Image Size [in cm**2]	.061	.014	18.778	1	<.001	1.062	1.034	1.092
	Colour of dress is yellow [Y=1,N=0]	-1.703	.810	4.418	1	.036	.182	.037	.891
	Age [in years]	.053	.045	1.405	1	.236	1.055	.966	1.151
	Constant	-11.667	2.670	19.090	1	<.001	.000		

a. Variable(s) entered on step 1: Image Size [in cm**2] on a 17" Computer Screen, Colour of dress is yellow [Y=1,N=0], Age [in years].

Fig. 11.9 Logistic regression results for online retail of dresses. *Source* Own graphic

Correlation Matrix

		Constant	Image Size [in cm**2]	Colour of dress is yellow [Y=1,N=0]	Age [in years]
Step 1	Constant	1.000	-.855	.526	-.528
	Image Size [in cm**2]	-.855	1.000	-.561	.024
	Colour of dress is yellow [Y=1,N=0]	.526	-.561	1.000	-.183
	Age [in years]	-.528	.024	-.183	1.000

Fig. 11.9 (continued)

88.442 for the full model. This is a first indication that the predictors (image size, colour, and age) make a significant contribution to the improvement of the model. This is also supported by the results of the likelihood ratio test, which shows a chi-square value of $\chi^2 = 41.047$ with three degrees of freedom and a p-value of $p < 0.001$, meaning that the predictors—image size, dress colour, and age—significantly provide important insights into customer decisions. In other words, there is at least one independent variable that significantly improves the full model relative to the null model. The Cox & Snell $R^2 = 0.337$ and the Nagelkerke $R^2 = 0.464$ also suggest that the model has a reasonable level of predictive power.

The next step is to determine which predictors significantly improve the model. The variable *image size (in cm^2)* has a significant and positive effect on purchase behaviour. The coefficient (B) of 0.061 indicates that larger images increase the likelihood of purchase. This effect is highly significant at $p < 0.001$. The odds ratio ($\text{Exp}(B) = 1.062$) shows that for each additional square centimetre in the size of the picture, the odds of buying a dress increase by 6.2% ($0.062 = 1.062 - 1$).

The colour of the dress, particularly whether the dress is yellow, is another significant ($p = 0.036$) predictor of buying behaviour. With $B = (-1.703)$, the coefficient is negative, which means that this predictor variable reduces the probability of purchase. The odds ratio ($\text{Exp}(B) = 0.182$) suggests that the odds of purchasing a yellow dress are about 81.8% ($0.818 = 1 - 0.182$) lower than purchasing a dress of another colour.

The third predictor, age, is included in the model in order to determine whether older or younger customers are more likely to be purchasers of a dress. However, this effect is not statistically significant ($p = 0.236$), which in turn means that its influence is not significantly different from zero.

The Hosmer–Lemeshow test yields a p-value of $p = 0.759$, which is well above the 0.05 threshold. This indicates that the model fits the data well, as there is no significant difference between the observed and predicted results.

The model's classification accuracy provides a measure of how well it predicts customer behaviour. Overall, the model correctly classifies 73% of cases, meaning it accurately predicts whether a customer will or will not purchase a dress in 73 out of 100 cases. The model performs slightly better at predicting customers who did not purchase a dress, with an accuracy of 78.5%, compared to customers who did make a purchase, with a sensitivity of 62.9%. This indicates that while the model is generally reliable, it is more effective at identifying non-purchasers than purchasers. Nonetheless, the overall classification accuracy is strong and shows that the model provides valuable insights into purchasing behaviour.

Overall, it can be concluded that the model is good. In practice, the variable *age* could have been removed in a subsequent step due to its insignificance. The results of this study provide the retailer with actionable insights into how product presentation affects customer behaviour. By focusing on image size optimization and rethinking the prominence of certain colours, such as yellow, the retailer can make data-driven decisions to improve the online shopping experience and increase sales on its platform.

The following sections show how statistical software can be used to produce the results shown in Fig. 11.9. It should be noted that it is not advisable to perform logistic regression with Excel. There is no built-in logistic regression tool and therefore important diagnostics such as pseudo-R^2, confidence intervals, p-values, or goodness-of-fit tests are not provided. Specialized software such as Stata, R, and SPSS offer much greater efficiency, reliability, and functionality for this type of analyses.

11.9.1 Logistic Regression Calculations with SPSS

In SPSS, load the dataset *Chapter9_mail_order_business_2.sav* (see supplementary material at https://doi.org/10.1007/978-3-031-78070-7_9). Open the *Logistic Regression* window shown in Fig. 11.10 by selecting *Analyze* → *Regression* → *Binary Logistic*. Then assign the dependent and independent variables. Click on *Options* and check boxes for relevant statistics (in our case *Hosmer–Lemeshow goodness-of-fit*, *Correlations of estimates*, *Iteration History, CI for exp(B)*). If necessary, a different value for the threshold of the confusion table can be entered in this window (see *Classification cutoff*). Then click *Continue*. You can save results by clicking the *Save* button and selecting options such as *Predicted Values* (probabilities and group memberships) or *Residuals* for later analysis. The *Categorical* button only needs to be clicked in the case of independent categorical variables that have more than two outcomes. There is no need to explicitly define independent variables as categorical if they are binary (i.e. a dummy variable with only two categories, such as zero and one). Once everything is set up, click *OK*. Results are shown in Fig. 11.9.

11.9.2 Logistic Regression Calculations with Stata

In Stata, load the dataset *Chapter9_mail_order_business_2.dta* (see supplementary material at https://doi.org/10.1007/978-3-031-78070-7_9) and access the regression menu by choosing *Statistics* → *Binary Outcomes* → *Logistic Regression*. Then enter the dependent variable in the *dependent variable* field and the independent variables in the *independent variable* field and click *OK* or *Submit*. If you want to use the command line instead of the menu, in our example you would type the following command: *logistic purchase_yn image_size_laptop yellow age*. This



Fig. 11.10 Logistic regression with SPSS. *Source* Own graphic

command produces a table with the odds ratios Exp(*B*) for each independent variable (not the log-odds coefficients!), the *p*-values to indicate whether the variables are statistically significant, the log-likelihood value and the pseudo-R^2 value to summarize the fit of the model. Enter *estat gof* in the command line to assess the goodness of fit of the logistic model using the Hosmer–Lemeshow test. The *estat classification* command activates the classification table. After running the logistic regression, you can save the predicted probabilities for each observation in your dataset using the *predict prob, p* command. Screenshots and results are shown in Fig. 11.11.

11.9.3 Logistic Regression Calculations with R Commander

If you have not done so already, open the R Commander by typing *library(Rcmdr)* in the R console and load the dataset *Chapter9_mail_order_business_2.RData* (see supplementary material at https://doi.org/10.1007/978-3-031-78070-7_9). Access the logistic regression menu by choosing *Statistics* → *Fit models* → *Generalized linear model…*

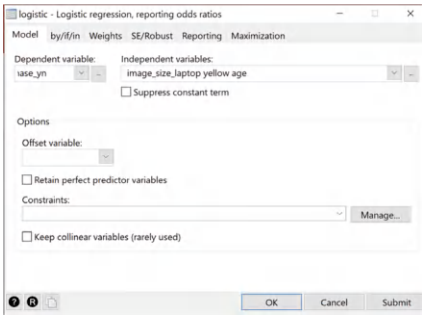
In the Generalized Linear Model window (see Fig. 11.12), select your dependent variable (purchase_yn) and your independent variables (image_size_laptop + yellow + age). Choose *Binomial* as the family and *logit* as the link function. Click *OK*.

By default, R outputs log-odds coefficients (logistic regression coefficients), *p*-values, and odds ratios for the independent variables (see Fig. 10.10). Finally, R implicitly performs the following syntax:

- GLM.1 < - glm(purchase_yn ~ image_size_laptop + yellow + age, family = binomial(logit), data = Dataset)
- summary(GLM.1)
- exp(coef(GLM.1)) # Exponentiated coefficients ("odds ratios").

The results of the likelihood ratio test, Mc Fadden R^2, and Hosmer–Lemeshow tests are not available from the R Commander menu and must be entered directly (Please refer to the commands marked in red at the bottom of Fig. 11.12):

1. to get the likelihood ratio test results:
   (a) GLM.1 < - glm(purchase_yn ~ image_size_laptop + yellow + age, family = binomial(logit), data = Dataset)
   (b) GLM.2 < - glm(purchase_yn ~ 1, family = binomial(logit), data = Dataset)
   (c) library(lmtest)
   (d) lrtest(GLM.2, GLM.1)
2. to get results for pseudo-R^2:
   (a) install.packages("pscl")
   (b) library(pscl)
   (c) pR2(GLM.1)




```
. logistic purchase_yn image_size_laptop yellow age
Logistic regression                               Number of obs     =        100
                                                  LR chi2(3)        =      41.05
                                                  Prob > chi2       =     0.0000
Log likelihood = -44.220931                       Pseudo R2         =     0.3170

------------------------------------------------------------------------------
      purchase_yn | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------+-----------------------------------------------------------
image_size_laptop |   1.062404   .0148413     4.33   0.000      1.03371    1.091894
           yellow |   .1821649   .1475803    -2.10   0.036      .0372277    .8913801
              age |   1.054587   .0472906     1.19   0.236      .9658557    1.15147
            _cons |   8.57e-06   .0000229    -4.37   0.000     4.57e-08     .001607
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Likelihood ratio test ($Chi^2 = 41.05$ with 3 df) is significant ($p < 0.005\%$). There is at least one independent variable that significantly improves the full model relative to the null model.

p-value: This effect is highly significant (p-value less than 0.5%)

The odds ratio shows that for each additional square centimetre in the size of the picture, the odds of buying a dress increase by 6.2%.

```
. estat gof
Logistic model for purchase_yn, goodness-of-fit test

       number of observations =       100
 number of covariate patterns =        99
              Pearson chi2(95) =     79.14
                   Prob > chi2 =      0.8795
```

The Hosmer-Lemeshow test yields a p-value of $p = 0.8795$, which is well above the 0.05 threshold. This indicates that the model fits the data well, as there is no significant differ-ence between the observed and predicted results.

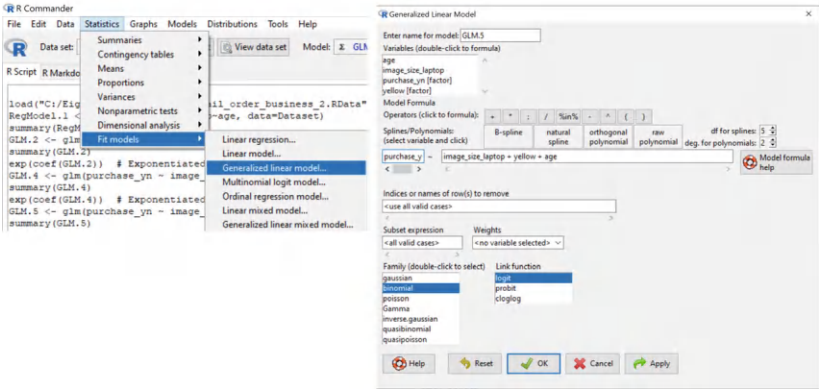
```
. estat classification
Logistic model for purchase_yn

             -------- True --------
Classified |         D             ~D  |      Total
-----------+--------------------------+-----------
     +     |        22             14  |         36
     -     |        13             51  |         64
-----------+--------------------------+-----------
   Total   |        35             65  |        100

Classified + if predicted Pr(D) >= .5
True D defined as purchase_yn != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   62.86%
Specificity                     Pr( -|~D)   78.46%
Positive predictive value       Pr( D| +)   61.11%
Negative predictive value       Pr(~D| -)   79.69%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   21.54%
False - rate for true D         Pr( -| D)   37.14%
False + rate for classified +   Pr(~D| +)   38.89%
False - rate for classified -   Pr( D| -)   20.31%
--------------------------------------------------
Correctly classified                        73.00%
--------------------------------------------------
```

Classification table: the model has an overall accuracy of 73%, meaning it correctly classifies 73% of both positive and negative cases. It has a sensitivity of 63%, meaning it identifies 22 out of 35 actual "Yes" purchases. The precision is 61%, meaning 22 out of 36 of the cases predicted as "Yes" are true positives.

Fig. 11.11 Logistic regression with Stata. *Source* Own graphic




```
glm(formula = purchase_yn ~ image_size_laptop + yellow + age,
    family = binomial(logit), data = Dataset)

Coefficients:
                  Estimate Std. Error z value  Pr(>|z|)
(Intercept)      -11.66706    2.67030  -4.369 0.0000125 ***
image_size_laptop  0.06053    0.01397   4.333 0.0000147 ***
yellow[T.yes]     -1.70284    0.81015  -2.102    0.0356 *
age                0.05315    0.04484   1.185    0.2359
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 129.489  on 99  degrees of freedom
Residual deviance:  88.442  on 96  degrees of freedom
AIC: 96.442

Number of Fisher Scoring iterations: 6

> exp(coef(GLM.5))  # Exponentiated coefficients ("odds ratios")
      (Intercept) image_size_laptop     yellow[T.yes]               age
   0.000008571574    1.062404121524     0.182164914463    1.054587205214

> lrtest(GLM.2, GLM.1)
Likelihood ratio test

Model 1: purchase_yn ~ 1
Model 2: purchase_yn ~ image_size_laptop + yellow + age
  #Df  LogLik Df  Chisq     Pr(>Chisq)
1   1 -64.745
2   4 -44.221  3 41.047 0.000000006389 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1> pR2(GLM.1)

> library(performance)
> performance_hosmer(GLM.1)
fitting null model for pseudo-r2
        llh      llhNull           G2     McFadden         r2ML         r2CU
-44.2209309 -64.7446639  41.0474660   0.3169950   0.3366647   0.4636789

> library(performance)
> performance_hosmer(GLM.1)
# Hosmer-Lemeshow Goodness-of-Fit Test

  Chi-squared: 4.991
           df: 8
      p-value: 0.759
```

p-value: effect is highly significant (p<0.01)

p-value: effect is significant (p<0.05)

The odds ratio shows that for each additional square centimetre in the size of the picture, the odds of buying a dress increase by 6.2%.

Likelihood ratio test ($Chi^2 = 41.05$ with 3 df) is significant ($p < 0.005\%$). There is at least one independent variable that significantly improves the full model relative to the null model.

Mc Fadden $R^2 = 0.317$

The Hosmer-Lemeshow test yields a p-value of $p = 0.759$, which is well above the 0.05 threshold. This indicates that the model fits the data well, as there is no significant difference between the observed and predicted results.

Fig. 11.12 Logistic regression with R Commander. *Source* Own graphic

3. to get Hosmer–Lemeshow test results:
   (a) install.package("performance")
   (b) library(performance)
   (c) performance_hosmer(GLM.1).

11.10 Chapter Exercises

Exercise 78
Answer the following multiple-choice questions. There is at least one correct answer.

1. What does the term "log-odds" represent in logistic regression? (a) The logarithm of the predicted probability; (b) the natural logarithm of the odds of the outcome; (c) the log-likelihood of the regression model; d) the difference between true and false positives.
2. If the odds ratio for a predictor is 1.5, how should it be read? (a) The odds increase by 1.5 times for every one-unit increase in the predictor; (b) the odds decrease by 1.5 times for every one-unit increase in the predictor; (c) the probability increases by 1.5% for every one-unit increase in the predictor; (d) the predictor has no impact on the odds.
3. What is the meaning of an odds ratio of 1 in logistic regression? (a) The predictor has no effect on the odds of the outcome; (b) the predictor doubles the odds of the outcome; (c) the predictor halves the odds of the outcome; (d) the predictor reverses the relationship.
4. The likelihood ratio test in logistic regression is used to compare: (a) the coefficients of two predictors; (b) the performance of two different models; (c) the observed versus expected frequencies; (d) the significance of individual predictors?
5. The Hosmer–Lemeshow test is used to evaluate: (a) the overall fit of the model; (b) the individual predictors; (c) multicollinearity; (d) goodness of fit of the logistic regression model?
6. A high p-value in the Hosmer–Lemeshow test ($p > 0.05$) indicates: (a) the model fits well; (b) the model fits poorly; (c) the predictors are collinear; (d) the pseudo R-square is low.
7. In logistic regression, which test helps to check for significant improvement when adding predictors to a model? (a) Likelihood ratio test; (b) Wald test; (c) Hosmer–Lemeshow test; (d) Breusch-Pagan test.
8. Which of the following is NOT a common effect of heteroscedasticity in logistic regression? (a) It can bias coefficient estimates; (b) it affects the interpretation of p-values; (c) it increases the variance of error terms; (d) it inflates the odds ratios.
9. Which of the following is NOT a key assumption of logistic regression? (a) The outcome variable is binary; (b) the observations are independent; (c) there

is a linear relationship between predictors and the log-odds; (d) the residuals must be normally distributed.

10. Which of the following is NOT a consequence of heteroscedasticity in logistic regression? (a) Standard errors become unreliable; (b) it violates the assumption of homoscedasticity; (c) it biases the coefficient estimates; (d) it increases the variance of the error terms.

Exercise 79

A company is interested in understanding which customer attributes—such as gender, age, regular customer status, and frequency of media contact—influence the likelihood of a purchase. All data are available in the *Chapter11_purchase* dataset (see supplementary material at https://doi.org/10.1007/978-3-031-78070-7_11). Interpret the results in Fig. 11.13.

11.11 Exercise Solutions

Solution 78

(b); (a); (a); (b); (d); (a); (a); (d); (d); (c).

Solution 79

The − 2 log-likelihood statistic improved from $LL_0 = 119.472$ for the null model to $LL_{full} = 120.293$ for the full model. This is a first indication that the predictors are making a significant contribution to the improvement of the model. This is also supported by the results of the likelihood ratio test, which shows a chi-square value of $\chi^2 = 18.997$ with four degrees of freedom and a p-value of $p < 0.001$, meaning that the predictors significantly provide important insights into customer decisions. Therefore, there is at least one independent variable that significantly improves the full model relative to the null model. The Cox & Snell $R^2 = 0.173$ and the Nagelkerke $R^2 = 0.231$ also suggest that the model has a small but reasonable level of predictive power.

The next step is to determine which predictors significantly improve the model. The variable *Individual Media Contacts* has a significant and positive effect on purchase behaviour. The coefficient (B) of 0.092 indicates that more media contacts of a client increase the likelihood of purchase. This effect is significant at $p = 0.015$. The odds ratio ($\text{Exp}(B) = 1.096$) shows that for each additional media contact, the odds of purchase increase by 9.6% ($0.096 = 1.096 - 1$). All other predictors are statistically not significant.

The Hosmer–Lemeshow test yields a p-value of $p = 0.703$, which is well above the 0.05 threshold. This indicates that the model fits the data well, as there is no significant difference between the observed and predicted results.

The model's classification accuracy provides a measure of how well it predicts customer behaviour. Overall, the model correctly classifies 69% of cases, meaning it accurately predicts whether a customer will go through with a purchase or not

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients: Constant	Gender	Age	Regular Customer [Y=1;N=0]	Individual Media Contacts
Step 1	1	120.293	-1.467	-.732	.027	.548	.062
	2	119.489	-1.797	-.844	.030	.653	.087
	3	119.472	-1.842	-.859	.030	.668	.092
	4	119.472	-1.843	-.859	.030	.669	.092

a. Method: Enter

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 138.469

d. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	18.997	4	<.001
	Block	18.997	4	<.001
	Model	18.997	4	<.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	119.472[a]	.173	.231

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	5.501	8	.703

Classification Table[a]

	Observed		Predicted: Purchase [Y=1;N=0] no purchase	purchase	Percentage Correct
Step 1	Purchase [Y=1;N=0]	no purchase	38	14	73.1
		purchase	17	31	64.6
	Overall Percentage				69.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Gender	-.859	.447	3.698	1	.054	.424	.176	1.017
	Age	.030	.020	2.169	1	.141	1.030	.990	1.071
	Regular Customer [Y=1;N=0]	.669	.460	2.109	1	.146	1.951	.792	4.810
	Individual Media Contacts	.092	.038	5.902	1	.015	1.096	1.018	1.181
	Constant	-1.843	.762	5.848	1	.016	.158		

a. Variable(s) entered on step 1: Gender, Age, Regular Customer [Y=1;N=0], Individual Media Contacts.

Correlation Matrix

		Constant	Image Size [in cm**2]	Colour of dress is yellow [Y=1,N=0]	Age [in years]
Step 1	Constant	1.000	-.855	.526	-.528
	Image Size [in cm**2]	-.855	1.000	-.561	.024
	Colour of dress is yellow [Y=1,N=0]	.526	-.561	1.000	-.183
	Age [in years]	-.528	.024	-.183	1.000

Fig. 11.13 Logistic regression Exercise 2. *Source* Own graphic

in 69 out of 100 cases. The model performs slightly better at predicting customers who did not purchase, with an accuracy of 73.1%, compared to customers who did make a purchase, with a sensitivity of 64.6%. This indicates that while the model is generally reliable, it is slightly more effective at identifying non-purchasers than purchasers. Nonetheless, the overall classification accuracy is strong and shows that the model provides valuable insights into purchasing behaviour.

Overall, it can be concluded that the model is satisfactory. In practice, the insignificant variables could have been subsequently removed. The company could invest in more frequent media contact, as this has been shown to directly contribute to a higher likelihood of purchase.

References

Cox, D. R., & Snell, E. J. (1989). *The analysis of binary data* (2nd ed.). Chapman and Hall.

Dalal, S. R., Fowlkes, E. B., & Hoadley, B. (1989). Risk analysis of the space shuttle: Pre-challenger prediction of failure. *Journal of the American Statistical Association, 84*(408), 945–957.

Hair, J. et al. (2006). *Multivariate Data Analysis* (6th ed.). Upper Saddle River, NJ: Prentice Hall International.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). Wiley.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis, 9*(2), 137–163.

McFadden, D. (1978). Quantitative methods for analyzing travel behaviour on individuals: Some recent developments. In D. Hensher & P. Stopher (Eds.), *Behavioral travel modelling* (pp. 279–318). Croom Helm.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*, 691–692.

Presidential Commission on the Space Shuttle Challenger Accident. (1986). *Report of the Presidential Commission on the space shuttle challenger accident* (Vol. I). Washington, DC: U.S. Government Printing Office.

Thompson, B. (2006). *Foundations of behavioral statistics. An insight-based approach.* The Guilford Press.



Time Series and Indices

12

In Chap. 10 we used a variety of independent variables to predict dress sales. All the trait values for sales (dependent variable) and for catalogue image size (independent variable) were recorded over the same period of time. Studies like these are called cross-sectional analyses. When the data is measured at successive time intervals, it is called a time series analysis or a longitudinal study. This type of study requires a time series in which data for independent and dependent variables are observed for specific points of time ($t = 1, \ldots, n$). In its simplest version, time is the only independent variable and is plotted on the x-axis. This kind of time series does nothing more than link variable data over different periods. Figure 12.1 shows an example with a graph of diesel fuel prices by year.

Frequently, time series studies involve a significantly more complicated state of affairs. Sometimes future demand does not depend on the time but on present or previous income. Let's look at an example. For the period t, the demand for a certain good y_t results from price (p_t), advertising expenses in the same period (a_t) and demand in the previous period (y_{t-1}). If the independent variable on the x-axis is not the time variable itself, but another independent variable bound to time, things become more difficult. For situations like these, see the helpful introductions offered by Greene (2017) and Wooldridge (2019).

The daily news bombards us with time series data: trends for things like unemployment, prices and economic growth. The announcement of new economic data is eagerly anticipated, and, when inauspicious (think: falling profits), can cause much distress (think: pearls of sweat beading on executives' foreheads). The reason time series have such a prominent role in the media is simple: they make discrete observations dynamic. Swoboda (1971, p. 96) aptly compares this process to film, which consists of individual pictures that produce a sense of motion when shown in rapid succession. Time series data are similar, as they allow us to recognize movements and trends and to project them into the future. Below we

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*,
Springer Texts in Business and Economics,
https://doi.org/10.1007/978-3-031-78070-7_12

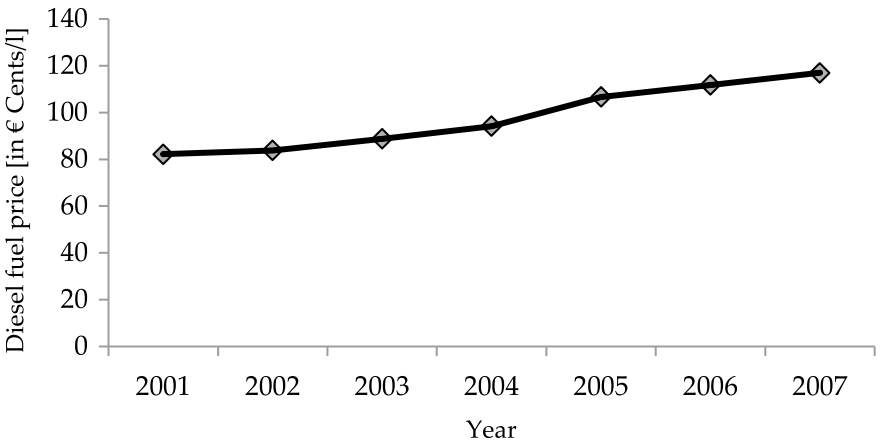


Fig. 12.1 Diesel fuel prices by year, 2001–2007

investigate the most frequently used technique for measuring dynamic phenomena: index figures.

12.1 Price Indices

The simplest way to express price changes over time is to indicate the *(unweighted) percentage price change* in one reporting period compared with an earlier one, known as the base period. Table 12.1 shows the average yearly prices for diesel and petrol in Germany. To find out the percentage increase for diesel fuel in the 2007 *reporting period* compared with the 2001 base period, we calculate what is called a *price relative*:

$$P^*_{\text{base year}=0,\ \text{reporting year}=t} = \frac{\text{Price in reporting year } (p_t)}{\text{Price in base year } (p_0)} \tag{12.1}$$

$$P^*_{2001,2007} = \frac{p_{2007}}{p_{2001}} = \frac{117.0}{82.2} = 1.42 \tag{12.2}$$

The price of diesel in 2007 was around 42% higher than in 2001. In principle, price relatives can be calculated for every possible base year and reporting year combination. Price relatives for the base year 2005 are also indicated in Table 12.1. According to these figures, the 2007 price increased by 10% over that of 2005, while the price in 2001 still lay 23% (=1.00 − 0.77) below the price from the base year 2005.

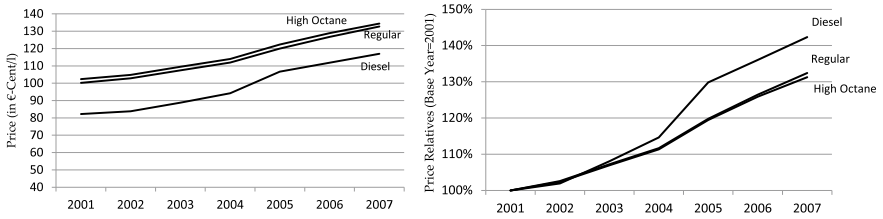
Table 12.1 Average prices for diesel and petrol in Germany

Price in cents/l	2001	2002	2003	2004	2005	2006	2007
High octane	102.4	104.8	109.5	114.0	122.3	128.9	134.4
Regular	100.2	102.8	107.4	111.9	120.0	126.7	132.7
Diesel	82.2	83.8	88.8	94.2	106.7	111.8	117.0
Price relative for diesel (base year 2001)	1.00	1.02	1.08	1.15	1.30	1.36	1.42
Price relative for diesel (base year 2005)	0.77	0.79	0.83	0.88	1.00	1.05	1.10
Sales in 1000*t* and (share of consumption in %)	2001	2002	2003	2004	2005	2006	2007
High octane	18,979 (33.6%)	18,785 (33.7%)	18,140 (33.7%)	17,642 (32.7%)	16,870 (32.5%)	16,068 (31.5%)	15,718 (31.2%)
Regular	8970 (15.9%)	8409 (15.1%)	7710 (14.3%)	7395 (13.7%)	6561 (12.6%)	6181 (12.1%)	5574 (11.1%)
Diesel	28,545 (50.5%)	28,631 (51.3%)	27,944 (52.0%)	28,920 (53.6%)	28,531 (54.9%)	28,765 (56.4%)	29,059 (57.7%)
All fuels	56,494	55,825	53,794	53,957	51,962	51,014	50,351
Quantity relative for diesel (base year 2001)	1.00	1.00	0.98	1.01	1.00	1.01	1.02

Source Association of the German Petroleum Industry. Based on the author's calculations

This fuel example illustrates the advantages of indexing. Index series make dynamic developments comparable and push absolute differences into the background. If one compares the absolute prices for diesel, high octane and regular over time (see Fig. 12.2, part 1) with their index series from the base year 2001 (see Fig. 12.2, part 2), the varying price dynamic becomes immediately apparent. The price boost for diesel—hard to discern in part 1—is pushed to the fore by the indexing (part 2), while absolute price differences can no longer be inferred from the figure.

To calculate the change in price between two years when neither is a base year, the base for the price relatives must be shifted. Let us consider the diesel fuel price



Part 1 Part 2

Fig. 12.2 Fuel prices over time

relatives for the base year 2001. What is the change of price between 2004 and 2007? At first glance, you might think the answer is 27% (1.42 − 1.15). But the correct answer is not 27% but 27 *percentage points* relative to the base year 2001. Here it would better to shift the base[1] for 2004 by dividing the old series of price relatives (base year: 2001) by the price relative of 2004:

$$P^*_{\text{new base year},\, t} = \frac{P^*_{\text{old base year}}}{P^*_{\text{new base year}}} \tag{12.3}$$

Now we can see that the percentage change between 2004 and 2007 is 23%:

$$P^*_{2004,2007} = \frac{P^*_{2001,2007}}{P^*_{2001,2004}} = \frac{1.42}{1.15} = 1.23 \tag{12.4}$$

This price relative—an unweighted percentage price change of a homogenous product—no longer applies when heterogeneous product groups exist. Let us leave aside this particular result (which is probably only interesting for drivers of diesel vehicles) and instead calculate how the prices of all fuel types (diesel, regular, and high octane) developed in total. For this case, we must use the so-called *weighted aggregated price index*. This index can determine the price trend of a product group, a branch or an entire national economy using a predefined market basket. The German consumer price index determined by the Federal Statistical Office of Germany consists of some 700 everyday products whose prices are collected monthly. The prices are weighted based on average consumption in a representative German household. For instance, rent (not including heating) has a share of 20.3% in the consumer price index. Of course, individual choices can lead to different rates of price increase than those experienced by the *average consumer*.[2]

The comparability of prices in different periods is ensured only if the contents of the market basket and the weights of its products remain the same. This is called a *fixed-weighted aggregated price index*. For the above example, the question is not how demand and price change in total but how the price for a specific quantity of diesel, regular, and super octane changes relative to the base year. In practice, of course, consumption does not remain constant over time. In the period of observation, for example, the share of diesel consumption rose continuously, while the share of consumption for the other fuels sank. There are two index options that use fixed weights:

[1] Strictly speaking, a base shift need only be undertaken when the market basket linked to the time series is changed (see Sect. 12.5).

[2] For more, see the information on consumer price statistics at https://www.destatis.de/EN/Home/_node.html. The site calculates the price increase rate for individuals with a personal inflation calculator.

1. The first type of index is called the *Laspeyres index*.[3] It is probably the best-known index and is used by the Federal Statistical Office of Germany and by many other Statistical Offices in the world. It identifies weights from average consumption in the base period ($t = 0$):

$$P_{0,t}^{L} = \frac{\sum_{i=1}^{n} \frac{p_{i,t}}{p_{i,0}} \cdot p_{i,0} \cdot q_{i,0}}{\sum_{i=1}^{n} p_{i,0} \cdot q_{i,0}} = \frac{\sum_{i=1}^{n} p_{i,t} \cdot q_{i,0}}{\sum_{i=1}^{n} p_{i,0} \cdot q_{i,0}} \tag{12.5}$$

Usually, Laspeyres index figures are multiplied by 100 or, as with the German DAX stock market index, by 1000. For example, the Federal Statistical Office of Germany expresses the inflation rate as $P_{0,t}^{L}$ multiplied by 100[4]:

$$100 \cdot P_{0,t}^{L} = 100 \cdot \frac{\sum_{i=1}^{n} p_{i,t} \cdot q_{i,0}}{\sum_{i=1}^{n} p_{i,0} \cdot q_{i,0}} \tag{12.6}$$

In the example with diesel and petrol, total demand is 28,545,000 tons of diesel ($q_{\text{diesel},2001}$), 8,970,000 tons of regular ($q_{\text{regular},2001}$) and 18,979,000 tons of high octane ($q_{\text{high-octane},2001}$) in 2001. Now imagine we wanted to know how the total fuel price in 2007 would have developed relative to 2001 if the weights—i.e. the share of consumption for each of the fuels—remained the same compared to 2001. First we weight the 2007 prices for diesel, regular and high octane with the average amounts consumed in 2001 ($q_{\text{i},2001}$) and add them together. This total goes in the numerator. Next we weight the amounts consumed in 2001 with the prices of 2001 ($p_{\text{i},2001}$) and add them together. This total goes in the denominator. Now we have the following weighted percentage change in price:

$$P_{\text{base year,report year}}^{L} = \frac{\sum_{i=1}^{n} p_{i,\text{report year}} \cdot q_{i,\text{ base year}}}{\sum_{i=1}^{n} p_{i,\text{ base year}} \cdot q_{i,\text{ base year}}} = P_{0,t}^{L} = \frac{\sum_{i=1}^{n} p_{i,t} \cdot q_{i,0}}{\sum_{i=1}^{n} p_{i,0} \cdot q_{i,0}} \tag{12.7}$$

$$P_{2001,2007}^{L} = \frac{134.4 \cdot 18,979 + 132.7 \cdot 8970 + 117.0 \cdot 28,545}{102.4 \cdot 18,979 + 100.2 \cdot 8970 + 82.2 \cdot 28,545} = 1.3647 \tag{12.8}$$

Alternatively, instead of the absolute amount consumed, we can use the share of consumption, since this expands the fraction only by the inverse of total tons consumed in the base year:

$$P_{0,t}^{L} = \frac{\sum_{i=1}^{n} p_{i,t} \cdot q_{i,0}}{\sum_{i=1}^{n} p_{i,0} \cdot q_{i,0}} = \frac{\sum_{i=1}^{n} p_{i,t} \cdot \frac{q_{i,0}}{\sum_{j=1}^{n} q_{j,0}}}{\sum_{i=1}^{n} p_{i,0} \cdot \frac{q_{i,0}}{\sum_{j=1}^{n} q_{j,0}}} = \frac{\sum_{i=1}^{n} p_{i,t} \cdot f_{q_{i,0}}}{\sum_{i=1}^{n} p_{i,0} \cdot f_{q_{i,0}}} \tag{12.9}$$

[3] Ernst Louis Etienne Laspeyres (1834–1913) was a court advisor and professor for economics at the universities of Basel, Riga, Dorpat, Karlsruhe, and Giessen. He got his name from his Portuguese ancestors, who came to Germany by way of France. He first used his price index to measure price trends in Hamburg.

[4] Later in this section, the index values are multiplied by 100 only when indicated.

$$P^{L}_{2001,2007} = \frac{134.4 \cdot 33.6\% + 132.7 \cdot 15.9\% + 117.0 \cdot 50.5\%}{102.4 \cdot 33.6\% + 100.2 \cdot 15.9\% + 82.2 \cdot 50.5\%} = 1.3647 \quad (12.10)$$

This tells us that price levels rose by 36.5% from 2001 to 2007 assuming that the shares of consumption for each of the fuels remained the same compared to the base year 2001.

When measuring price changes with the Laspeyres index, one should be aware of the problems that can arise. Some are general problems affecting all weighted aggregated indices. The first is the representativeness of market basket items. For instance, if the price of diesel increases and the price of petrol stays the same, the index we created will indicate that the average price of fuel has increased, but this price increase won't affect car drivers who don't buy diesel. Similarly, a homeowner won't feel a rise in the consumer price index caused by climbing rent prices. A renter might argue, by contrast, that the rise indicated in the consumer price index is nowhere close to actual price increases. The greater the difference between forms of consumption, the more often this problem occurs. Of course, the purpose of an aggregated index is not to express the personal price changes experienced by Mr. Jones or Ms. Smith. The point is to measure the sum of expenditures of all households and from them derive the average shares of consumption. This figure identifies neither the price changes experienced by individual household nor the price changes experienced by rich households and poor households. It might be the case that there is no household in the whole economy whose consumption corresponds exactly to that of the "representative" household. The consumer price index is nevertheless consistent for the total sum of households. To get around this problem, the Federal Statistical Office of Germany has created an internet site where people can calculate their individual rates of price increase by indicating shares of total expenditure for their own household.

Another general problem of indices relates to retail location and product quality. Price differences can occur not only between regions but even between city districts, where, say, the price of 250 g of butter can vary by tens of cents, depending on its quality, the type of retail location, and the average income of consumers. As a result, something as minor as changing stores can create significant artificial fluctuations in price. This is why officials who collect prices for the consumer price index are required to use the same locations and product qualities whenever possible.

Aside from such general problems exhibited by aggregate indices, the Laspeyres index has its own oddities caused by changing consumer habits. If the locations people shop at change significantly after the market basket is set for the base period (e.g. from small retailers to warehouse stores), the price changes indicated by the index may differ from actual changes. The same thing can happen if consumers begin substituting consumer goods in the market basket with non-indexed items or if product shares in the total of the consumer expenditures covered by the index change. Especially in rapidly changing sectors such as the digital industry, comparisons with base period hardware prices can be misleading. Changing consumer preferences create a problem for the fixed weighting of market basket items

from a distant base period. To isolate actual price changes from changes in quality, National Statistical Offices change the contents of the consumer price index basket frequently—typically about every five years. In 2015, for instance, the Federal Statistical Office of Germany changed the base year from 2010 to 2015.

2. The second option for a fixed-weighted index is the *Paasche index*.[5] It solves the problem of out-of-date market baskets by setting a new market basket for every period. In this way, the market basket product shares in the total of the consumer expenditures covered by the index precisely reflect the year under observation. Yet creating a new market basket each year is time consuming, and is one of the disadvantages of this method. The Paasche index compares current period expenditure with a hypothetical base period expenditure. This hypothetical value estimates the value one would have had to pay for a current market basket during the base period. The total expenditure of the current period and hypothetical expenditure of the base period form the Paasche index's numerator and denominator, respectively:

$$P^P_{\text{base year, report year}} = \frac{\sum_{i=1}^{n} p_{i,\text{ report year}} \cdot q_{i,\text{report year}}}{\sum_{i=1}^{n} p_{i,\text{base year}} \cdot q_{i,\text{ report year}}} = P^P_{0,t} = \frac{\sum_{i=1}^{n} p_{i,t} \cdot q_{i,t}}{\sum_{i=1}^{n} p_{i,0} \cdot q_{i,t}} \tag{12.11}$$

In the following example, we again calculate the rise in fuel prices between 2001 and 2007, this time using the Paasche index. In the 2007 period, the fuel market basket consists of 29,059,000 tons of diesel ($q_{\text{Diesel,2007}}$), 5,574,000 tons of regular ($q_{\text{regular,2007)}}$ and 15,718,000 tons of high octane ($q_{\text{high-octane,2007}}$). The total expenditure results from weighting fuel prices for diesel, regular and high octane by their consumption levels and then adding them together (numerator). The total expenditure is then related to the shares of total expenditure in the reporting period as measured by base period prices ($p_{i,2001}$) (denominator). This produces the following:

$$P^P_{2001,2007} = \frac{\sum_{i=1}^{n} p_{i,2007} \cdot q_{i,2007}}{\sum_{i=1}^{n} p_{i,2001} \cdot q_{i,2007}} \tag{12.12}$$

$$P^P_{2001,2007} = \frac{134.4 \cdot 15{,}718 + 132.7 \cdot 5574 + 117.0 \cdot 29{,}059}{102.4 \cdot 15{,}718 + 100.2 \cdot 5574 + 82.2 \cdot 29{,}059} = 1.3721 \tag{12.13}$$

This is then weighted by the shares of the total expenditure:

$$P^P_{2001,2007} = \frac{134.4 \cdot 31.2\% + 132.7 \cdot 11.1\% + 117.0 \cdot 57.7\%}{102.4 \cdot 31.2\% + 100.2 \cdot 11.1\% + 82.2 \cdot 57.7\%} = 1.3721 \tag{12.14}$$

[5] The German economist Hermann Paasche (1851–1925) taught at universities in Aachen, Rostock, Marburg and Berlin. In addition to his achievements in economics, Paasche was an engaged member of the Reichstag, serving as its Vice President for more than a decade.

Based on this calculation, price levels rose by 37.2% from 2001 to 2007 assuming that the shares of expenditure for each of the fuels remained the same compared to the reporting period 2007. Compared with the results of the Laspeyres index (36.5%), the inflation rate of the Paasche index is higher. This means that consumers shifted their demand to products whose prices rose at a higher-than-average rate. Though diesel is still cheaper than other fuels in absolute terms—this ultimately explains the increase of shares in total expenditure from 50.5 to 57.7% between 2001 and 2007—its price increased by around 42%, while the prices of regular and high octane increased by 32% and 31%, respectively. Accordingly, during the reporting period, consumers tended to purchase more products whose prices increased by a higher-than-average rate than consumers did during the base period.[6] In the opposite case, when the inflation rate indicated by the Laspeyres index is larger than that indicated by the Paasche index, demand develops in favour of products whose prices increase at a lower-than-average rate. In this case, consumers substitute products whose prices increase at a higher-than-average rate with those whose prices increase at a lower-than-average rate. On account of this economic rationality, the Laspeyres index is almost always larger than the Paasche index, even if the needn't always be the case, as our example shows. With some consumer goods, especially expensive lifestyle products, demand increases though prices increase at a higher-than-average rate. To sum up, the Laspeyres price index is higher than the Paasche index when price changes and consumption changes negatively correlate; it is lower than the Paasche index when price changes and consumption changes positively correlate (see Rinne, 2008, p. 106).

Because the indices produce different results, Irving Fisher (1867–1947) proposed calculating the geometric mean of the two values, resulting in the so-called *Fisher index*:

$$P_{0,t}^{F} = \sqrt{P_{0,t}^{L} \cdot P_{0,t}^{P}} \tag{12.15}$$

This index seeks to find a "diplomatic solution" to conflicting approaches, but it lacks a clear market basket concept, as it relies on different baskets with different products and weights. The Paasche index, too, faces the general problem having to define anew the shares of the total expenditure covered by the market basket each year, which ultimately requires a recalculation of inflation rates, including those of past years. This means that past inflation rates do not remain fixed but change depending on the current market basket.

[6] The shift in expenditure is also expressed by the rise of new registrations for diesel vehicles in Germany (from 34.5 to 47.8%) and in Europe (from 36.7 to 53.6%) (ACEA, European Automobile Manufacturers' Association: http://www.acea.be/index.php/collection/statistics).

12.2 Quantity Indices

Next to the price index are a number of other important indices, of which the quantity index is the most important. Just as with the simple price relative, a change in the quantity of a homogenous product can be expressed by an unweighted quantity relative. Table 12.1 shows the quantity relative for the change in diesel sales:

$$Q^*_{0,t} = \frac{\text{Quantity in the report year } (q_t)}{\text{Quantity in the base year } (q_0)} \tag{12.16}$$

$$Q^*_{2001,2003} = \frac{q_{t=2003}}{q_{t=2001}} = \frac{27{,}944}{28{,}545} = 0.98 \tag{12.17}$$

Accordingly, the demand for diesel declined by 2% (= 1.00 − 0.98) from 2001 to 2003. If we now put aside homogenous products and consider instead quantity changes for a market basket at constant prices, we must use the *weighted aggregated quantity index*. Here too, we can use either the Laspeyres index or the Paasche index, though both follow the same basic idea.

How do the weighted quantities of a defined market basket change between a base period and a given observation period, assuming prices remain constant? The only difference between the Laspeyres quantity index and the Paasche quantity index is that the former presumes a market basket defined in the base period and its constant item prices, while the latter serves as the basis for the market basket and the constant prices of the reporting period. With both concepts, we may only use absolute quantities from the market basket, not relative values:

$$\text{Laspeyres quantity index: } Q^L_{0,t} = \frac{\sum_{i=1}^{n} q_{i,t} \cdot p_{i,0}}{\sum_{i=1}^{n} q_{i,0} \cdot p_{i,0}} \tag{12.18}$$

$$\text{Paasche quantity index: } Q^P_{0,t} = \frac{\sum_{i=1}^{n} q_{i,t} \cdot p_{i,t}}{\sum_{i=1}^{n} q_{i,0} \cdot p_{i,t}} \tag{12.19}$$

$$\text{Fisher quantity index: } Q^F_{0,t} = \sqrt{Q^L_{0,t} \cdot Q^P_{0,t}} \tag{12.20}$$

Important applications for quantity indices include trends in industrial production and capacity workload. Quantity indices can also be used to answer other questions. For instance, how did diesel sales between 2001 and 2007 develop with constant prices from 2001 versus constant prices from 2007 (see Table 12.1)?

Laspeyres quantity index (constant prices from 2001):

$$Q^L_{2001,2007} = \frac{\sum_{i=1}^{n} q_{i,2007} \cdot p_{i,2001}}{\sum_{i=1}^{n} q_{i,2001} \cdot p_{i,2001}} \tag{12.21}$$

$$Q^L_{2001,2007} = \frac{15{,}718 \cdot 102.4 + 5574 \cdot 100.2 + 29{,}059 \cdot 82.2}{18{,}979 \cdot 102.4 + 8970 \cdot 100.2 + 28{,}545 \cdot 82.2} = 0.8782 \tag{12.22}$$

Paasche quantity index (constant prices from 2007):

$$Q^P_{2001,2007} = \frac{\sum_{i=1}^{n} q_{i,2007} \cdot p_{i,2007}}{\sum_{i=1}^{n} q_{i,2001} \cdot p_{i,2007}} \tag{12.23}$$

$$Q^P_{2001,2007} = \frac{15{,}718 \cdot 134.4 + 5574 \cdot 132.7 + 29{,}059 \cdot 117}{18{,}979 \cdot 134.4 + 8970 \cdot 132.7 + 28{,}545 \cdot 117} = 0.8830 \tag{12.24}$$

Diesel sales in 2007 weighted by 2001 base period prices (Laspeyres quantity index) compared with those of 2001 declined by 12.2% (= 1.00 − 0.8782), while 2007 diesel sales weighted by prices of the 2007 observation period declined by (1.00 − 0.883 =) 11.7% (Paasche quantity index). Here too, the values of the quantity indices differ.

12.3 Value Indices (Sales Indices)

After identifying indices for price and quantity, it makes sense to calculate a value index for the market basket. Ultimately, the value of a consumer good is nothing more than the mathematical product of price and quantity. Interestingly, the *value index* (frequently called the *sales index*) can be derived neither from the product of the Laspeyres price and quantity indices alone nor from the product of the Paasche price and quantity indices alone.[7] Only the product of the Fisher price and quantity indices produces the correct value index. Alternatively, one can multiply the Paasche quantity index by the Laspeyres price index or the Laspeyres quantity index by the Paasche price index:

$$V_{0,t} = \frac{\sum_{i=1}^{n} p_{i,t} \cdot q_{i,t}}{\sum_{i=1}^{n} p_{i,0} \cdot q_{i,0}} = Q^F_{0,t} \cdot P^F_{0,t} = Q^L_{0,t} \cdot P^P_{0,t} = Q^P_{0,t} \cdot P^L_{0,t} \tag{12.25}$$

According to this equation, fuel sales in 2007 rose by 20.5% relative to those of 2001. The calculations are as follows:

$$V_{2001,2007} = Q^L_{2001,2007} \cdot P^P_{2001,2007} = 0.8782 \cdot 1.3721 = 1.2050, \text{ or} \tag{12.26}$$

$$V_{2001,2007} = Q^P_{2001,2007} \cdot P^L_{2001,2007} = 0.8830 \cdot 1.3647 = 1.2050 \tag{12.27}$$

[7] $V_{0,t} = \frac{\sum_{i=1}^{n} p_{i,t} \cdot q_{i,t}}{\sum_{i=1}^{n} p_{i,0} \cdot q_{i,0}} \neq P^L_{0,t} \cdot Q^L_{0,t} = \frac{\sum_{i=1}^{n} p_{i,t} \cdot q_{i,0}}{\sum_{i=1}^{n} p_{i,0} \cdot q_{i,0}} \cdot \frac{\sum_{i=1}^{n} q_{i,t} \cdot p_{i,0}}{\sum_{i=1}^{n} q_{i,0} \cdot p_{i,0}}$

12.4 Deflating Time Series by Price Indices

An important task of price indices is to adjust time series for inflation. Many economic times series—gross national product, company sales, warehouse stock—reflect changes in a given monetary unit, often indicating a rising trend. This can point to a real growth in quantity, but it can also indicate hidden inflation-based nominal growth, which may also be associated with a decline in quantity. Frequently, increases in both quantity and price are behind increases in value.

For these reasons, managers are interested in real parameter changes adjusted for inflation, which express value trends at constant prices. Table 12.2 provides sample trends for average employee salaries at two companies, each in a different country with different inflation rates. Compared with the base year, the nominal salary in company one increased by 0.5% (= 106.5 − 106.0) from 2013 to 2014. But the inflation rate for this period was 1.5% (= 105.5 − 104.0) compared with the base year. If factors out inflation with the help of the price index (compared with the base year), then the average salary declined by 1%. Adjustments for inflation are made by dividing nominal values by the price index. For the real (inflation-adjusted) average salary in 2014 L_t^{real} [in €], we thus find:

$$L_t^{\text{real}} = \frac{L_t^{\text{nominal}}}{P_{0,t}^L} \rightarrow L_{2014}^{\text{real}} = \frac{L_{2014}^{\text{nominal}}}{P_{0,2014}^L} = \frac{1917.00}{1.055} = €\,1817.06 \tag{12.28}$$

In 2013 the real average salary was €1834.62 per month (see Table 12.2). This means that, in 2014, employees lost some purchasing power compared to 2013. While nominal average monthly salaries between 2010 and 2018 increased by

Table 12.2 Sample salary trends for two companies

Year	Company 1					Company 2					
	Nominal salary		Price	Real salary		Nominal salary		Price	Real salary		
	[in €]	Index [2000 = 100]	Index [2000 = 100]	[in €]	Index [2000 = 100]	[in €]	Index [2002 = 100]	Index [2002 = 100]	[in €]	Index [2002 = 100]	Index [2000 = 100]
2010	1800.00	100.0	100.0	1800.00	100.0	1850.00	98.3	99.0	1868.69	99.3	100.0
2011	1854.00	103.0	102.0	1817.65	101.0	1868.50	99.3	99.7	1874.12	99.6	100.3
2012	1845.00	102.5	103.0	1791.26	99.5	1881.45	100.0	100.0	1881.45	100.0	100.7
2013	1908.00	106.0	104.0	1834.62	101.9	1868.50	99.3	101.0	1850.00	98.3	99.0
2014	1917.00	106.5	105.5	1817.06	100.9	1877.75	99.8	102.5	1831.95	97.4	98.0
2015	1926.00	107.0	106.5	1808.45	100.5	1951.75	103.7	103.0	1894.90	100.7	101.4
2016	1962.00	109.0	108.0	1816.67	100.9	1979.50	105.2	103.0	1921.84	102.1	102.8
2017	1998.00	111.0	109.0	1833.03	101.8	1998.00	106.2	103.5	1930.43	102.6	103.3
2018	2025.00	112.5	109.5	1849.32	102.7	2025.75	107.7	104.0	1947.84	103.5	104.2

Source Author's research

12.5%, from €1800 to €2025, the real salary in 2018 was only

$$Y_{2018}^{\text{real}} = \frac{2025.00}{1.095} = €1849.32 \tag{12.29}$$

an increase of 2.7%. It should be noted that real values always change dependent on the base year of the price index. Hence, comparisons of real values always need to be anchored to the base year, and presented in terms of indexed values instead of absolute values. See, for instance, the last column in Table 12.2.

12.5 Shifting Bases and Chaining Indices

As I describe above, the Federal Statistical Office of Germany prepares a new market basket about every five years. The point is to take into account large changes in product markets. Strictly speaking, a measurement of price and quantity indices is only possible when based on the same market basket. Hence, over longer time series, it is impossible to calculate inflation or adjust for inflation, as product markets undergo dynamic changes. This is where base shifts and chained indices come into play. We already learned about the base shift technique in Sect. 12.1, where we shifted the price relative of diesel fuel from the base year 2001 to the new base year 2004 by dividing the price relative of 2001 by the price relative of 2004. We can proceed analogously with any new base year (τ) for any index series. Index values for all years change according to the following formula:

$$I_{\tau,t}^{\text{new}} = \frac{I_{0,t}^{\text{old}}}{I_{0,\tau}^{\text{old}}} \tag{12.30}$$

Let us consider the example from Table 12.2. The index for the change of real income values in company #2 is based on 2012 (see the second-to-last column). If we now want to base this index series on the year 2010 so as to compare it with the corresponding index series of company #1, we must divide every index value of company #2 by the index value for 2010. This produces the final column in Table 12.2. Although the nominal income for company #2 has risen less than in company #1, its real increase is 4.2%, which is greater than the real increase of company #1 (2.7%).

The technique of chaining indices allows indices with different and time-restricted market baskets to be joined, forming one long index series. The only requirement is that each of the series to be chained overlaps with its neighbour in an observation period (τ). For the *forward extrapolation*, the index with older observations (I^1 between periods zero and τ) remains unchanged and the base of the younger overlapping index series (I^2) shifts to the older index. We do this by multiplying the values of the younger index series with the overlapping value of the older index series (at time τ):

$$\text{Forward extrapolation: } \tilde{I}_{0,t} = \begin{cases} I^1_{0,t} & \text{for } t \leq \tau \\ I^1_{0,\tau} \cdot I^2_{\tau,t} & \text{for } t > \tau \end{cases} \tag{12.31}$$

With the *backward extrapolation*, the index with the younger observations (I^2 starting out time τ) remains unchanged and the values of the older overlapping index series (I^1) are divided by the overlapping value of the younger index (at time τ):

$$\text{Backward extrapolation: } \tilde{I}_{0,t} = \begin{cases} \frac{I^1_{0,\tau}}{I^2_{\tau,t}} & \text{for } t < \tau \\ I^2_{\tau,t} & \text{for } t \geq \tau \end{cases} \tag{12.32}$$

If more than two index series are joined, we must gradually join the oldest series to the youngest series in the forward extrapolation and the youngest series to the oldest series in the backward extrapolation. Table 12.3 gives a sample of chain indices for backward and forward extrapolations.

Table 12.3 Chain indices for forward and backward extrapolations

		2015	2016	2017	2018	2019
Index 1		1.05	1.06			
Index 2			1.00	1.4	1.05	
Index 3					1.00	1.01
Chain index	Backward extrapolation	1.05/ $(1.06 \cdot 1.05)$ =	1.00/ 1.05 =	1.04/1.05 =	1.00 =	1.01 =
		0.94	0.95	0.99	1.0	1.01
	Forward extrapolation	1.05 =	1.06 =	$1.06 \cdot 1.04$ =	$1.06 \cdot 1.05$ =	$1.06 \cdot 1.05 \cdot 1.01$ =
		1.05	1.06	1.10	1.11	1.12

12.6 Chapter Exercises

Exercise 80
The following table presents price and sales trends for consumer goods A, B, C, and D in years 1 and 3.

Good	Price 1	Price 1	Price 3	Price 3
A	6	22	8	23
B	27	4	28	5
C	14	7	13	10
D	35	3	42	3

(a) Calculate the Laspeyres price and quantity indices for reporting year 3 using base year 1. Interpret your results.
(b) Calculate the Paasche price and quantity indices for reporting year 3 using base year 1. Interpret your results.
(c) Why is the inflation indicated by the Paasche index usually lower?
(d) Calculate the Fisher price and quantity indices for reporting year 3 using base year 1.
(e) Calculate and interpret the value index for reporting year 3 using base year 1.
(f) What is the per cent of annual price increase after calculating the Laspeyres price index?

Exercise 81
You are given the following information:

	2019	2020	2021	2022	2023
Nominal values	$100,000	$102,000	$105,060	$110,313	$114,726
Nominal value index [2019 = 100]					
Real values					
Real value index [2019 = 100]					
Price index 1 [2018 = 100]	101.00	102.00	102.50		
Price index 2 [2021 = 100]			100.00	103.00	103.50
Price index 3 [2018 = 100]					
Price index 4 [2019 = 100]					

(a) Calculate the nominal value index [2019 = 100].

(b) Chain the price trends for base year 2018.
(c) With the resulting index series, shift the base to 2019.
(d) Calculate the real value changes and the real value index for the base year 2019.

12.7 Exercise Solutions

Solution 80

Good	Price 1	Quantity 1	Price 3	Quantity 3	$p_3 \cdot q_1$ 0000	$p_1 \cdot q_1$ 0000	$p_3 \cdot q_3$ 0000	$p_1 \cdot q_3$ 0000
A	6	22	8	23	176	132	184	138
B	27	4	28	5	112	108	140	135
C	14	7	13	10	91	98	130	140
D	35	3	42	3	126	105	126	105
					505	443	580	518

(a) $P_{1,3}^{L} = \frac{\sum_{i=1}^{4} p_{i,3} \cdot q_{i,1}}{\sum_{i=1}^{4} p_{i,1} \cdot q_{i,1}} = \frac{(8 \cdot 22)+(28 \cdot 4)+(13 \cdot 7)+(42 \cdot 3)}{(6 \cdot 22)+(27 \cdot 4)+(14 \cdot 7)+(35 \cdot 3)} = \frac{505}{443} = 1.14$

(b) $Q_{1,3}^{L} = \frac{\sum_{i=1}^{4} q_{i,3} \cdot p_{i,1}}{\sum_{i=1}^{4} q_{i,1} \cdot p_{i,1}} = \frac{(23 \cdot 6)+(5 \cdot 27)+(10 \cdot 14)+(3 \cdot 35)}{(22 \cdot 6)+(4 \cdot 27)+(7 \cdot 14)+(3 \cdot 35)} = \frac{518}{443} = 1.17$

(c) The inflation rate between the two years is 14%. During the same period, sales of the four goods assessed with the prices of the first year increased by 17%.

(d) $P_{1,3}^{P} = \frac{\sum_{i=1}^{n} p_{i,3} \cdot q_{i,3}}{\sum_{i=1}^{4} p_{i,1} \cdot q_{i,3}} = \frac{(8 \cdot 23)+(28 \cdot 5)+(13 \cdot 10)+(42 \cdot 3)}{(6 \cdot 23)+(27 \cdot 5)+(14 \cdot 10)+(35 \cdot 3)} = \frac{580}{518} = 1.12$

(e) $Q_{1,3}^{P} = \frac{\sum_{i=1}^{4} q_{i,3} \cdot p_{i,3}}{\sum_{i=1}^{4} q_{i,1} \cdot p_{i,3}} = \frac{(23 \cdot 8)+(5 \cdot 28)+(10 \cdot 13)+(3 \cdot 42)}{(22 \cdot 8)+(4 \cdot 28)+(7 \cdot 13)+(3 \cdot 42)} = \frac{580}{505} = 1.15$

(f) The inflation rate between the two years is 12%. During the same period, sales of the four goods assessed with the prices of the third year increased by 15%.

(g) The inflation shown by the Paasche index is lower because demand shifts in favour of products with lower-than-average rising prices. In the given case, the consumption shifts (substitution) in favour of products B and C. The price of product B rose by only 3.7%—a lower-than-average rate—while the price of product C sank by 7.1% (substitution of products with greater-than-average rising prices through products B and C).

(h) $P_{1,3}^{F} = \sqrt{P_{1,3}^{L} \cdot P_{1,3}^{P}} = \sqrt{1.14 \cdot 1.12} = 1.13$

(i) $Q_{1,3}^{F} = \sqrt{Q_{1,3}^{L} \cdot Q_{1,3}^{P}} = \sqrt{1.17 \cdot 1.15} = 1.16$

(j) $W_{1,3} = Q^F_{1,3} \cdot P^F_{1,3} = 1.16 \cdot 1.13 = Q^L_{1,3} \cdot P^P_{1,3} = 1.17 \cdot 1.12 = Q^P_{1,3} \cdot P^L_{1,3} = 1.15 \cdot 1.14 = 1.31$

(k) The sales growth in the third year is 31% more than the first year.

(l) $\overline{p}_{\text{geom}} = \sqrt[n]{\prod_{i=1}^{n}(1+p_i)} - 1 = \sqrt[2]{(1+0.14)} - 1 = 0.0677 \rightarrow 6.77\%$ price rate increase.

Solution 81

Result for (a) to (d)

Example calculations:

	2019	2020	2021	2022	2023
Nominal values	\$100,000	\$102,000	\$105,060	\$110,313	\$114,726
Nominal value index [2019 = 100]	100.00	102.00	105.06	110.31	114.73
Real values	\$100,000	\$101,000	\$103,523	\$105,533	\$109,224
Real value index [2019 = 100]	100.00	101.00	103.52	105.53	109.22
Price index 1 [2018 = 100]	101.00	102.00	102.50		
Price index 2 [2021 = 100]			100.00	103.00	103.50
Price index 3 [2018 = 100]	101.00	102.00	102.50	105.58	106.09
Price index 4 [2019 = 100]	100.00	100.99	101.49	104.53	105.04

- Nominal value index [2019 = 100] for 2021: $W^{\text{nominal}}_{2019,2021} = \frac{\$105,060}{\$100,000} \cdot 100 = 105.06$
- Price index [2018 = 100] for 2022:

$$\tilde{P}_{2018,2022} = P_{2018,2021} \cdot P_{2021,2022} = 102.50 \cdot 103.00 = 105.58$$

- Shifting the base of the price index [2018 = 100] to [2019 = 100] for 2022:

$$\tilde{P}^{[2019=100]}_{2019,2022} = \frac{P^{[2018=100]}_{2018,2022}}{P^{[2018=100]}_{2018,2019}} = \frac{105.58}{101.00} \cdot 100 = 104.53$$

- Real value change for 2022: $V^{\text{real}}_{2022} = \frac{V^{\text{nominal}}_{2022}}{\tilde{P}^{[2015=100]}_{2019,2022}} = \frac{110,313}{1.0453} = \$105,533$
- Real value index [2019 = 100] for 2022: $V^{\text{nominal}}_{2019,2022} = \frac{\$105,533}{\$100,000} \cdot 100 = 105.53.$

References

Greene, W. H. (2017). *Econometric analysis* (8th ed.). Pearson Education.
Rinne, H. (2008). *Taschenbuch der Statistik* (4th ed.). Verlag Harri Deutsch.
Swoboda, H. (1971). *Exakte Geheimnisse: Knaurs Buch der modernen Statistik*. Knaur.
Wooldridge, J. M. (2019). *Introductory econometrics. A modern approach* (7th ed.). Cengage Learning.



Cluster Analysis

13

Before we turn to the subject of cluster analysis, think for a moment about the meaning of the word *cluster*. The term refers to a group of individuals or objects that converge around certain point and are thus closely related in their position. In astronomy, there are clusters of stars; in chemistry, clusters of atoms. Economic research often relies on techniques that consider groups within a total population. For instance, firms that engage in target group marketing must first divide consumers into segments, or clusters of potential customers. Indeed, in many contexts, researchers and economists need accurate methods for delineating homogenous groups within a set of observations. Groups may contain individuals (such as people or their behaviours) or objects (such as firms, products, or patents). This chapter thus takes a cue from *Goethe's Faust* (1987, Line 1943–45): "You soon will [understand]; just carry on as planned/You'll learn reductive demonstrations/ And all the proper classifications".

If we want to compare individuals or objects, we must do more than merely sample them. We must determine the dimensions of comparison, which is to say, the independent variables. Should individuals be grouped by age and height? Or by age, weight, and height?

A cluster is a group of individuals or objects with similar (i.e. homogenous) traits. The property traits of one cluster differ strongly from those of other clusters. The aim of cluster analysis is to identify homogeneous clusters within a set of heterogeneous individuals or objects.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78070-7_13.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*,
Springer Texts in Business and Economics,
https://doi.org/10.1007/978-3-031-78070-7_13

In this way, cluster analysis is an exploratory data analysis technique. "The term exploratory is important here," Everitt and Rabe-Hesketh write, "since it explains the largely absent '*p*-values', ubiquitous in many areas of statistics. […] Clustering methods are intended largely for generating rather than testing hypothesis" (2004, p. 267). This quote speaks to a frequent misunderstanding regarding cluster analysis: although it is able to group observations in a complex dataset, cluster analysis cannot determine whether the resulting groups differ significantly from each other. The mere fact that groups exist does not prove that significant differences exist between them.

Another misunderstanding about cluster analysis is the belief that there is only *one* cluster analysis technique. In reality, there are many clustering methods. Indeed, detractors claim there are as many clustering methods as users of cluster analysis. This claim has merit, as there is an incredible variety of distance measures and linkage algorithms can be used for a single clustering method (as we'll see later). Nevertheless, we can identify two general types of clustering methods:

(a) hierarchical cluster analysis
(b) k-means cluster analysis.

The following sections offer a brief introduction to both types of cluster analysis.

13.1 Hierarchical Cluster Analysis

Hierarchical clustering can be agglomerative or divisive. *Agglomerative methods* begin by treating every observation as a single cluster. For n observations, there are n clusters. Next, the distance between each cluster is determined and those closest to each other aggregated into a new cluster. The two initial clusters are never separated from each other during the next analytical steps. Now there are $n - 1$ clusters remaining. This process continues to repeat itself so that with each step the number of remaining clusters decreases and a cluster hierarchy gradually forms.[1] At the same time, however, each new step sees an increase in the difference between objects within a cluster, as the observations to be aggregated grow further apart. Researchers must decide at what point the level of heterogeneity outweighs the benefits of aggregation.

[1] By contrast, *divisive clustering methods* start by collecting all observations as one cluster. They proceed by splitting the initial cluster into two groups, and continue by splitting the subgroups, repeating this process down the line. The main disadvantage of divisive methods is their high level of computational complexity. With agglomerative methods, the most complicated set of calculations comes in the first step: for n observations, a total of $n(n - 1)/2$ distance measurements must be performed. With divisive methods containing two non-empty clusters, there are a total of $2^{(n-1)} - 1$ possible calculations. The greater time required for calculating divisive hierarchical clusters explains why this method is used infrequently by researchers and not included in standard statistics software.

	beer	cost	calories	alcohol
1	Budweiser	.43	144	4.70
2	Löwenbräu	.48	157	4.90
3	Michelob	.50	162	5.00
4	Kronenbourg	.73	170	5.20
5	Heineken	.77	152	5.00
6	Schmidt's	.30	147	4.70
7	Pabst Blue Ribbon	.38	152	4.90
8	Miller Light	.43	99	4.30
9	Bud Light	.44	113	3.70
10	Coors Light	.46	102	4.10
11	Dos Equis	.70	145	4.50
12	Beck's	.76	150	4.70
13	Rolling Rock	.36	144	4.70
14	Pabst Extra Light	.38	68	2.30
15	Tuborg	.43	155	5.00
16	Olympia Gold Light	.46	72	2.90
17	Schlitz Light	.47	97	4.20

Fig. 13.1 Beer dataset. *Source* Bühl (2019, p. 636)

Using a dataset employed by Bühl (2019, p. 636), let's take a look at the methods of hierarchical cluster analysis and the problems associated with them.[2]

Our sample dataset on 17 beers (see Fig. 13.1) contains the variables *cost per fl. oz.* and *calories per fl. oz.* Cluster analysis helps us determine how best to group the beers into clusters.

Using agglomerative clustering, we begin by seeing each beer as an independent cluster and measuring the distances between them. But what should be our reference point for measurement?

In the following section, we determine the shortest distance between the beers *Dos Equis* and Bud Light. If we take most direct route—as the crow flies—and split it into a vertical distance ($= a$) and a horizontal distance ($= b$), we get a right triangle (see Fig. 13.2). Using the Pythagorean theorem ($a^2 + b^2 = c^2$), the direct distance can be expressed as the root of the sum of the squared horizontal and vertical distances:

$$\text{Distance(Dos Equis, Bud Light)} = \sqrt{a^2 + b^2} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{13.1}$$

[2] Please refer to the dataset *Chapter13_beer.sav*, *Chapter13_beer.RData* or *Chapter13_beer.dta* (see supplementary material at https://doi.org/10.1007/978-3-031-78070-7_13).

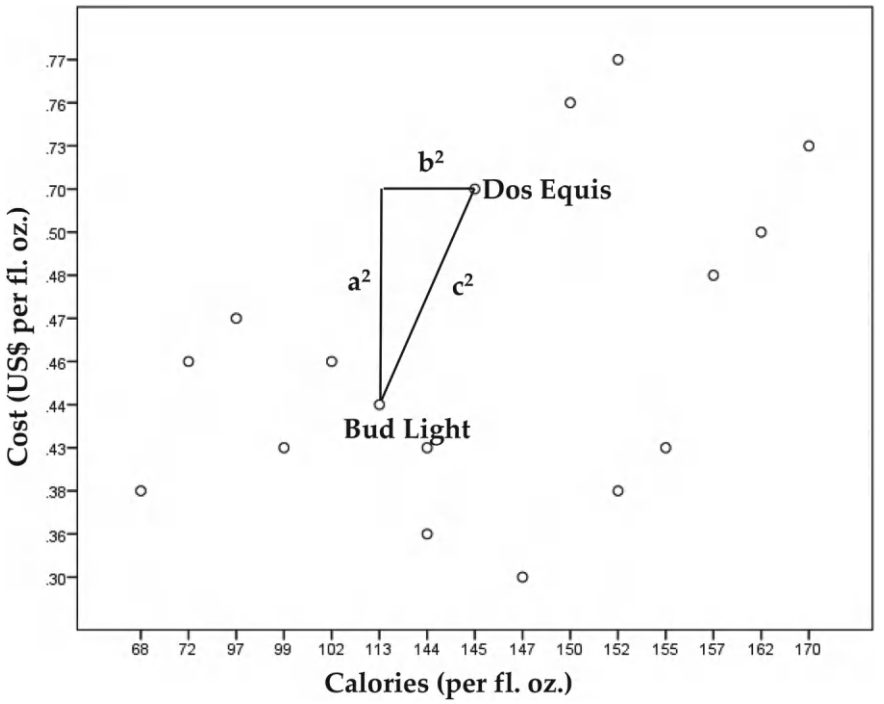


Fig. 13.2 Distance calculation 1

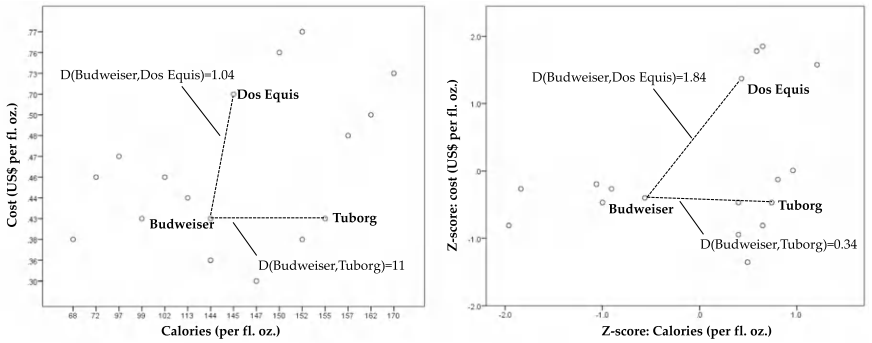
$$\text{Distance(Dos Equis, Bud Light)} = \sqrt{(0.70 - 0.44)^2 + (145 - 113)^2} = 32. \tag{13.2}$$

If more than two variables are used for comparing properties, we can no longer use the Pythagorean theorem as before. Here we need to expand the Pythagorean theorem for r-dimensional spaces by determining the *Euclidean distance* between two observations[3]:

$$\text{Distance}(A, B) = \sqrt{\left(x_1^A - x_1^B\right)^2 + \left(x_2^A - x_2^B\right)^2 + \cdots + \left(x_r^A - x_r^B\right)^2}. \tag{13.3}$$

Using this information, we can now determine the distances between, say, *Tuborg*, *Dos Equis*, and *Budweiser*. In part 1 of Fig. 13.3, the distance between *Budweiser* and *Tuborg* is eleven units, while the distance between *Budweiser* and *Dos Equis* is only 1.04 units. These results contradict the intuitive impression made by the figure. *Budweiser* and *Tuborg* seem much closer to each other than *Budweiser* and *Dos Equis.*

[3] In the case of two dimensions, the Euclidean distance and the Pythagorean theorem provide the same results.



Part 1: Non-standardized values

Part 2: Standardized values

Fig. 13.3 Distance calculation 2

In this case, our visual intuition does not deceive us. The variables *cost per fl. oz.* and *calories per fl. oz.* display two completely different units of measure. The calorie values are in the hundreds, while the costs range between 0.30 and 0.77. This means that differences in calories—e.g. the eleven units separating *Tuborg* and *Budweiser*—have a stronger impact on the distance than the differences in cost—e.g. the 0.27 units separating *Dos Equis* and *Budweiser*. And if we change the unit of measure from calories to kilocalories, the distance values shift dramatically, even as the cost difference remains the same.

This teaches us an important lesson: distance measurements in cluster analysis must rely on comparable units of measure. If the properties are in different units of measures, the variables must be made "unit free" before being measured. Usually, this is done by applying a *z-transform* to all variables in order to standardize them.[4] These functions are available in most statistics programmes. Sometimes the z-transform is overlooked, even in professional research studies. A warning sign is when only the variables with large values for a group (e.g. firm size, company expenditures) are significant. This need not indicate a lack of standardization, though researchers must be on the alert.

After the variables in our beer example are subjected to a z-transform, we arrive at the results in part 2 of Fig. 13.3. The distance between *Tuborg* and *Budweiser* is now 0.34—less than the distance between *Budweiser* and *Dos Equis* (1.84)—which agrees with the visual impression made by the figure.

The Euclidean distance is just one possible distance measure. There is a variety of other ways to measure the distance between two observations. One method is a similarity measure such as phi. The more similar the observations are to each other, the closer their distance. Every distance measure can be transformed into

[4] In standardization—sometimes also called z-transform—the mean of x is subtracted from each x variable value and the result divided by the standard deviation (S) of the x variable: $z_i = \frac{x_i - \bar{x}}{S}$.

Interval	Distance	Euclidean distance, squared Euclidean distance, Chebychev, block, Minkowski, Mahalanobis
	Similarity	Cosine, Pearson correlation
Counts	Distance	Chi-square measure
	Similarity	Phi-square measure
Binary	Distance	Euclidean distance, squared Euclidean distance, size difference, pattern difference, variance, dispersion, shape
	Similarity	Phi 4-point correlation, lambda, Anderberg's D, dice, Hamann, Jaccard, Kulczynski 1, Kulczynski 2, Lance and Williams, Ochiai, Rogers and Tanimoto, Russel and Rao, Sokal and Sneath 1, Sokal and Sneath 2, Sokal and Sneath 3, Sokal and Sneath 4, Sokal and Sneath 5, Yule's Y, and Yule's Q

Fig. 13.4 Distance and similarity measures

a similarity measure by creating an inverse value and vice versa. Distance and similarity measures are generally known as proximity measures.

Despite the analogous relationship between distance and similarity measures, distance measures are mostly used to emphasize differences between observations, while similarity measures emphasize their symmetries. Which proximity measure is appropriate depends on the scale. Figure 13.4 presents the most important distance and similarity measures grouped by scale.

It is important to note that only one of the possible proximity measures may be used in a given hierarchical cluster analysis. For instance, the chi-square may not be used for a mixed set of count and interval scaled variables. If two different variable scales exist at the same time, we must find a proximity measure permitted for both. If, say, we have binary and metric variables, we must use the squared Euclidean distance. Backhaus et al. (2021, p. 552) propose two additional strategies for dealing with the occurrence of metric and nonmetric variables. The first strategy involves calculating proximity measures for differing scales separately and then determining a weighted or unweighted arithmetic mean. In the second strategy, metric variables are transformed at a lower scale. For instance, the variable *calories per fl. oz.* can be broken down into different binary calorie variables.[5]

Let us return again to our beer example. Using the squared Euclidean distance, we obtain the *distance matrix* in Fig. 13.5.

[5] Say we wanted to dichotomize *calories per fl. oz.* using three calorie variables. Calorie variable 1 assumes the value of one when the calories in a beer lie between 60 and 99.99 cal, otherwise it is equal to zero. Calorie variable 2 assumes the value one when the calories in a beer lie between 100 and 139.99 cal, otherwise it is equal to zero. Calorie variable 3 assumes the value one when the calories in a beer lie between 140 and 200 cal, otherwise it is equal to zero.

	Budweiser	Löwenbräu	Michelob	Kronenbourg	Heineken	Schmidt's
1:Budweiser		0.279	0.541	4.832	5.430	0.793
2:Lowenbrau	0.279		0.043	3.065	3.929	1.601
3:Michelob	0.541	0.043		2.518	3.482	2.075
4:Kronenbourg	4.832	3.065	2.518		0.387	9.097
5:Heineken	5.430	3.929	3.482	0.387		10.281
6:Schmidts	0.793	1.601	2.075	9.097	10.281	
7:Pabst Blue Ribbon	0.178	0.488	0.765	6.001	7.063	0.321
8:Miller Light	1.956	3.366	4.062	9.049	8.081	3.011
9:Budweiser Light	0.933	1.945	2.487	7.044	6.526	2.027
10:Coors Light	1.746	2.941	3.552	7.852	6.877	3.145
11:Dos Equis	3.386	2.387	2.137	0.646	0.275	7.433
12:Becks	5.091	3.688	3.278	0.428	0.009	9.834
13:Rolling Rock	0.228	0.832	1.223	7.010	7.867	0.176
14:Pabst Extra Light	5.696	8.117	9.205	15.739	13.879	6.326
15:Tuborg	0.117	0.120	0.275	4.396	5.376	0.847
16:Olympia Gold Light	5.050	6.999	7.900	12.663	10.645	6.623
17:Schlitz Light	2.208	3.483	4.123	8.287	7.101	3.757

	Pabst Blue Ribbon	Miller Light	Bud Light	Coors Light	Dos Equis	Beck's
1:Budweiser	0.178	1.956	0.933	1.746	3.386	5.091
2:Lowenbrau	0.488	3.366	1.945	2.941	2.387	3.688
3:Michelob	0.765	4.062	2.487	3.552	2.137	3.278
4:Kronenbourg	6.001	9.049	7.044	7.852	0.646	0.428
5:Heineken	7.063	8.081	6.526	6.877	0.275	0.009
6:Schmidts	0.321	3.011	2.027	3.145	7.433	9.834
7:Pabst Blue Ribbon		2.830	1.637	2.712	4.802	6.709
8:Miller Light	2.830		0.194	0.050	5.429	7.569
9:Budweiser Light	1.637	0.194		0.135	4.128	6.077
10:Coors Light	2.712	0.050	0.135		4.461	6.405
11:Dos Equis	4.802	5.429	4.128	4.461		0.191
12:Becks	6.709	7.569	6.077	6.405	0.191	
13:Rolling Rock	0.080	2.184	1.226	2.169	5.369	7.464
14:Pabst Extra Light	6.817	1.044	2.123	1.414	10.483	13.201
15:Tuborg	0.125	3.030	1.709	2.756	3.482	5.081
16:Olympia Gold Light	6.480	0.746	1.643	0.869	7.823	10.057
17:Schlitz Light	3.299	0.078	0.289	0.029	4.682	6.619

	Rolling Rock	Pabst Extra Light	Tuborg	Olympia Gold Light	Schlitz Light
1:Budweiser	0.228	5.696	0.117	5.050	2.208
2:Lowenbrau	0.832	8.117	0.120	6.999	3.483
3:Michelob	1.223	9.205	0.275	7.900	4.123
4:Kronenbourg	7.010	15.739	4.396	12.663	8.287
5:Heineken	7.867	13.879	5.376	10.645	7.101
6:Schmidts	0.176	6.326	0.847	6.623	3.757
7:Pabst Blue Ribbon	0.080	6.817	0.125	6.480	3.299
8:Miller Light	2.184	1.044	3.030	0.746	0.078
9:Budweiser Light	1.226	2.123	1.709	1.643	0.289
10:Coors Light	2.169	1.414	2.756	0.869	0.029
11:Dos Equis	5.369	10.483	3.482	7.823	4.682
12:Becks	7.464	13.201	5.081	10.057	6.619
13:Rolling Rock		5.599	0.344	5.473	2.696
14:Pabst Extra Light	5.599		7.428	0.313	1.189
15:Tuborg	0.344	7.428		6.697	3.324
16:Olympia Gold Light	5.473	0.313	6.697		0.608
17:Schlitz Light	2.696	1.189	3.324	0.608	

Fig. 13.5 Distance matrix (squared Euclidean distance)

	beer
1	Budweiser
2	Löwenbräu
3	Michelob
4	Kronenbourg
5	Heineken
6	Schmidt's
7	Pabst Blue Ribbon
8	Miller Light
9	Bud Light
10	Coors Light
11	Dos Equis
12	Beck's
13	Rolling Rock
14	Pabst Extra Light
15	Tuborg
16	Olympia Gold Light
17	Schlitz Light

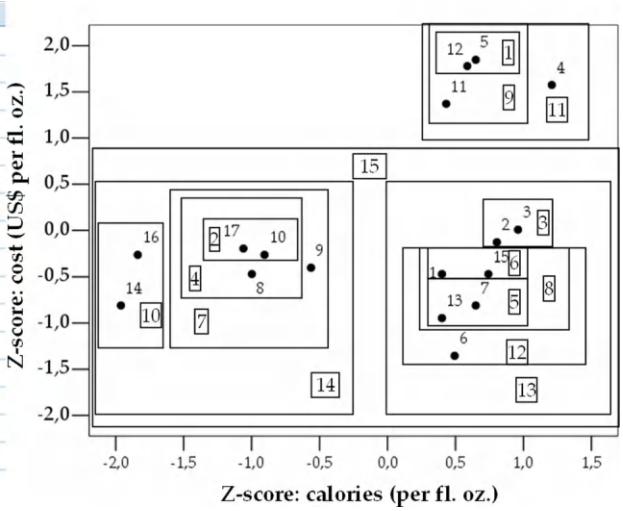


Fig. 13.6 Sequence of steps in the linkage process

After determining the distance between each observation, we aggregate the closest pair into a cluster. These are *Heineken* (#5) and *Becks* (#12), which are separated by 0.009.

The new cluster configuration consists of 15 observations and a cluster containing *Heineken* and *Becks*. Now we once again subject the (clusters of) beers to a distance measurement and link the beers closest to each other. These turn out to be *Schlitz Light* (#17) and *Coors Light* (#10). The configuration then consists of 13 different data objects and two clusters of two beers each.

We continue to repeat the distance measurement and linkage steps. We can link beers with other beers, beers with clusters, or clusters with other clusters. Figure 13.6 shows the sequence of steps in the linkage process.

With every step, the heterogeneity of linked objects tends to rise. In the first step, the distance between *Heineken* and *Becks* is only 0.009; by the 10th step, the linkage of *Pabst Extra Light* (#14) and *Olympia Gold Light* (#16) exhibits a distance of 0.313. The sequence of linkage steps and their associated distance values can be taken from the *agglomeration schedule* (see Fig. 13.7). For each step, the combined observations are given under *cluster combined* and the linkage distances under *Coefficients*. The columns *Cluster Combined* specify the (cluster of) beers which are merged during a given iteration. For example, Becks (#12) and Heineken (#5) are combined during the first iteration with a distance of 0.04. This new cluster will be part of another subsequent iteration in stage #9 (see column *Next Stage*). The column *Stage Cluster First Appears* specifies in which previous stage a (cluster of) beer already appeared. If one of the linked objects is a cluster, the number of an observation from within the cluster will be used as its stand-in.

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	5	12	.004	0	0	9
2	10	17	.019	0	0	4
3	2	3	.040	0	0	11
4	8	10	.078	0	2	7
5	7	13	.118	0	0	8
6	1	15	.177	0	0	11
7	8	9	.318	4	0	14
8	6	7	.471	0	5	13
9	5	11	.625	1	0	12
10	14	16	.781	0	0	14
11	1	2	1.045	6	3	13
12	4	5	1.370	0	9	16
13	1	6	2.470	11	8	15
14	8	14	3.907	7	10	15
15	1	8	15.168	13	14	16
16	1	4	32.000	15	12	0

Fig. 13.7 Agglomeration schedule

In stage #11, for example, those clusters of beer are merged that have already undergone an iteration in stages #6 and #3.

But we have yet to answer one question: if clusters with multiple beers arise during the cluster analysis, where should we set the points for measuring distance within a cluster? There is a wide variety of possibilities, known as *linkage methods*. There are five common linkage methods for agglomerative hierarchical clustering alone:

1. The *single linkage method* uses the closest two observations of two clusters as the basis for distance measurement. It is known as a *merge-the-closest-point strategy* (see (1) in Fig. 13.8). This technique tends to form long and snakelike chains of clusters (see (2) in Fig. 13.8).
2. The *complete linkage method* (see (3) in Fig. 13.8), by contrast, uses the furthest two observations of two clusters as the basis for distance measurement. This method generates wide yet compact cluster solutions. This technique may not be used when *elongated* cluster solutions exist in the dataset.
3. The *centroid linkage method* (see (4) in Fig. 13.8) calculates the midpoint for each cluster from its observations. This produces the centroid—the cluster's centre of gravity—which serves as the basis for distance measurement.
4. Centroid linkage should not be confused with the *average linkage method* (see (5) in Fig. 13.8), which determines the average distance between the observations of two clusters. Generally, this technique forms neither chains nor

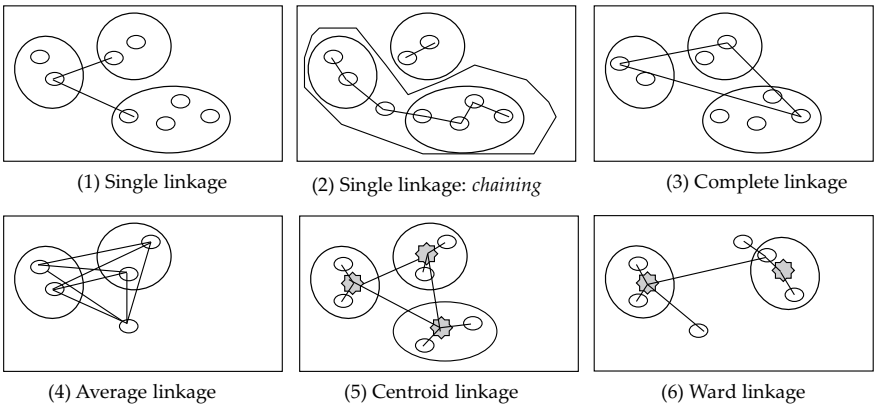


Fig. 13.8 Linkage methods

wide cluster solutions. Kaufman and Rousseeuw (1990) describe it as a robust method independent of available data.

5. *Ward's method* (proposed by Joe H. Ward (1963); see (6) in Fig. 13.8) links clusters that optimize a specific criterion: the error sum of squares. This criterion minimizes the total within-cluster variance. As with other hierarchical methods, it begins by seeing every observation as its own cluster. In this case, the error sum of squares assumes the value of zero, as every observation equals the cluster mean. Let's illustrate the next linkage step with an example. Assume the initial observation values are 2, 4, and 5. We begin by identifying the error sum of squares: $\text{QS} = (2-2)^2 + (4-4)^2 + (5-5)^2 = 0$. Next we calculate the error sum of squares for all possible combinations of next linkages. From this, we choose clusters that lead to the fewest increases in the error sum of squares. For our example, the following clusters are possible:
   (a) the observation values two and four with a mean of three,
   (b) the observation values two and five with a mean of 3.5, or
   (c) the observation values four and five with a mean of 4.5.

These clusters yield the following error sums of squares:

(1) $\text{QS} = [(2-3)^2 + (4-3)^2] + (5-5)^2 = 2$
(2) $\text{QS} = [(2-3.5)^2 + (5-3.5)^2] + (4-4)^2 = 4.5$
(3) $\text{QS} = (2-2)^2 + [(4-4.5)^2 + (5-5.5)^2] = 0.5.$

The value for the third linkage is the lowest. Its aggregated cluster raises the error sum of squares for all clusters by 0.5, the least of them all.

When several variables are used for clustering, the sum of squares is determined not by the cluster mean but by the cluster centroid. Figure 13.8 presents the basic idea behind each linkage method.

Though each method follows a logical rationale, they rarely lead to the same cluster solution. Dilation techniques like the complete linkage method tend to produce equal sized groups; contraction techniques like the single linkage method tend to build long, thin chains. "We can make use of the chaining effect to detect [and remove] outliers, as these will be merged with the remaining objects—usually at very large distances—in the last step of the analysis" (Sarstedt & Mooi, 2019, p. 311). Ward's method, centroid linkage, and average linkage exhibit no dilating or contracting qualities, hence their status as "conservative" methods. In scientific practice, the usual recommendation is to use single linkage first. After excluding possible outliers, we can move on to Ward's method. Ward's method has established itself as the preferred technique for metric variables, and multiple studies have confirmed the quality of the cluster solutions generated by this technique (Berg, 1981, p. 96).

As the heterogeneity of linked observations increases with each step, we must keep in mind that at a certain number of iterations the differences outweigh the utility of linkage. Recall again the definition of a cluster: a group of individuals or objects with similar (homogenous) traits. What are some criteria for when to stop the linkage process?

Though researchers ultimately have to make this decision themselves, there are three criteria to help ensure the objectivity of their results.

1. It is better to end with a cluster number for which heterogeneity increases in jumps. The agglomeration schedule can provide some indications of when such jumps occur (see the column *coefficients* in Fig. 13.7; here the jump occurs between the coefficients 3.907 and 15.168, which suggests a three-cluster solution). Dendrograms and scree plots are two other forms of visual identification.

The term *dendrogram* comes from the Greek word for tree. It's called this because it presents cluster solutions in branch-like form (see Fig. 13.9). The length of each branch equals the heterogeneity level of the cluster, with values normalized to a scale between zero and 25. Reading the dendrogram of the beer example in Fig. 13.9 from left to right, we see that the beers 4, 5, 11, and 12 are clustered with short branches, or a low heterogeneity level. The same is true of the beer cluster 1, 2, 3, and 15. When the latter cluster is linked with the beer cluster 6, 7, and 13, heterogeneity increases somewhat. The linkage of light beers (8, 9, 10, 14, 16, 17) with affordable regular beers (1, 2, 3, 6, 7, 13, 15) implies a comparatively high level of heterogeneity (long branches).

When determining the optimal number of clusters, we proceed as a gardener who begins pruning his tree at the first big branch on the left. This "pruning" is indicated by the dotted line in Fig. 13.9. The number of branches to be pruned corresponds to the number of clusters—in this case, three.

In a *scree plot,* the number of clusters is plotted from lowest to highest on the x-axis and their respective heterogeneity jumps are plotted on the y-axis. A homogenous cluster solution usually occurs when the number of clusters produces

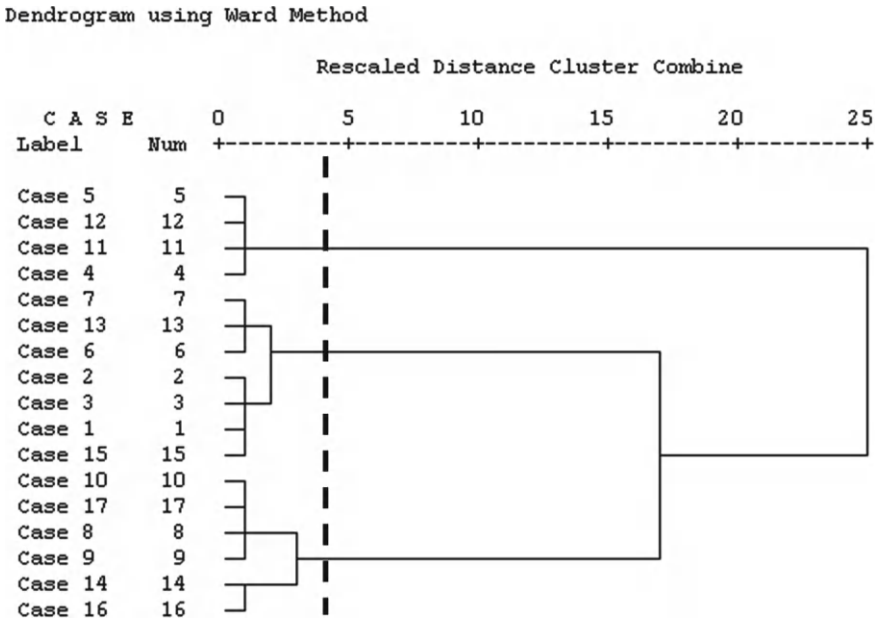


Fig. 13.9 Dendrogram

a line that converges asymptotically on the abscissa. Our beer example yields the scree plot in Fig. 13.10 (confirmed by a three-cluster solution):

Though scree plots and dendrograms are frequently used in empirical research, they do not always yield unambiguous results.

2. The second criterion is obtained by calculating the quotient of the sum of all variances within all clusters and the variance of the total sample. This is called the *F*-value. If the quotient for all clusters and variables is less than one, the dispersion of group properties is low compared with the total number of observations. Cluster solutions with *F*-values less than one produce large intragroup homogeneity and small intergroup homogeneity. Cluster solutions with *F*-values over one have undesirably high heterogeneity levels.

Some statistics programmes do not calculate *F*-values automatically during cluster analysis. In such cases, *F*-values must be determined by calculating variances individually. Figure 13.11 provides the corresponding *F*-values for our example. In the two-cluster solution, the *F*-value for the variable *calories* in cluster one is noticeably greater than one:

$$F = \frac{1117.167}{1035.110} = 1.079. \tag{13.4}$$

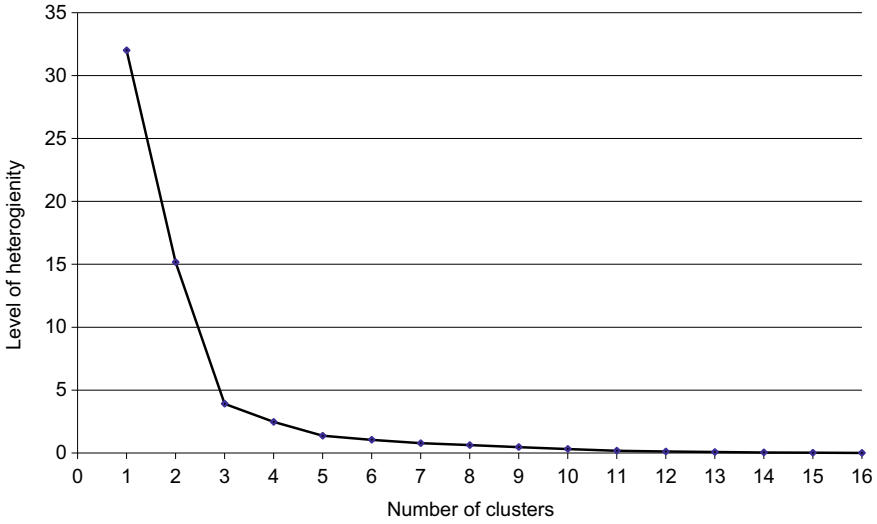


Fig. 13.10 Scree plot identifying heterogeneity jumps

	Variance		F-value	
	calories	cost	calories	cost
1 Cluster	1035.110	0.022	1.000	1.000
2 Cluster	1117.167	0.003	1.079	0.136
3 Cluster	47.620	0.005	0.046	0.227
4 Cluster	47.619	0.005	0.046	0.227
5 Cluster	57.667	0.001	0.056	0.045

Fig. 13.11 *F*-value assessments for cluster solutions 2 to 5

Only with the three-cluster solution are all *F*-values smaller than one, which is to say, only the three-cluster solution produces homogenous clusters.

3. The last procedure for checking individual cluster solutions is called *discriminant analysis*. Since I do not treat this method explicitly in this book, I will sketch its relationship to cluster quality only briefly. Discriminant analysis uses mathematical functions (known as discriminant functions) to present the information of independent variables in compressed form. The comparison of the given cluster classification with the classification predicted by the discriminant function provides the number of incorrectly classified observations. In my experience, an error rate over 10% produces qualitatively unusable results. In our beer example, all cluster solutions between two and five clusters can be correctly classified using discriminant analysis. A sample result for the three-cluster solution is provided in Fig. 13.12.

Classification Results[a,c]

		Ward Method	Predicted Group Membership 1	2	3	Total
Original	Count	1	7	0	0	7
		2	0	4	0	4
		3	0	0	6	6
	%	1	100.0	.0	.0	100.0
		2	.0	100.0	.0	100.0
		3	.0	.0	100.0	100.0
Cross-validated[b]	Count	1	7	0	0	7
		2	0	4	0	4
		3	0	0	6	6
	%	1	100.0	.0	.0	100.0
		2	.0	100.0	.0	100.0
		3	.0	.0	100.0	100.0

a. 100.0% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

c. 100.0% of cross-validated grouped cases correctly classified.

Fig. 13.12 Cluster solution and discriminant analysis

Discriminant analysis delivers some clues for how to interpret cluster solutions. Variance analysis, too, can help generate different *cluster profiles*. Let us consider the three-cluster solution *graphically* in Fig. 13.13.

Cluster three contains all light beers with a lower-than-average calorie count and a lower-than-average cost. Cluster one contains all low-cost regular beers with a higher-than-average calorie count. The premium beers in cluster two exhibit both higher-than-average costs and higher-than-average calorie counts. Based on this chart, the three-cluster solution appears to offer logical conclusions. But can we assume that that the groups are significantly different from one another statistically? In Sect. 9.5.1 we learned about how analysis of variance (ANOVA) can be used to check group differences for significance.

When we apply this technique using the results of our cluster solution—with cluster membership as the independent variable and *costs* and *calories* as the dependent variables in two different ANOVAs—we ascertain significant differences among the three groups (see first table in Fig. 13.14). According to the post-hoc method, premium beers are significantly more expensive than other beers (see third table in Fig. 13.14) and light beers have significantly fewer calories than other beers (see second table in Fig. 13.14).

A test of these methods should make the advantages and disadvantages of cluster analysis readily apparent. On the one hand, cluster analysis is not an inference technique and, as such, has no prerequirements (e.g. the existence of a normal

	beer
1	Budweiser
2	Löwenbräu
3	Michelob
4	Kronenbourg
5	Heineken
6	Schmidt's
7	Pabst Blue Ribbon
8	Miller Light
9	Bud Light
10	Coors Light
11	Dos Equis
12	Beck's
13	Rolling Rock
14	Pabst Extra Light
15	Tuborg
16	Olympia Gold Light
17	Schlitz Light

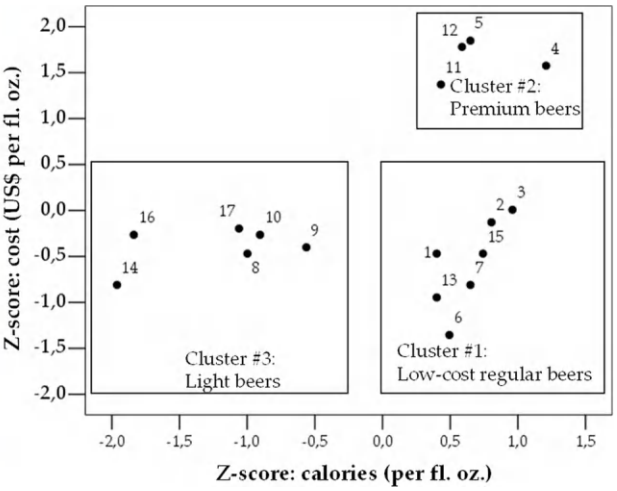


Fig. 13.13 Cluster interpretations

distribution). On the other hand, it is unable to verify the statistical significance of the results.

The absence of typical usage requirements (e.g. a normal distribution of variables) does not mean we can use cluster analysis arbitrarily. A few requirements still apply:

- The sample must be representative.
- Multicollinearity problems must be avoided. We discussed this problem in the chapter on regression analysis. Because each variable possesses the same weight in cluster analysis, the existence of two or more multicollinear variables leads to a high likelihood that this dimension is represented twice or more in the model. Observations that exhibit similarity for this dimension have a higher chance of ending up in a common cluster.
- Agglomerative methods with large datasets cannot be calculated with traditional desktop software. In these instances, a k-means cluster analysis should be used instead.

13.2 K-Means Cluster Analysis

K-means clustering is another method of analysis that groups observations into clusters. The main difference between k-means clustering and hierarchical clustering is that users of k-means clustering decide on the number of clusters at the beginning. In the first step, we determine an initial partition by assigning observations to clusters. We need not worry about whether our assignments are arbitrary

Tests of Between-Subjects Effects

Dependent Variable: Zscore: Cost (US$ per fl. oz.)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	14.250[a]	2	7.125	57.004	.000
Intercept	.748	1	.748	5.987	.028
CLU3_1	14.250	2	7.125	57.004	.000
Error	1.750	14	.125		
Total	16.000	17			
Corrected Total	16.000	16			

a. R Squared = .891 (Adjusted R Squared = .875)

Multiple Comparisons

Dependent Variable: Zscore: Calories (per fl. oz.)

	(I) Ward Method	(J) Ward Method	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Scheffe	Low Cost Beers	Premium Beers	-.0832549	.24605562	.945	-.7561072	.5895974
		Light Beers	1.8567698*	.21840521	.000	1.2595290	2.4540105
	Premium Beers	Low Cost Beers	.0832549	.24605562	.945	-.5895974	.7561072
		Light Beers	1.9400247*	.25340218	.000	1.2470828	2.6329666
	Light Beers	Low Cost Beers	-1.8567698*	.21840521	.000	-2.4540105	-1.2595290
		Premium Beers	-1.9400247*	.25340218	.000	-2.6329666	-1.2470828
Tamhane	Low Cost Beers	Premium Beers	-.0832549	.18786364	.967	-.7846054	.6180956
		Light Beers	1.8567698*	.24041856	.001	1.0818537	2.6316858
	Premium Beers	Low Cost Beers	.0832549	.18786364	.967	-.6180956	.7846054
		Light Beers	1.9400247*	.28275410	.000	1.0902359	2.7898134
	Light Beers	Low Cost Beers	-1.8567698*	.24041856	.001	-2.6316858	-1.0818537
		Premium Beers	-1.9400247*	.28275410	.000	-2.7898134	-1.0902359

Based on observed means.
The error term is Mean Square(Error) = .154.

*. The mean difference is significant at the 0.05 level.

Multiple Comparisons

Dependent Variable: Zscore: Cost (US$ per fl. oz.)

	(I) Ward Method	(J) Ward Method	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Scheffe	Low Cost Beers	Premium Beers	-2.2389633*	.22159444	.000	-2.8449252	-1.6330014
		Light Beers	-.1946925	.19669285	.623	-.7325596	.3431747
	Premium Beers	Low Cost Beers	2.2389633*	.22159444	.000	1.6330014	2.8449252
		Light Beers	2.0442709*	.22821066	.000	1.4202166	2.6683251
	Light Beers	Low Cost Beers	.1946925	.19669285	.623	-.3431747	.7325596
		Premium Beers	-2.0442709*	.22821066	.000	-2.6683251	-1.4202166
Tamhane	Low Cost Beers	Premium Beers	-2.2389633*	.20972786	.000	-2.8548881	-1.6230385
		Light Beers	-.1946925	.20183014	.738	-.7876445	.3982596
	Premium Beers	Low Cost Beers	2.2389633*	.20972786	.000	1.6230385	2.8548881
		Light Beers	2.0442709*	.14130301	.000	1.5996270	2.4889147
	Light Beers	Low Cost Beers	.1946925	.20183014	.738	-.3982596	.7876445
		Premium Beers	-2.0442709*	.14130301	.000	-2.4889147	-1.5996270

Based on observed means.
The error term is Mean Square(Error) = .125.

*. The mean difference is significant at the 0.05 level.

Fig. 13.14 Test of the three-cluster solution with two ANOVAs

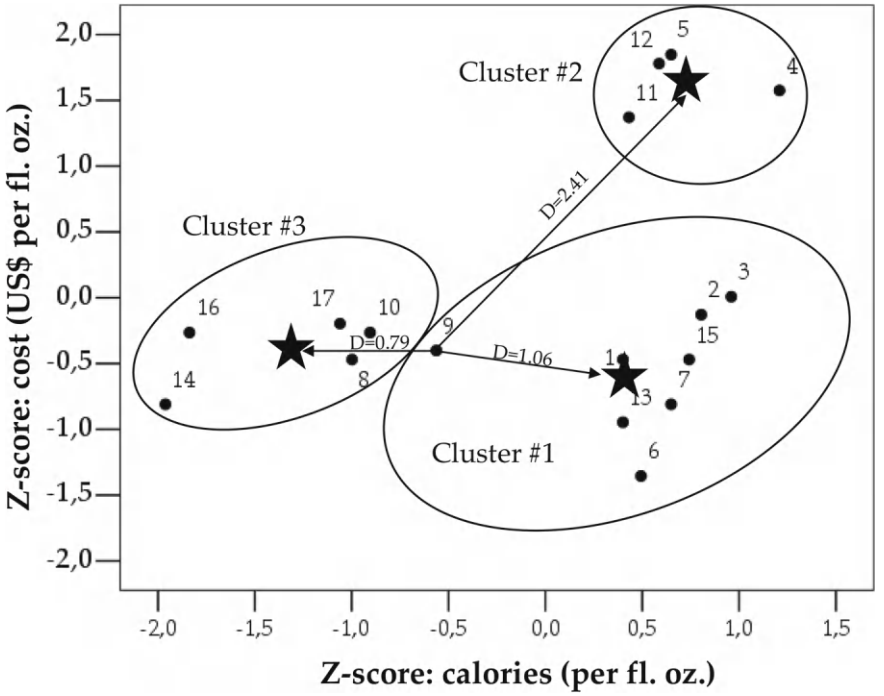


Fig. 13.15 Initial partition for k-means clustering

or logical. The only problem with poor assignments is that they increase calculating time. The better the clustering is in the initial partition, the faster we obtain the final result.

After we set the initial partition, we can then start thinking about the quality of the clustering. Let us turn again to our beer example. Assume we have set the three clusters shown in Fig. 13.15. This clustering is different from that provided by hierarchical analysis. Here *Bud Light* (#9) is grouped with low-cost beers, not with light beers.

We begin by calculating the centroid for each of the clusters.[6] Every observation should be close to the centroid of its own cluster—once again, a cluster is by definition a group of objects with similar traits—and at the very least should be closer to the centroid of its own cluster than to the centroid of a neighbouring cluster. When we reach observation #9 (see Fig. 13.15), we notice that *Bud Light* has an Euclidean Distance of 1.06 from its own centroid[7] (cluster #1), an Euclidean

[6] The centroid is determined by calculating the mean for every variable for all observations of each cluster separately.

[7] Euclidean Distance of #9 to Centroid CLU#1: $\sqrt{(-0.571-(-0.401))^2+(0.486-(-0.563))^2}=1.06$.

Distance of 2.41 from the centroid of cluster #2,[8] and an Euclidean Distance of 0.79 from the centroid of cluster #3.[9] We thus assign *Bud Light* to the light beers in cluster #3. This changes the location of the centroids of both clusters, so we must again verify that all observations lie closer to their own centroid than to the centroid of the neighbouring cluster. If they do, then we have arrived at the optimal clustering. If not, the observations must be reassigned and the centroids calculated anew.

There are a variety of other strategies for improving the quality of k-means clustering. The author prefers variance to centroid distance for assessing assignments. When using this technique, we first determine the sum of squared errors for the initial partition. Then we check which change in assignment reduces the sum of squared errors the most. We repeat this process until the total error variance can no longer be minimized.

K-means clustering has the following requirements:

- Knowledge about the best number of clusters. Different cluster solutions can be tested and their quality compared using a suitable method, such as hierarchical cluster analysis, discriminant analysis, or variance analysis.
- Metric variables must be z-transformed and checked for multicollinearities before clustering.

Due to the high computational complexity of hierarchical agglomerative methods, researchers with large datasets often must turn to k-means clustering. Researchers using hierarchical clustering can also use k-means clustering to test the quality of a given cluster solution.

13.3 Cluster Analysis Using SPSS, Stata, and R

13.3.1 Cluster Analysis with SPSS

This section uses the SPSS dataset *Chapter13_beer.sav* (see supplementary material at https://doi.org/10.1007/978-3-031-78070-7_13). Follow the steps outlined in Figs. 13.16 and 13.17.

[8] Euclidean Distance of #9 to Centroid CLU#2: $\sqrt{(1.643-(-0.401))^2+(0.719-(-0.563))^2}=2.41$.

[9] Euclidean Distance of #9 to Centroid CLU#3: $\sqrt{(-0.401-(-0.401))^2+(-1.353-(-0.563))^2}=0.79$.

➔ Select *Analyze* → *Classify* → *Hierarchical Cluster…* to open the dialogue box

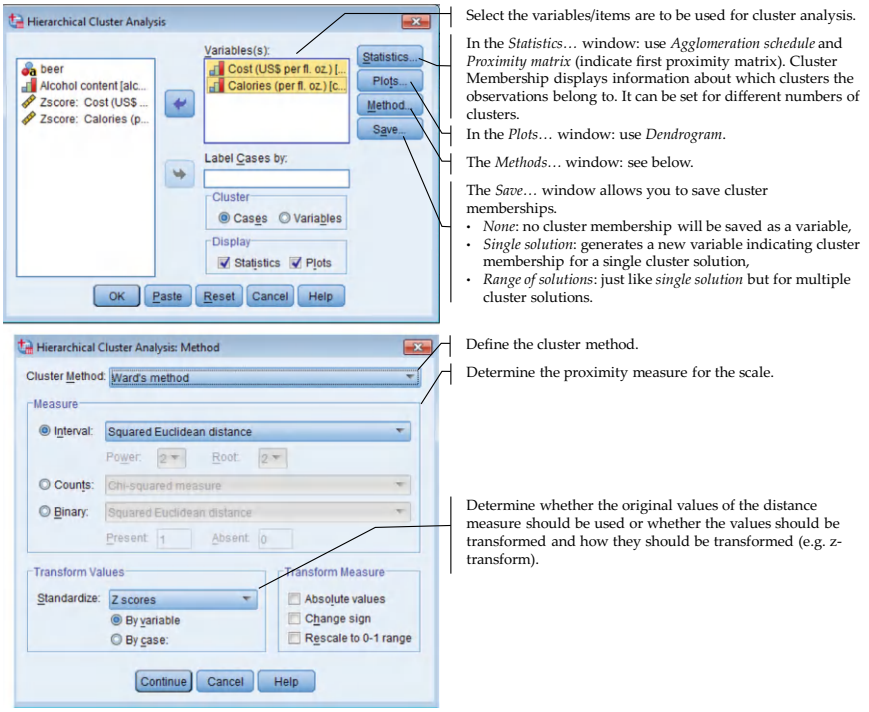


Fig. 13.16 Hierarchical cluster analysis with SPSS

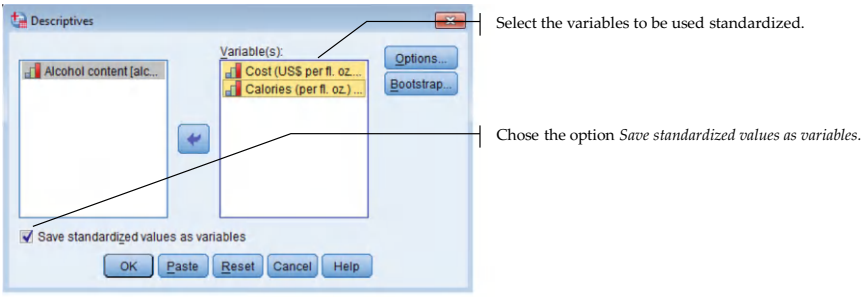
13.3.2 Cluster Analysis with Stata

This section uses the Stata dataset *Chapter13_beer.dta* (see supplementary material at https://doi.org/10.1007/978-3-031-78070-7_13). Follow the steps outlined in Fig. 13.18.

13.3.3 Cluster Analysis with R Commander

This section uses the R dataset *Chapter13_beer.RData* (see supplementary material at https://doi.org/10.1007/978-3-031-78070-7_13). Follow the steps outlined in Fig. 13.19, for hierarchical cluster analysis, for k-means cluster analysis follow the steps in Fig. 13.20.

➔ Select *Analyze* → *Descriptive statistics* → *Descriptives…* to standardize variables



➔ Then select *Analyze* → *Classify* → *K-Means Cluster…* to obtain the K-means cluster analysis

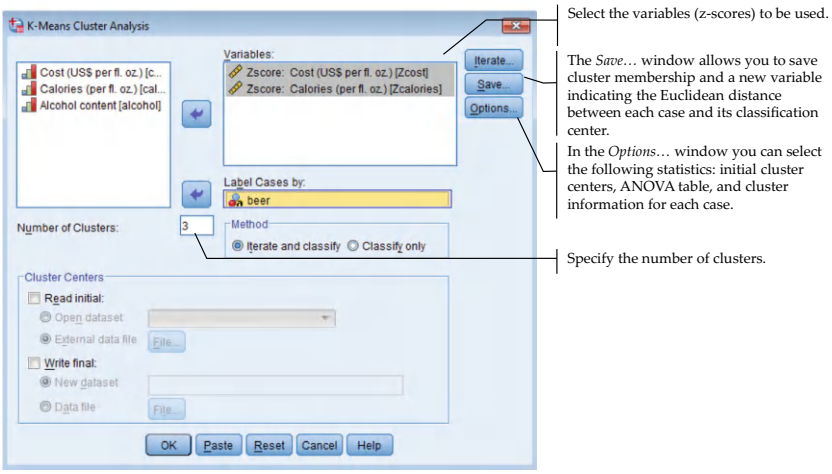


Fig. 13.17 K-means cluster analysis with SPSS

13.4 Chapter Exercises

Exercise 82

The share of the population living in urban areas and infant deaths per 1000 births were collected for 28 European countries. Afterwards the data underwent a hierarchical cluster analysis (see dataset *Chapter13_Child mortality and urban population.sav*, *Chapter13_Child mortality and urban population.RData* or *Chapter13_Child mortality and urban population.dta*; supplementary material at https://doi.org/10.1007/978-3-031-78070-7_13). The results are presented in Fig. 13.21.

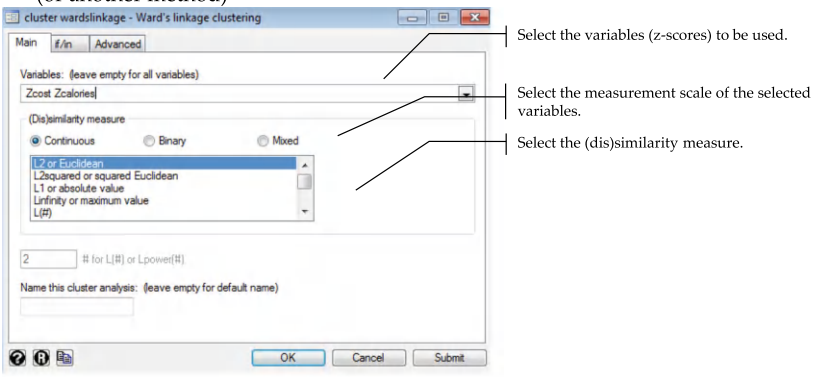
(a) Sketch the basic steps of the cluster analysis method from the agglomeration table (see Fig. 13.21).

Z-transform commands:

➔ Independent variables may have to be standardized. Use the command *"egen float z_variable1 = std(variable_1), mean(0) std(1)."*

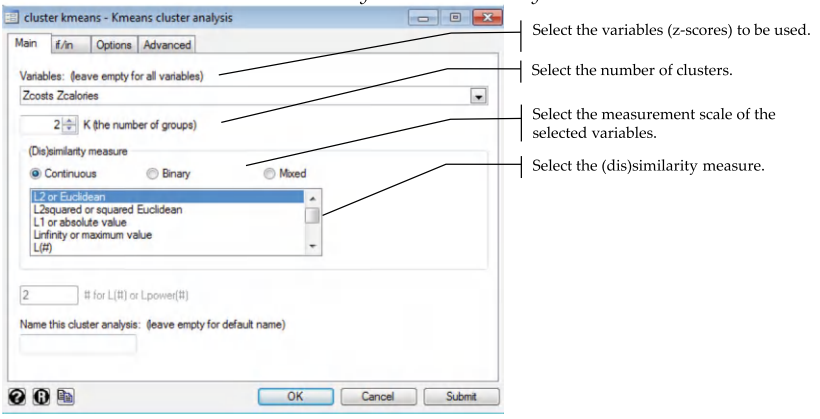
Hierarchical cluster analysis commands:

➔ Select *Statistics* → *Multivariate analysis* → *Cluster analysis* → *Cluster data* → *Ward´s linkage* (or another method)



K-means cluster analysis commands:

➔ Select *Statistics* → *Multivariate analysis* → *Cluster analysis* → *Cluster data* → *Kmeans*



Postclustering commands:

➔ Select *Statistics* → *Multivariate analysis* → *Cluster analysis* → *Postclustering*

- *Dendrogram*: displays dendrogram.
- *Summary variables from cluster analysis*: save a specific cluster solution.

Important syntax commands for variance analysis:

cluster averagelinkage; cluster centroidlinkage; cluster completelinkage; cluster wardslinkage; cluster singlelinkage; cluster medianlinkage; cluster waveragelinkage; cluster dendrogramm; cluster generate; cluster kmeans

Fig. 13.18 Cluster analysis with Stata

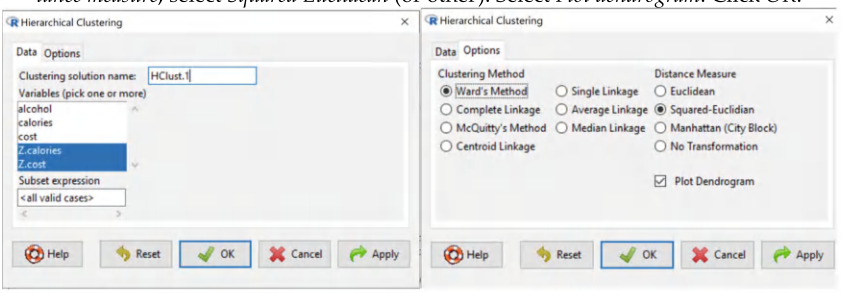
Z-transform commands:

➔ In R Commander, click on *Data* → *Manage variables in active dataset* → *Standardize variables…*. In the *Variables* box, select the variables to be standardized (*cost* and *calories* in our case). This command generates new z-transformed variables (*Z.calories* and *Z.cost* in our case).

Hierarchical cluster analysis commands:

➔ Select *Statistics* → *Dimensional analysis* → *Cluster analysis* → *Hierarchical cluster analysis…*

➔ Select the variables to be included in the analysis (*Z.calories* and *Z.cost* in our case) and click on *Options*: Under *Clustering method*, select *Ward's method* (or other) and under *Distance measure*, select *Squared Euclidean* (or other). Select *Plot dendrogram*. Click *OK*.



➔ The dendrogram will appear in the window of R.

➔ Then click *Statistics* → *Dimensional analysis* → *Cluster analysis* → *Summarize hierarchical clustering…*

➔ Select one of the cluster solutions on the left and specify the number of clusters on the right. Click on *OK* (see below). A bi-plot of the cluster solution is provided in the window of R.

➔ Then click *Statistics* → *Dimensional analysis* → *Cluster analysis* → *Add hierarchical clustering to data set…*This command generates a new variable indicating cluster membership for a specified number of clusters (see below).

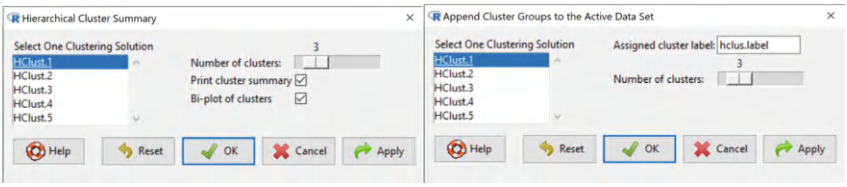


Fig. 13.19 Hierarchical cluster analysis with R Commander

(b) How many clusters make sense from a methodological standpoint? Explain your answer.

Exercise 83

A market research institute studied the relationship between income and personal satisfaction (see dataset *Chapter13_Income_satifaction.sav*, *Chapter13_Income_satifaction.RData* or *Chapter13_Income_satifaction.dta*; supplementary material at

Z-transform commands:

➔ In R Commander, click on *Data* → *Manage variables in active dataset* → *Standardize variables*…. In the *Variables* box, select the variables to be standardized (*cost* and *calories* in our case). This command generates new z-transformed variables (*Z.calories* and *Z.cost* in our case).

K-means cluster analysis commands:

➔ Select *Statistics* → *Dimensional analysis* → *Cluster analysis* → *k-means cluster analysis*…

➔ Select the variables to be included in the analysis (*Z.calories* and *Z.cost* in our case) and click on *Options*: Specify the *Number of clusters* (three in our case) and the maximum of iterations (20 in our case). Select *Print cluster summary* (to get results for the k-means clustering) and *Bi-plot of clusters*. *Assign clusters to the data set* generates a new variable indicating cluster membership for a specified number of clusters (see below). Click *OK*.

➔ The bi-plot of the cluster solution is provided in the window of R.

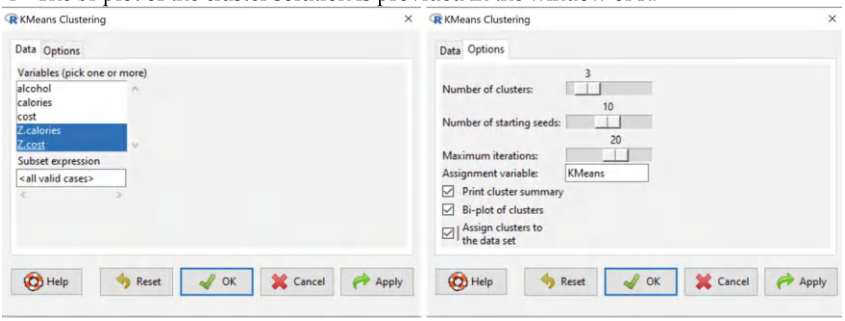


Fig. 13.20 K-means cluster analysis with R Commander

https://doi.org/10.1007/978-3-031-78070-7_13). They used hierarchical clustering to analyse their data (see Fig. 13.22).

(a) Assume you decide for a four-cluster solution. Circle the four clusters in the Fig. 13.23.
(b) Characterize the contents of the four clusters.
(c) Assume you decide for a five-cluster solution. Circle the five clusters into Fig. 13.23.
(d) Which approach makes the most sense from a methodological standpoint?
(e) Now, the market research institute used K-means clustering to analyse their data. Please interpret the table *Final Cluster Centres* in Fig. 13.24. What is the difference between the three-cluster-solution of the hierarchical and the K-means approach?

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	7	23	.008	0	0	21
2	4	16	.016	0	0	12
3	9	24	.031	0	0	20
4	13	26	.048	0	0	13
5	11	18	.083	0	0	18
6	12	22	.120	0	0	17
7	5	15	.162	0	0	14
8	14	20	.204	0	0	22
9	3	19	.255	0	0	13
10	6	25	.307	0	0	14
11	27	28	.407	0	0	16
12	4	10	.560	2	0	15
13	3	13	.742	9	4	20
14	5	6	.927	7	10	23
15	4	8	1.138	12	0	19
16	21	27	1.379	0	11	22
17	2	12	1.692	0	6	19
18	11	17	2.008	5	0	21
19	2	4	2.531	17	15	26
20	3	9	3.095	13	3	25
21	7	11	3.695	1	18	23
22	14	21	5.270	8	16	24
23	5	7	7.057	14	21	25
24	1	14	9.591	0	22	27
25	3	5	13.865	20	23	26
26	2	3	25.311	19	25	27
27	1	2	54.000	24	26	0

Fig. 13.21 Hierarchical cluster analysis. *Source* Bühl (2019, p. 636)

13.5 Exercise Solutions

Solution 82

- First the variables must be z-transformed and then the distance or similarity measures determined. Next, the distance between the remaining objectives must be measured and linked with its nearest objects. This step is repeated until the heterogeneity exceeds an acceptable level.
- A four-cluster-solution makes sense, since further linkage raises heterogeneity levels excessively. The last heterogeneity jump rises from 9.591 to 13.865.

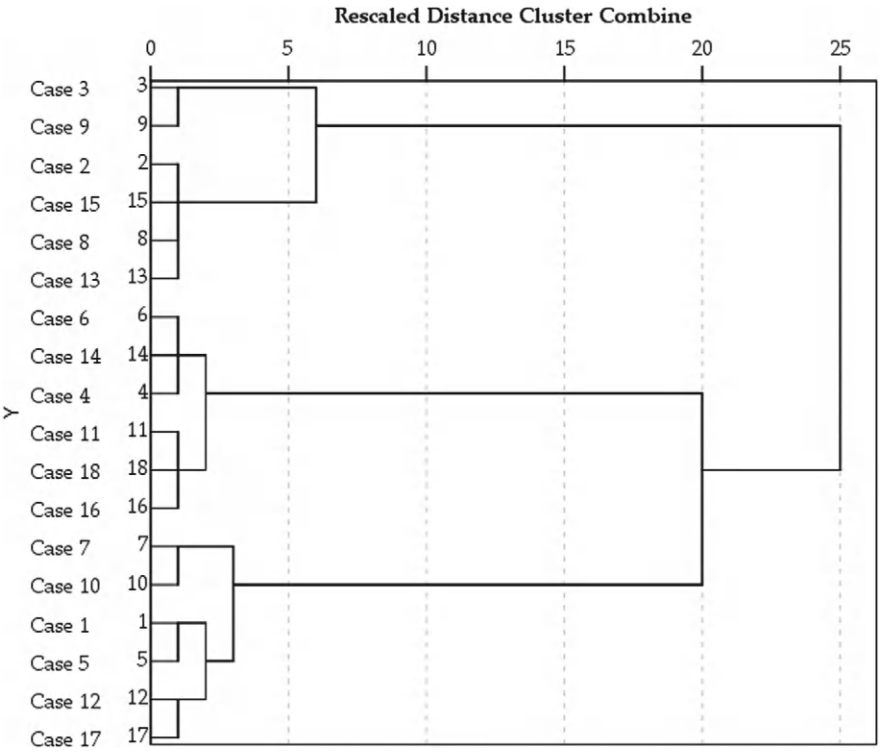


Fig. 13.22 Dendrogram

Solution 83

(a) See Fig. 13.25.
(b) Cluster #1: more dissatisfied customers with high income; Cluster #2: dissatisfied customers with middle income; Cluster #3: dissatisfied customers with low income; Cluster #4: satisfied customers with medium to high income.
(c) Cluster #1 of solution (a) is divided into two clusters (see dotted circles in Cluster #1).
(d) Four clusters, since heterogeneity barely increases between four and five clusters.
(e) See Fig. 13.26.

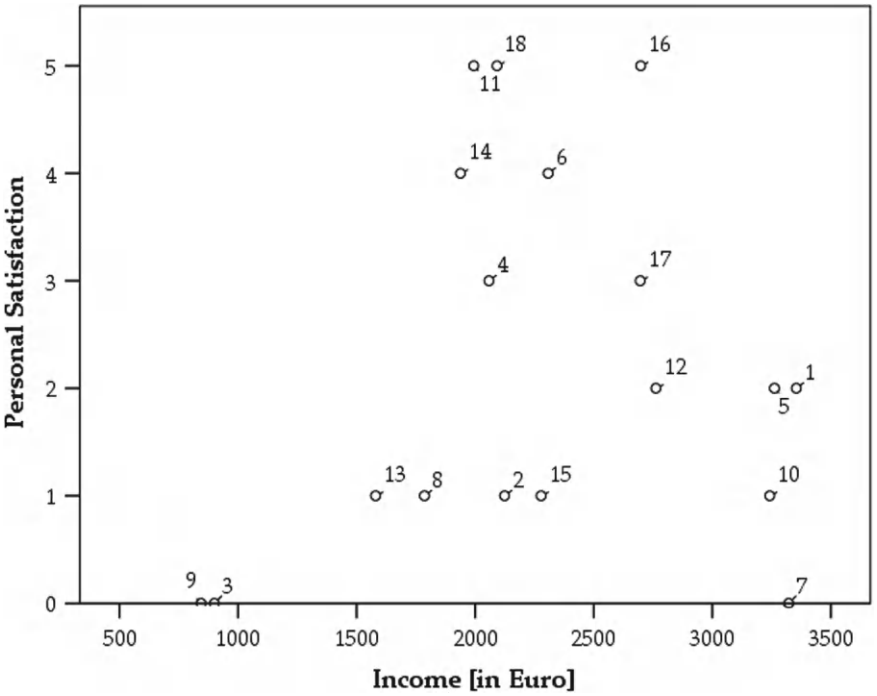


Fig. 13.23 Cluster memberships

Final Cluster Centers

	Cluster		
	1	2	3
Zscore: Income [in euros]	.81388	-.04781	-1.34062
Zscore: Personal satisfaction	-.52984	1.08662	-.97436

Fig. 13.24 Final cluster centres and cluster memberships

Cluster Membership

Case number	Cluster	Distance	Case number	Cluster	Distance
1	1	.717	10	1	.473
2	1	1.047	11	2	.595
3	3	.574	12	1	.447
4	2	.697	13	3	.490
5	1	.620	14	2	.427
6	2	.107	15	1	.847
7	1	.912	16	2	.761
8	3	.730	17	2	.871
9	3	.639	18	2	.531

Fig. 13.24 (continued)

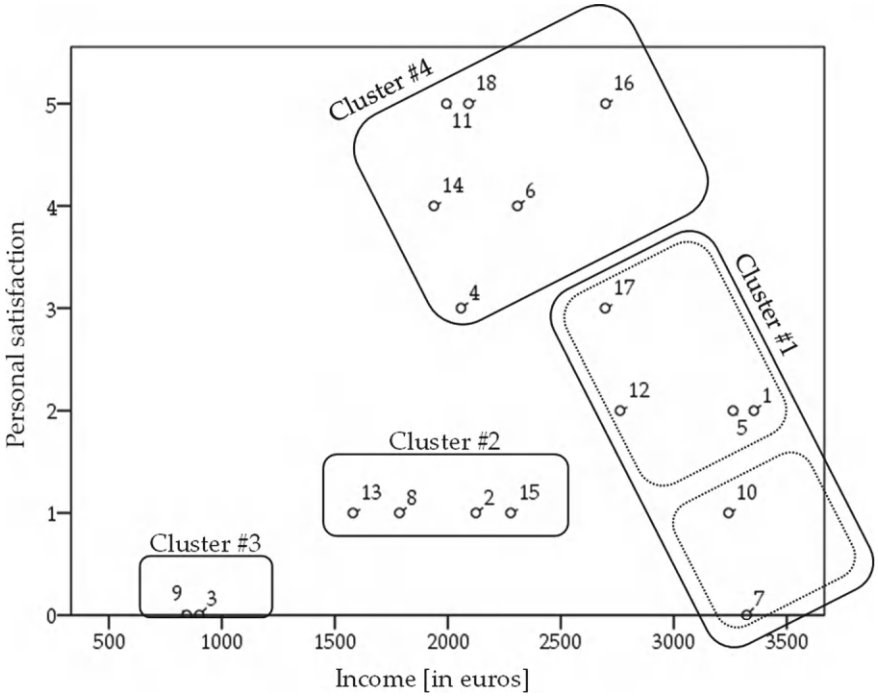


Fig. 13.25 Cluster analysis (1)

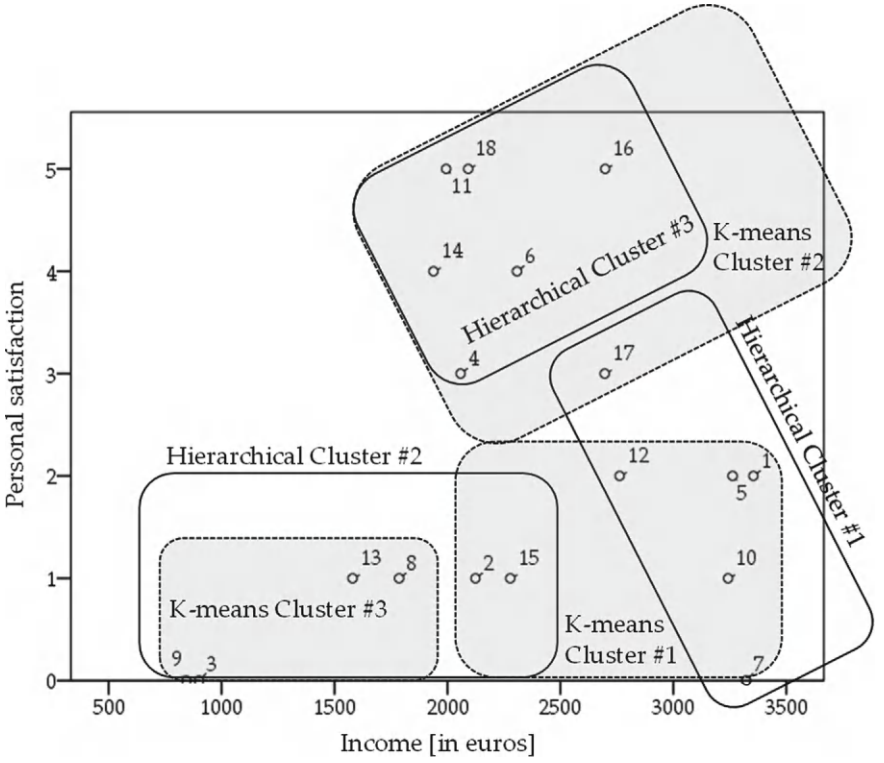


Fig. 13.26 Cluster analysis (2)

References

Backhaus, K., Erichson, B., Gensler, S., Weiber, R., & Weiber, T. (2021). *Multivariate Analysemethoden. Eine Anwendungsorientierte Einführung* (16th ed.). Springer.

Berg, S. (1981). *Optimalität bei Cluster-Analysen.* Münster: Dissertation, Fachbereich Wirtschafts- und Sozialwissenschaften, Westfälische Wilhelms-Universität Münster.

Bühl, A. (2019). *SPSS: Einführung in die moderne Datenanalyse ab SPSS 25* (16th ed.). Pearson Studium.

Everitt, B. S., & Rabe-Hesketh, S. (2004). *A handbook of statistical analyses using Stata* (3rd ed.). Chapman & Hall.

Goethe, J. W. (1987). *Faust part one.* Translated with an Introduction and Notes by David Luke. Oxford University Press.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data.* Wiley.
Mooi, E., & Sarstedt, M. (2019). *A concise guide to market research. The process, data, and methods using IBM SPSS statistics* (3rd ed.). Springer.
Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*, 236–244.



Factor Analysis

14

14.1 Factor Analysis: Foundations, Methods, Interpretations

Frequently, empirical studies rely on a wide variety of variables—so-called item batteries—to describe a certain state of affairs. An example for such a collection of variables is the study of preferred toothpaste attributes by Malhotra et al. (2017, p. 712). Thirty people were asked the questions in Fig. 14.1.

Assuming these statements are accurate descriptions of the original object—preferred toothpaste attributes—we can decrease their complexity by reducing them to some underlying dimensions or factors. Empirical researchers use two basic approaches for doing so:

1. The first method adds the individual item values to produce a total index for each person. The statement scores—which in our example range from one to seven—are simply added together for each person. One problem with this method occurs when questions are formulated negatively, as with question five in our example. Another problem with this method is that it assumes the one dimensionality of the object being investigated or the item battery being applied. In practice, this is almost never the case. In our example, the first, third, and fifth statements describe health benefits of toothpaste, while the others describe social benefits. Hence, this method should only be used for item batteries or scales already checked for one dimensionality.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78070-7_14.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*, Springer Texts in Business and Economics,
https://doi.org/10.1007/978-3-031-78070-7_14

1.	It is important to buy a toothpaste that prevents cavities.	Disagree completely	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	Agree completely
2.	I like a toothpaste that gives me shiny teeth.	Disagree completely	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	Agree completely
3.	A toothpaste should strengthen your gums.	Disagree completely	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	Agree completely
4.	I prefer a toothpaste that freshens breath.	Disagree completely	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	Agree completely
5.	Prevention of tooth decay is not an important benefit offered by toothpaste.	Disagree completely	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	Agree completely
6.	The most important consideration in buying toothpaste is attractive teeth.	Disagree completely	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	Agree completely

Fig. 14.1 Toothpaste attributes

2. A second method of data reduction—known as factor analysis—is almost always used to carry out this check. Factor analysis uses correlation among individual items to reduce them to a small number of independent dimensions or factors, without presuming the one dimensionality of the scale. The correlation matrix of items indicates which statements exhibit similar patterns of responses. These items are then bundled into factors. Figure 14.2 shows that the health attributes *preventing cavities*, *strengthening gums,* and *not preventing tooth decay* are highly correlated. The same is true for the social attributes *whitening teeth*, *freshening breath*, and *making teeth attractive*. Hence, the preferred toothpaste attributes should be represented by two factors, not by one.

If those surveyed do not show similar patterns in their responses, then the high level of data heterogeneity and low level of data correlation render the results unusable for factor analysis. Backhaus et al. (2021, p. 418) give five criteria for determining whether the correlation matrix is suitable for running a factor analysis:

	cavity	whiteness	gums	fresh	decay	attract
cavity	1.0000					
whiteness	-0.0532	1.0000				
gums	0.8731	-0.1550	1.0000			
fresh	-0.0862	0.5722	-0.2478	1.0000		
decay	-0.8576	0.0197	-0.7778	-0.0066	1.0000	
attract	0.0042	0.6405	-0.0181	0.6405	-0.1364	1.0000

Fig. 14.2 Correlation matrix of the toothpaste attributes

Table 14.1 Measure of sampling adequacy (MSA) score intervals

MSA	[1.0; 0.9]	[0.9; 0.8]	[0.8; 0.7]	[0.7; 0.6]	[0.6; 0.5]	[0.5; 0.0]
Score	Marvellous	Meritorious	Middling	Mediocre	Miserable	Unacceptable

Source Kaiser and Rice (1974, p. 111)

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.660
Bartlett's Test of Sphericity	Approx. Chi-Square	111.314
	df	15
	Sig.	.000

Fig. 14.3 Correlation matrix check

1. Most of the correlation coefficients of the matrix must exhibit significant values.
2. The inverse of the correlation matrix must display a diagonal matrix with as many values close to zero for the non-diagonal elements as possible.
3. The Bartlett test (sphericity test) verifies whether the variables correlate. It assumes a normal distribution of item values and a χ^2 distribution of the test statistics. It checks the randomness of correlation matrix deviations from an identity matrix. A clear disadvantage with this test is that it requires a normal distribution. For any other form of distribution, the Bartlett test should not be used.
4. A factor analysis should not be performed when, in an anti-image covariance matrix (AIC),[1] more than 25% of elements below the diagonal have values larger than 0.09.
5. The Kaiser–Meyer–Olkin measure (or KMO measure) is generally considered by researchers to be the best method for testing the suitability of the correlation matrix for factor analysis, and it is recommended that it be performed before every factor analysis. It expresses a measure of sample adequacy (MSA) between zero and one. Calculated by all standard statistics software packages, MSA works for the sampling adequacy test for the entire correlation matrix as well as for each individual item. The KMO/MSA should be bigger or equal to 0.5. Table 14.1 suggests how KMO might be interpreted.

If the correlation matrix (see Fig. 14.3) turns out to be suitable for factor analysis, we can assume that regular patterns exist between responses and questions. This turns out to be the case for our toothpaste attribute survey, which possesses an acceptable MSA (0.660) and a significant result for the Bartlett test ($p < 0.05$).

[1] A discussion of the anti-image covariance matrix (AIC) lies beyond the scope of this book, though most software programmes are able to calculate it.

After checking the correlation matrix, we must identify its communalities. The communalities depend on the method of factor extraction, i.e. on the assumptions of the model. There are many types of factor analysis. Two are used most frequently:

- *Principal component analysis* assumes that individual variables can be described by a linear combination of the factors, i.e. that factors represent variable variances in their entirety. If there is a common share of variance for a variable determined by all factors, a communality of 100% (or 1) results. This desirable outcome occurs seldom in practice, as item batteries can rarely be reduced to a few factors representing a variance of all items. With principal component analysis, a communality less than one indicates a loss of information in the representation.
- *Principal factor analysis*, by contrast, assumes that variable variances can be separated into two parts. One part is determined by the joint variance of all variables in the analysis. The other part is determined by the specific variance for the variable in question. The total variance among the observed variable cannot be accounted for by its common factors. With principal factor analysis, the factors explain only the first variance component—the share of variance formed commonly by all variances—which means that the communality indicator must be less than one.

The difference in assumptions implied by the two different extraction methods can be summarized as follows: in principal component analysis, the priority is placed on representing each item exactly; in principal factor analysis, the hypothetical dimensions behind the items are determined so the correlation of individual items can be interpreted. This difference serves as the theoretical starting point for many empirical studies. For instance, the point of our toothpaste example is to identify the hypothetical factors behind the survey statements. A number of authors therefore see an advantage of method principal factor analysis over method principal component analysis (see Russell, 2002; Widaman, 1993). Therefore, one should use the principal factor analysis technique.

To check the quality of item representations by the factors, we need to use the factor loading matrix. The factor loading indicates the extent to which items are determined by the factors. The sum of all squared factor loadings for a factor is called the *eigenvalue*. Eigenvalues allow us to weigh factors based on the empirical data. When we divide the eigenvalue of an individual factor by the sum of eigenvalues of all extracted factors, we get a percentage value reflecting the perceived importance for all surveyed persons.

Say we extract from the toothpaste example two factors, one with an eigenvalue of 2.57 and the other with an eigenvalue of 1.87. This results in an importance of 42.84% for factor one and 31.13% for factor two. Later I will explain this importance in more detail (see Fig. 14.4).

The sum of a factor's eigenvalues strongly depends on the selection of items. The square of the factor loading matrix reproduces the variables' correlation

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotated Sums of Squared Loadings		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
1	2.73	45.52	45.52	2.57	42.84	42.84	2.54	42.34	42.34
2	2.22	36.97	82.49	1.87	31.13	73.96	1.90	31.62	73.96
3	.44	7.36	89.85						
4	.34	5.69	95.54						
5	.18	3.04	98.58						
6	.09	1.42	100.00						

Fig. 14.4 Eigenvalues and stated total variance for toothpaste attributes

matrix. If there are no large deviations (≤ 0.05) between the reproduced and the original correlation matrix, then the reproduction—the representability of the original data—is considered very good. Figure 14.5 shows the reproduced correlation matrix and the residuals from the original matrix for the toothpaste attribute survey. There is only one deviation above the level of difference (0.05), and it is minor (0.051). This means that both factors are highly representative of the original data.

	Toothpaste should…	prevent cavities	whiten teeth	strengthen gums	freshen breath	not prevent tooth decay	make teeth attractive
Reprod. Correlation	… prevent cavities	.928(b)	-.075	.873	-.110	-.850	.046
	… whiten teeth	-.075	.562(b)	-.161	.580	-.012	.629
	… strengthen gums	.873	-.161	.836(b)	-.197	-.786	-.060
	… freshen breath	-.110	.580	-.197	.600(b)	.019	.645
	… not prevent tooth decay	-.850	-.012	-.786	.019	.789(b)	-.133
	… make teeth attractive	.046	.629	-.060	.645	-.133	.723(b)
Residual(a)	… prevent cavities		.022	.000	.024	-.008	-.042
	… whiten teeth	.022		.006	-.008	.031	.012
	… strengthen gums	.000	.006		-.051	.008	.042
	… freshen breath	.024	-.008	-.051		-.025	-.004
	… not prevent tooth decay	-.008	.031	.008	-.025		-.003
	… make teeth attractive	-.042	.012	.042	-.004	-.003	

a Residuals are calculated between observed and reproduced correlations. There is one redundant residual with absolute values larger than 0.05 (at 6.0%).

b Reproduced communalities.

Fig. 14.5 Reproduced correlations and residuals

Though the number of factors can be set by the researcher himself (which is the reason why factor analysis is often accused of being susceptible to manipulation) some rules have crystallized over time. The most important of these is the *Kaiser criterion.* This rule takes into account all factors with an eigenvalue greater than one. Since eigenvalues less than one describe factors that do a poorer job of explaining variance than individual items do, this criterion is justified, hence its widespread acceptance. For instance, in our toothpaste example (see Fig. 14.4) an extraction of the third factor results in a smaller explanatory value than by adding one of the six items. Hence, a two-factor solution is more desirable in this case.

The Kaiser criterion is often accompanied by a scree plot in which the eigenvalues are plotted against the number of factors into a coordinate system in order of decreasing eigenvalues and increasing number of factors. When the curve forms an *elbow* towards a less steep decline, all further factors after the one starting the elbow are omitted. The plot in Fig. 14.6 applies to a three-factor solution.

After we set the number of factors, we interpret the results based on the individual items. Each item whose factor loading is greater than 0.5 is assigned to a factor. Figure 14.7 shows the factor loadings for attributes from our toothpaste

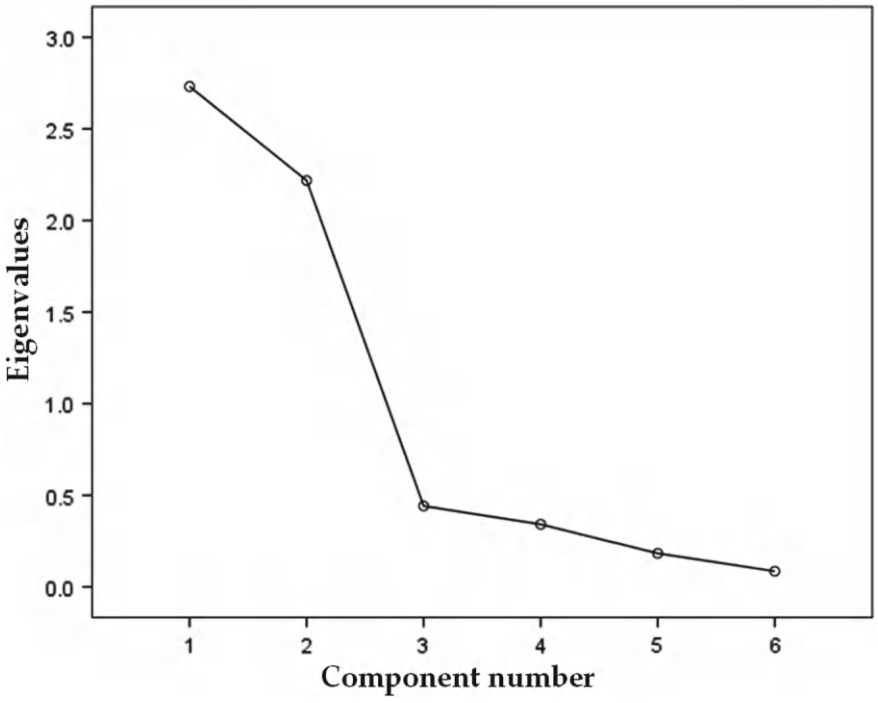


Fig. 14.6 Scree plot of the desirable toothpaste attributes

Toothpaste should	Unrotated factors		Rotated factors	
	1	2	1	2
… prevent cavities	.949	.168	.963	-.030
… whiten teeth	-.206	.720	-.054	.747
… strengthen gums	.914	.038	.902	-.150
… freshen breath	-.246	.734	-.090	.769
… not prevent tooth decay	-.849	-.259	-.885	-.079
… make teeth attractive	-.101	.844	.075	.847

Extraction method: Principal factor analysis

Rotation method: Varimax with Kaiser standardization

Fig. 14.7 Unrotated and rotated factor matrix for toothpaste attributes

example. Each variable is assigned to exactly one factor. The variables *prevent cavities*, *strengthen gums,* and *not prevent tooth decay* are loaded on factor 1, which describe the toothpaste's health-related attributes.

When positive factor loadings obtain, high factor values accompany high item values. When negative factor loadings obtain, low item values lead to high factor values and vice versa. This explains the negative sign in front of the factor loading for the variable *not prevent tooth decay*. People who assigned high values to *prevent cavities* and *strengthen gums* assigned low values to *not prevent tooth decay*. That is to say, those surveyed strongly prefer a toothpaste with health-related attributes.

The second factor describes the social benefits of toothpaste: *whiten teeth*, *freshen breath*, and *make teeth attractive*. Here too, the items correlate strongly, allowing the surveyed responses to be expressed by the second factor.

Sometimes, an individual item possesses factor loadings greater than 0.5 for several factors at the same time, resulting in a *multiple loading*. In these cases, we must take it into account for all the factors. If an item possesses factor loadings less than 0.5 for all its factors, we must either reconsider the number of factors or assign the item to the factor with the highest loading.

The factor matrix is normally rotated to facilitate the interpretation. In most cases, it is rotated orthogonally. This is known as a *varimax rotation* and it preserves the statistical independence of the factors. Figure 14.8 shows the effect of the varimax rotation on the values of a factor matrix. The variable *freshen breath* has an unrotated factor loading of − 0.246 for factor one (health attributes) and of 0.734 for factor two (social attributes). The varimax method rotates the total coordinate system from its original position but preserves the relationship between the individual variables. The rotation calibrates the coordinate system anew. Factor one now has the value of − 0.090 and factor two the value of 0.769 for the item *freshen breath*. The varimax rotation reduces the loading of factor one and increases the loading of factor two, making factor assignments of items more obvious. This is the basic idea of the varimax method: the coordinate system is rotated

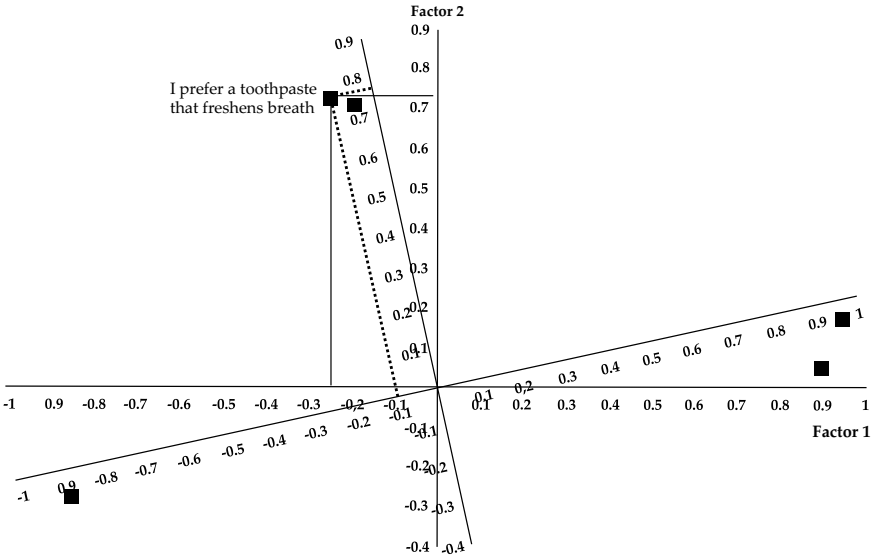


Fig. 14.8 Varimax rotation for toothpaste attributes

until the sum of the variances of the squared loadings is maximized. In most cases, this simplifies the interpretation.[2]

After setting the number of factors and interpreting their results, we must explain how the factor scores differ among the surveyed individuals. Factor scores generated by regression analysis provide some indications. The factor score of factor i can be calculated on the basis of linear combinations of the n original z-scores (z_j) of the surveyed person weighted with the respective values (α_{ij}) from the factor score coefficient matrix (see Fig. 14.9)

$$F_i = \alpha_{i1} \cdot z_1 + \alpha_{i2} \cdot z_2 + \alpha_{i3} \cdot z_3 + \alpha_{i4} \cdot z_4 + \cdots + \alpha_{in} \cdot z_n \qquad (14.1)$$

For each factor, every person receives a standardized value that assesses the scores given by individuals vis-à-vis the average scores given by all individuals. When the standardized factor score is positive, the individual scores are greater than the average of all responses, and vice versa. In the toothpaste dataset, person

[2] There are other rotation methods in addition to varimax, e.g. quartimax, equamax, promax, and oblimin. Even within varimax rotation, different calculation methods can be used, yielding minor (and usually insignificant) differences in the results.

Toothpaste should	Factor 1	Factor 2
… prevent cavities	.628	.101
… whiten teeth	-.024	.253
… strengthen gums	.217	-.169
… freshen breath	-.023	.271
… not prevent tooth decay	-.016	-.059
… make teeth attractive	.083	.500

Extraction method: Principal axis factoring
Rotation method: Varimax with Kaiser normalization

Fig. 14.9 Factor score coefficient matrix

#3[3] has a value of

$$F_1 = 0.628 \cdot 1.04 - 0.024 \cdot (-1.38) + 0.217 \cdot 1.41 - 0.023 \cdot (-0.07) - 0.166 \cdot (-1.31) + 0.083 \cdot (-0.84) = 1.14 \quad (14.2)$$

for factor one and a value of

$$F_2 = 0.101 \cdot 1.04 + 0.253 \cdot (-1.38) - 0.169 \cdot 1.41 + 0.271 \cdot (-0.07) - 0.059 \cdot (-1.31) + 0.5 \cdot (-0.84) = (-0.84) \quad (14.3)$$

for factor 2. This indicates a higher-than-average preference for health benefits and a lower-than-average preference for social benefits.

Factor analysis may only be used for metrically scaled variables. Some researchers have described certain conditions under which ordinal scales permit metric variables (Carifio & Perla, 2008; Pell, 2005). In any case, different item measurements (5-point scale vs. 7-point scale, say) require prior standardization. Factor values may only be calculated for persons or observations for which no missing values exist for any of the items being analysed. However, it is possible to impute missing data, enabling a broad analysis with the complete dataset. Different imputation techniques like mean imputation, regression imputation, stochastic imputation, multiple imputation, etc. are recommended in literature (see Enders, 2010).

[3] *Prevent cavities*: agree = 6 → $z = 1.04$; *whiten teeth*: agree = 2 → $z = -1.38$; *strengthen gums*: totally agree = 7 → $z = (1.41)$; *freshen breath*: neither agree or disagree = 4 → $z = (-0.07)$; *not prevent tooth decay*: totally disagree = 1 → $z = (-1.31)$; *make teeth attractive*: somewhat disagree = 3 → $z = (-0.84)$.

14.2 Factor Analysis with SPSS

This section uses the SPSS dataset *Chapter14_toothpaste_attributes.sav* (see supplementary material at https://doi.org/10.1007/978-3-031-78070-7_14). Select *Analyze* → *Dimension Reduction* → *Factor*… to open the *Factor Analysis* dialogue box. In the menu that opens, first select the variables (items) are to be used for the factor analysis. Follow then the steps outlined in Fig. 14.10. See Sect. 14.1 for interpretation of the results.

14.3 Factor Analysis with Stata

This section uses the Stata dataset *Chapter14_toothpaste_attributes.dta* (see supplementary material at https://doi.org/10.1007/978-3-031-78070-7_14). Select *Statistics* → *Multivariate analysis* → *Factor and principal component analysis* → *factor analysis* to open the *Factor Analysis* dialogue box. In the menu that opens (*Model*), first select the variables (items) are to be used for the factor analysis. Follow then the steps outlined in Fig. 14.11. See Sect. 14.1 for interpretation of the results.

14.4 Factor Analysis with R Commander

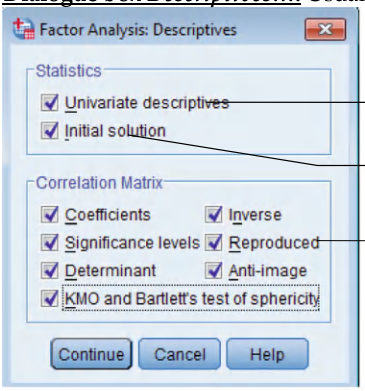
This section uses the R dataset *Chapter14_toothpaste_attributes.RData* (see supplementary material at https://doi.org/10.1007/978-3-031-78070-7_14). Select *Statistics* → *Dimensional analysis* → *Factor analysis*… to open the *Factor Analysis* dialogue box. For the principal component analysis, *Principal-components analysis*… should have been selected in the last step. In the menu that opens (*Data*), select the variables (items) are to be used for factor analysis. Click on *Options* and follow then the steps outlined in Fig. 14.12. See Sect. 14.1 for interpretation of the results.

14.5 Chapter Exercises

Exercise 84
Interpret the results of the following factor analysis about university students.

KMO and Bartlett's test		
Kaiser–Meyer–Olkin measure of sampling adequacy		0.515
Bartlett's test of sphericity	Approx. Chi-Square	37.813
	df	15
	Sig	0.001

Dialogue box *Descriptives…*: Usually, all options should be selected.

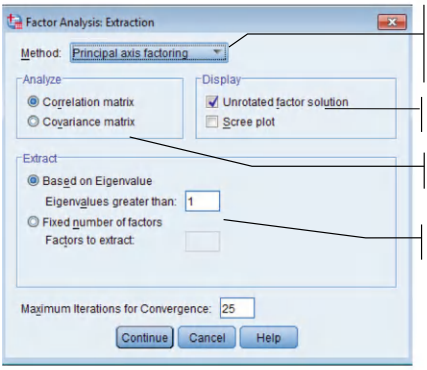


Displays mean, standard deviation, and number of valid cases for each variable.

Displays initial communalities, eigenvalues, and the percentage of variance explained.

Displays coefficients, significance levels, determinant, KMO and Bartlett's test of sphericity, inverse, reproduced, and anti-image correlation matrix.

Dialogue box *Extraction*:



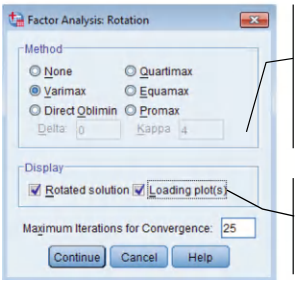
Specify the method of factor extraction. Available methods are principal components, unweighted least squares, generalized least squares, maximum likelihood, principal axis factoring, alpha factoring, and image factoring.

Displays the unrotated factor solution and a scree plot of the eigenvalues.

Specify either a correlation matrix or a covariance matrix. Usually, correlation matrix should be selected.

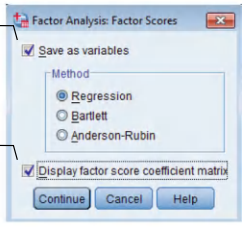
Base extraction on Eigenvalue (usually equal to one) or indicate a specific number of factors to extract.

Dialogue boxes *Rotation* and *Scores*:



Select the method of factor rotation. Available methods are varimax, direct oblimin, quartimax, equamax, or promax. Usually, varimax should be selected.

Dispays the rotated factor solution and loading plots for the first two or three factors.

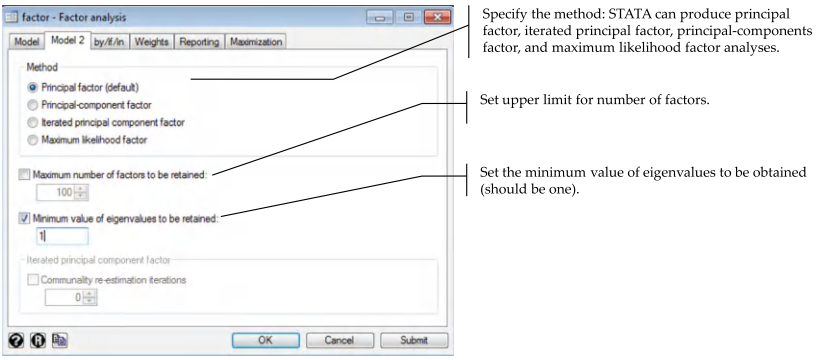


Creates one new variable for each factor in the data set.

Displays the factor score coefficient matrix.

Fig. 14.10 Factor analysis with SPSS

Under Model 2:



Rotation Commands

- Type in the command line *rotate, varimax* and hit <enter>.

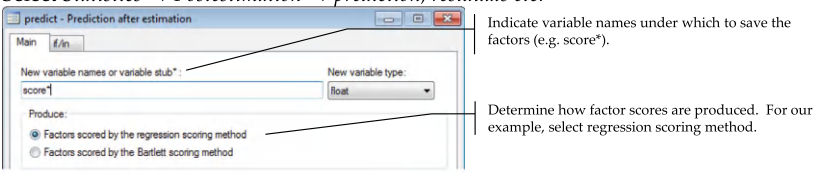
Reports Commands

Statistics → Multivariate analysis → Postestimation reports and statistics

- In the *Reports and statistics* subcommand: Select commands (KMO statistics, etc.). Click *Submit* to initiate the commands.

Saving Factor Scores

Select *Statistics → Postestimation → prediction, residuals etc.*



Important syntax commands for factor analysis:

factor; factor postestimation; pca; pca postestimation; rotate; rotatemat; scoreplot; screeplot; alpha; canon; estimates, estat; predict

Fig. 14.11 Factor analysis with Stata

Anti-image matrices							
		Intelligence quotient	Independent preparation	Motivation	Self-confidence	Assessment preparation	Contact hours
Anti-image covariance	Intelligence quotient	0.397	− 0.100	− 0.121	− 0.114	0.076	0.052
	Independent preparation (in hours)	− 0.100	0.191	0.115	− 0.095	− 0.065	− 0.112

(continued)

(continued)

Anti-image matrices							
		Intelligence quotient	Independent preparation	Motivation	Self-confidence	Assessment preparation	Contact hours
	Motivation [1: *very low* to 50: *very high*]	− 0.121	0.115	0.202	− 0.139	0.059	− 0.124
	Self-confidence [1: *very low* to 50: *very high*]	− 0.114	− 0.095	− 0.139	0.416	− 0.017	0.104
	Assessment preparation (in hours)	0.076	− 0.065	0.059	− 0.017	0.391	− 0.061
	Contact hours (in hours)	0.052	− 0.112	− 0.124	0.104	− 0.061	0.114
Anti-image correlation	Intelligence quotient	0.643[a]	− 0.362	− 0.427	− 0.281	0.192	0.246
	Independent preparation (in hours)	− 0.362	0.487[a]	0.584	− 0.338	− 0.237	− 0.755
	Motivation [1: *very low* to 50: *very high*]	− 0.427	0.584	0.385[a]	− 0.479	0.210	− 0.815
	Self-confidence [1: *very low* to 50: *very high*]	− 0.281	− 0.338	− 0.479	0.536[a]	− 0.042	0.475
	Assessment preparation (in hours)	0.192	− 0.237	0.210	− 0.042	0.816[a]	− 0.288
	Contact hours (in hours)	0.246	− 0.755	− 0.815	0.475	− 0.288	0.450[a]

[a] Measures of sampling adequacy (MSA)

Communalities		
	Initial	Extraction
Intelligence quotient	0.603	0.725
Independent preparation (in hours)	0.809	0.713
Motivation [1 = very low to 50 = very high]	0.798	0.622
Self-confidence [1 = very low to 50 = very high]	0.584	0.556
Assessment preparation (in hours)	0.609	0.651
Contact hours (in hours)	0.886	0.935

Extraction method: Principal axis factoring

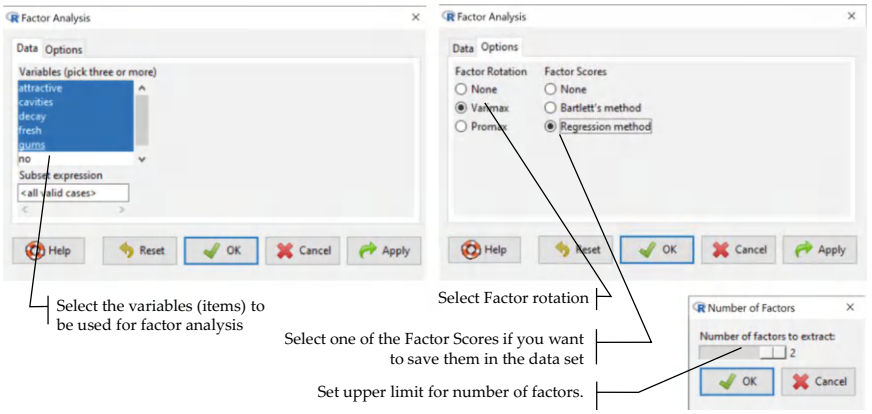


Fig. 14.12 Factor analysis with R Commander

Total variance explained									
Factor	Initial Eigenvalues			Extraction sums of squared loadings			Rotation sums of squared loadings		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
1	2.54	42.39	42.39	2.32	38.62	38.62	2.27	37.88	37.88
2	2.24	37.32	79.72	1.88	31.39	70.01	1.93	32.13	70.01
3	0.57	9.51	89.23						
4	0.34	5.74	94.97						
5	0.24	4.04	99.01						
6	0.06	0.99	100.00						

Extraction method: Principal axis factoring

Rotated factor matrix[a]		
	Factor	
	1	2
Intelligence quotient	− 0.004	0.851
Independent preparation (in hours)	0.839	0.091
Motivation [1 = very low to 50 = very high]	0.264	0.743
Self-confidence [1 = very low to 50 = very high]	− 0.166	0.727
Assessment preparation (in hours)	0.759	− 0.273
Contact hours (in hours)	0.946	0.201

Extraction method: Principal axis factoring
Rotation method: Varimax with Kaiser normalization[a]
[a] Rotation converged in 3 iterations

14.6 Exercise Solutions

Solution 84

- The KMO test: KMO measure of sampling adequacy = 0.515 (> 0.5) and Bartlett's test of sphericity is significant ($p = 0.001 < 0.05$), so the correlations between the items are large enough. Hence, it is meaningful to perform a factor analysis.
- Anti-image correlation matrix: In this matrix, the individual MSA values of each item on the diagonal should be bigger than 0.5. In the given case, some MSA values are smaller than 0.5. Those items should be omitted step by step.
- Total variance explained table: Components one and two have eigenvalues > 1. A two-factor solution thus seems to be appropriate. The two factors are able to explain 70% of the total variance.
- Communalities: 70.2% of the total variance in *Intelligence Quotient* is explained by the two underlying factors, etc.
- Rotated component matrix: Factor 1: Individual workload; Factor 2: Individual capacity.

References

Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical Education, 42*, 1150–1152.

Backhaus, K., Erichson, B., Gensler, S., Weiber, R., & Weiber, T. (2021). *Multivariate Analysemethoden. Eine Anwendungsorientierte Einführung* (16th ed.). Springer.

Enders, C. K. (2010). *Applied missing data analysis.* Guilford Press.

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement, 34*, 111–117.

Malhotra, N. K., Nunan, D., & Birks, D. F. (2017). *Marketing research. An applied approach* (5th ed.). Pearson.

Pell, G. (2005). Use and misuse of Likert scales. *Medical Education, 39*, 970.

Russell, D. W. (2002). In search of underlying dimension: The use (and abuse) of factor analysis. *Personality and Social Psychological Bulletin, 28*(12), 1629–1646.

Widaman, K. F. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research, 28*(3), 263–311.

List of Formulas

Measures of Central Tendency

Mean (from raw data): $\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$

Mean (from a frequency table): $\bar{x} = \frac{1}{n}\sum_{v=1}^{k} x_v \cdot n_v = \sum_{v=1}^{k} x_v \cdot f_v$

Mean (from classed data): $\bar{x} == \frac{1}{n}\sum_{v=1}^{k} n_v m_v = \sum_{v=1}^{k} f_v m_v$
where m_v is the mean of class number v.

Geometric Mean: $\bar{x}_{\text{geom}} = \sqrt[n]{(x_1 \cdot x_2) \cdot \cdots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$

Geometric Mean for Change Rates:
$\bar{p}_{\text{geom}} = \sqrt[n]{(1+p_1) \cdot (1+p_2) \cdot \cdots \cdot (1+p_n)} - 1 = \sqrt[n]{\prod_{i=1}^{n}(1+p_i)} - 1$

Harmonic Mean (unweighted) for k observations: $\bar{x}_{\text{harm}} = \frac{k}{\sum_{i=1}^{k} \frac{1}{x_i}}$

Harmonic Mean (weighted) for k observations: $\bar{x}_{\text{harm}} = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$

Median (from classed data): $\tilde{x} = x_{0.5} = x_{i-1}^{\text{UP}} + \frac{0.5 - F\left(x_{i-1}^{\text{UP}}\right)}{f(x_i)}\left(x_i^{\text{UP}} - x_i^{\text{LOW}}\right)$

Median (from raw data) for an odd number of observations (n): $\tilde{x} = x_{\left(\frac{n+1}{2}\right)}$

Median (from Raw Data) for an even number of observations (n): $\tilde{x} = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)$

Quantile (from Raw Data) using the Weighted Average Method: We first have to determine the product $(n+1) \cdot p$. The result consists of an integer before the decimal mark and a decimal fraction after the decimal mark (i, f). The integer (i) helps indicate the values between which the desired quantile lies—namely, between the observations (i) and $(i+1)$, assuming that (i) represents the ordinal numbers of the ordered dataset. The figures after the decimal mark can be used to locate the position between the values with the following formula: $\tilde{x} = (1-f) \cdot x_{(i)} + f \cdot x_{(i+1)}$

Quantile (from classed data): $x_p = x_{i-1}^{*} + \frac{p - F(x_{i-1}^{*})}{f_i}\Delta x_i$.

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-031-78070-7

Dispersion Parameters

Interquartile Range: $\text{IQR} = x_{0.75} - x_{0.25}$
Mid-Quartile Range: $\text{MQR} = 0.5(x_{0.75} - x_{0.25})$
Range: $\text{Range} = \text{Max}(x_i) - \text{Min}(x_i)$
Median Absolute Deviation: $\text{MAD} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \tilde{x}|$
Empirical Variance: $\text{Var}(x)_{\text{emp}} = S^2_{\text{emp}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$
Empirical Standard Deviation: $S_{\text{emp}} = \sqrt{\text{Var}(x)_{\text{emp}}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2}$
(Unbiased Sample) Variance: $\text{Var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$
(Unbiased Sample) Standard Deviation: $S = \sqrt{\text{Var}(x)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$
Coefficient of Variation: $V = \frac{S}{|\overline{x}|}, \quad \overline{x} \neq 0$

Measurement of Concentration

Concentration Ratio CR_g: The percentage of a quantity (e.g. revenues) achieved by g statistical units with the highest trait values.
Herfindahl Index: $H = \sum_{i=1}^{n} f(x_i)^2$
Gini Coefficient for Unclassed Ordered Raw Data: $\text{GINI} = \frac{2\sum_{i=1}^{n} i \cdot x_i - (n+1)\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i}$
Gini Coefficient for Unclassed Ordered Relative Frequencies: $\text{GINI} = \frac{2\sum_{i=1}^{n} i \cdot f_i - (n+1)}{n}$
Normalized Gini Coefficient ($\text{GINI}_{\text{norm.}}$): Normalized by Multiplying Each of the Above Formulas by $\frac{n}{n-1}$

Skewness and Kurtosis

Skewness (Yule & Pearson): $\text{Skew} = \frac{3 \cdot (\overline{x} - \tilde{x})}{S}$
Skewness (Third Central Moment): $\text{Skew} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{S^3}$
Kurtosis: $\text{Kurt} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{S^4}$

Bivariate Association

Chi-Square: $\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{\left(n_{ij} - n_{ij}^e\right)^2}{n_{ij}^e}$
Phi: $\text{PHI} = \sqrt{\frac{\chi^2}{n}}$
Contingency Coefficient: $C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \in [0; 1]$

Cramer's V: $V = \sqrt{\frac{\chi^2}{n \cdot (\min(k,m)-1)}} = \varphi \cdot \sqrt{\frac{1}{\min(k,m)-1}} \in [0;1]$

Covariance: $\text{cov}(x;y) = S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}$

Bravais-Pearson Correlation: $r = \frac{S_{xy}}{S_x S_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2\right)}}$

Partial Correlation: $r_{xy.z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{\left(1-r_{xz}^2\right)\cdot\left(1-r_{yz}^2\right)}}$

Point-Biserial Correlation: $r_{\text{pb}} = \frac{\overline{y}_1 - \overline{y}_0}{S_y}\sqrt{\frac{n_0 \cdot n_1}{n^2}}$

with

- n_0: number of observations with the value $x = 0$ of the dichotomous trait
- n_1: number of observations with the value $x = 1$ of the dichotomous trait
- n: total sample size $n_0 + n_1$
- $\overline{y}_0$: mean of metric variables (y) for the cases $x = 0$
- $\overline{y}_1$:mean of metric variables (y) for the cases $x = 1$
- S_y: standard deviation of the metric variable (y)

Spearman's Rank Correlation:

$\rho = \frac{S_{xy}}{S_x S_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(R(x_i)-\overline{R(x)})(R(y_i)-\overline{R(y)})}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(R(x_i)-\overline{R(x)})^2\right)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}(R(y_i)-\overline{R(y)})^2\right)}}$

Spearman's Rank Correlation (Short-Hand Version):

$\rho = 1 - \frac{6\cdot\sum_{i=1}^{n} d_i^2}{n\cdot(n^2-1)}$ with $d_i = (R(x_i) - R(y_i))$

Spearman's Rank Correlation (Short-Hand Version With Rank Ties):

$\rho_{\text{korr}} = \frac{2\cdot\left(\frac{N^3-N}{12} - N\right) - T - U - \sum_{i=1}^{n} d_i^2}{2\cdot\sqrt{\left(\frac{N^3-N}{12} - T\right)\cdot\left(\frac{N^3-N}{12} - U\right)}}$

With

- T as the length of b tied ranks among x variables: $T = \frac{\sum_{i=1}^{b}(t_i^3 - t_i)}{12}$, where t_i equals the number of tied ranks in the ith of b groups for the tied ranks of the x variables.
- U as the length of c tied ranks of y variables: $U = \frac{\sum_{i=1}^{c}(u_i^3 - u_i)}{12}$, where u_i equals the number of tied ranks in the ith of c groups for the tied ranks of the y variables.

Kendall's τ_a (Without Rank Ties): $\tau_a = \frac{P-I}{n\cdot(n-1)/2}$

Kendall's τ_b (With Rank Ties): $\tau_b = \frac{P-I}{\sqrt{\left(\frac{n\cdot(n-1)}{2} - T\right)\left(\frac{n\cdot(n-1)}{2} - U\right)}}$

where

- T is the length of the b tied ranks of x variables: $T = \frac{\sum_{i=1}^{b} t_i(t_i-1)}{2}$, and t_i is the number of tied ranks in the ith of b groups of tied ranks for the x variables.
- and U is the length of c tied ranks of the y variables: $U = \frac{\sum_{i=1}^{c} u_i(u_i-1)}{2}$, and u_i is the number of tied ranks in the ith of c groups of tied ranks for the y variables.

Biserial Rank Correlation (Without Rank Ties): $r_{\text{bis}R} = \frac{2}{n}\cdot\left(\overline{R(y_1)} - \overline{R(y_0)}\right)$

Regression Analysis

Intercept of a Bivariate Regression Line: $\alpha = \overline{y} - \beta \cdot \overline{x}$

Slope of a Bivariate Regression Line: $\beta = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\text{cov}(x;y)}{S_x^2} = \frac{r \cdot S_y}{S_x} = \frac{n\sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$

Coefficients of a Multiple Regression: $\beta = \left(X'X\right)^{-1}X'y$

R^2/Coefficient of Determination: $R^2 = \frac{\text{RSS}}{\text{TSS}} = \frac{\text{SS}_{\hat{Y}}}{\text{SS}_Y} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{SS}_e}{\text{SS}_Y} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$

Adjusted R^2/Coefficient of Determination: $R^2_{\text{adj}} = R^2 - \frac{\left(1-R^2\right)(k-1)}{(n-k)} = 1 - \left(1 - R^2\right)\frac{n-1}{n-k}$

Logistic Regression

Odds: $\text{odds} = \frac{p}{(1-p)}$

Log-odds: $\ln(\text{odds}) = \ln\left(\frac{p}{(1-p)}\right)$

Individual a priori probability: $p = \frac{1}{\left(1+e^{-(\beta_0+\beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k)}\right)}$

Odds ratio: $\text{Exp}(B) = \frac{\text{odds}(x_1+1)}{\text{odds}(x_1)} = \frac{e^{(\beta_0+\beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k)} \cdot e^{\beta_1}}{e^{(\beta_0+\beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k)}} = e^{\beta_1}$

Likelihood ratio test: $\chi^2 = (-2\text{LL}_0) + (2\text{LL}_{\text{full}})$

Wald test: $W = \left(\frac{\widehat{\beta}}{\text{S.E.}\left(\widehat{\beta}\right)}\right)^2$

McFadden R^2: $\text{R}^2_{\text{McF}} = 1 - \left(\frac{\text{LL}_{\text{full}}}{\text{LL}_0}\right)$

Cox-Snell R^2: $R^2_{\text{CS}} = 1 - \left(\frac{L_0}{L_{\text{full}}}\right)^{\frac{2}{n}}$

Nagelkerke R^2: $R^2_N = 1 - \left(\frac{R^2_{\text{CS}}}{1 - L_0^{(2/k)}}\right)$

Index Numbers

Laspeyres Price Index: $P^L_{0,t} = \frac{\sum_{i=1}^{n} \frac{p_{i,t}}{p_{i,0}} \cdot p_{i,0} \cdot q_{i,0}}{\sum_{i=1}^{n} p_{i,0} \cdot q_{i,0}} = \frac{\sum_{i=1}^{n} p_{i,t} \cdot q_{i,0}}{\sum_{i=1}^{n} p_{i,0} \cdot q_{i,0}}$

Laspeyres Quantity Index: $Q^L_{0,t} = \frac{\sum_{i=1}^{n} q_{i,t} \cdot p_{i,0}}{\sum_{i=1}^{n} q_{i,0} \cdot p_{i,0}}$

Paasche Price Index: $P^P_{0,t} = \frac{\sum_{i=1}^{n} p_{i,t} \cdot q_{i,t}}{\sum_{i=1}^{n} p_{i,0} \cdot q_{i,t}}$

Paasche Quantity Index: $Q^P_{0,t} = \frac{\sum_{i=1}^{n} q_{i,t} \cdot p_{i,t}}{\sum_{i=1}^{n} q_{i,0} \cdot p_{i,t}}$

Fisher Price Index: $P^F_{0,t} = \sqrt{P^L_{0,t} \cdot P^P_{0,t}}$

Fisher Quantity Index: $Q^F_{0,t} = \sqrt{Q^L_{0,t} \cdot Q^P_{0,t}}$

Value Index: $V_{0,t} = \frac{\sum_{i=1}^{n} p_{i,t} \cdot q_{i,t}}{\sum_{i=1}^{n} p_{i,0} \cdot q_{i,0}} = Q^F_{0,t} \cdot P^F_{0,t} = Q^L_{0,t} \cdot P^P_{0,t} = Q^P_{0,t} \cdot P^L_{0,t}$

Deflating Time Series by Price Index: $L^{\text{real}}_t = \frac{L^{\text{nominal}}_t}{P^L_{0,t}}$

Base Shift of an Index: $I^{\text{new}}_{\tau,t} = \frac{I^{\text{old}}_{0,t}}{I^{\text{old}}_{0,\tau}}$

Chaining an Index (Forward Extrapolation): $\tilde{I}_{0,t} = \begin{cases} I^1_{0,t} & \text{for } t \leq \tau \\ I^1_{0,\tau} \cdot I^2_{\tau,t} & \text{for } t > \tau \end{cases}$

Chaining an Index (Backward Extrapolation): $\tilde{I}_{0,t} = \begin{cases} \frac{I^1_{0,\tau}}{I^2_{\tau,t}} & \text{for } t < \tau \\ I^2_{\tau,t} & \text{for } t \geq \tau \end{cases}$

Combinatorics

Permutation of N elements without repetition: $P^N_N = N!$

Permutation of N elements with repetition (k different groups of elements exist): $P^N_{n_1;\ldots;n_k} = \frac{N!}{n_1! n_2! \cdot \cdots \cdot n_k!}$, with $\sum_{i=1}^{k} n_i = N$

Combination without repetition (order does not matter): $C^N_n = \binom{N}{n} = \frac{N!}{n!(N-n)!}$

Combination with repetition (order does not matter): $\tilde{C}^N_n = \binom{N+n-1}{n}$

Variation without repetition (order matters): $V^N_n = n! \binom{N}{n} = \frac{N!}{(N-n)!}$

Variation with repetition (order matters): $\tilde{V}^N_n = N^n$

Calculating Probabilities

Addition principle of nondisjoint events: $P\left(\bigcup_{i=1}^{m} A_i\right) = \sum_{i=1}^{m} P(A_i)$

Complementary events: $A \cap B = \{\}$; $P(A \cap B) = 0$

Inclusion–exclusion principle of nondisjoint events: $P(A \cup B) = P(A) + P(B) - (A \cap B)$

Law of multiplication with independent events: $P(A \cap B) = P(A) \cdot P(B)$

Law of multiplication: $P(A \cap B) = P(A|B) \cdot P(B)$

Law of total probability: $P(A) = P(A|B) \cdot P(B) + P(A|C) \cdot P(C) + \cdots + P(A|Z) \cdot P(Z)$

Bayes' theorem: $P(A|B) = P(B|A) \cdot P(A)/P(B)$

Discrete Distributions

Binomial distribution: $B(n,k,p) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = \frac{n!}{k!(n-k)!} \cdot p^k \cdot (1-p)^{n-k}$

- Expected value of a random variable in a binomial distribution: $E(X) = n \cdot p$
- Variance of a random variable in a binomial distribution:$\mathrm{Var}(X) = n \cdot p(1-p)$
- Random variables in binomial distributions approximately follow a …
  - … normal distribution $\left(N\left(n \cdot p\sqrt{n \cdot p(1-p)}\right)\right)$, if $n \cdot p \cdot (1-p) > 9$.
  - … Poisson distribution ($\mathrm{Po}(n \cdot p)$), if $n \cdot p \leq 10$ and $n \geq 1500 \cdot p$.

Hypergeometric distribution: $H(N, M, n, x) = \dfrac{\binom{N-M}{n-x} \cdot \binom{M}{x}}{\binom{N}{n}}$

- Expected value of a random variable in a hypergeometric distribution: $E(X) = n \cdot M/N$
- Variance of a random variable in a hypergeometric distribution: $\mathrm{Var}(X) = n\frac{M}{N}\left(1 - \frac{M}{N}\right) \cdot \frac{N-n}{N-1}$
- Random variables in a hypergeometric distribution ($H(N;M;n)$) approximately follow …
  - … a normal distribution $\left(N\left(\underbrace{n\frac{M}{N}}_{E(x)}; \underbrace{\sqrt{\frac{n(N-n)}{N-1} \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)}}_{\sigma}\right)\right)$, if $n > 30$ and $0.1 < \frac{M}{N} < 0.9$
  - … a Poisson distribution $\left(Po\left(\underbrace{n\frac{M}{N}}_{E(x)}\right)\right)$, if $0.1 \geq \frac{M}{N}$ or $\frac{M}{N} \geq 0.9$ and $n > 30$ and $\frac{n}{N} < 0.05$,
  - … a binomial distribution $\left(B\left(n; \frac{M}{N}\right)\right)$, if $n > 10$ and $0.1 < \frac{M}{N} < 0.9$ and $\frac{n}{N} < 0.05$.

Poisson Distribution: $P(X) = \frac{\lambda^x}{x!}e^{-\lambda}$

- Expected value of a random variable in a Poisson distribution: $E(X) = \mu = \lambda$
- Variance of a random variable in a Poisson distribution: $\mathrm{Var}(X) = \lambda$
- The Poisson distribution with $\lambda = n \cdot p$ is derived from the binomial distribution, which itself can be approximated using the continuous normal distribution. When $\lambda \geq 10$, therefore, the Poisson distribution can also be approximated using the continuous normal distribution with ($N(\mu = \mu; \sigma = \sqrt{\mu})$).

Continuous Distributions

Continuous uniform distribution: $f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a < X \leq b \\ 0 & \text{otherwise} \end{cases}$

- Expected value of a random variable in a continuous uniform distribution: $E(X) = \frac{a+b}{2}$
- Variance of a random variable in a continuous uniform distribution: $\text{Var}(X) = \frac{(b-a)^2}{12}$

Normal distribution: $f_x(x) = \frac{1}{\sigma\sqrt{2 \cdot \Pi}} \cdot e^{-\frac{(x-\mu)^2}{2 \cdot \sigma^2}}$

- Expected value of a random variable in a normal distribution: $E(X) = \mu$
- Variance of a random variable in a normal distribution: $\text{Var}(X) = \sigma^2$
- Standardized variable (z-transformation): $P\left(X_{\text{lower}} \leq X \leq X_{\text{upper}}\right) = P\left(\frac{X_{\text{lower}}-\mu}{\sigma} \leq Z \leq \frac{X_{\text{upper}}-\mu}{\sigma}\right)$
- Random variables in a normal distribution are reproductive, which means that the merging of two (or more) random variables in a binomial distribution leads to another random variable in a normal distribution (see example for two variables): $N\left(\mu_1 + \mu_2; \sqrt{\sigma_1^2 + \sigma_2^2}\right)$

Confidence Interval for the Mean

Calculating confidence intervals for means: see Fig. 8.7.
Length of a two-sided confidence interval for means: $E = 2 \cdot z_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{\bar{x}} = 2 \cdot z_{1-\frac{\alpha}{2}} \frac{S_{\text{theor}}}{\sqrt{n}} = 2 \cdot z_{1-\frac{\alpha}{2}} \frac{S_{\text{emp}}}{\sqrt{n-1}}$

Planning sample size for confidence intervals for means: $n = \frac{2^2 \cdot z^2_{1-\frac{\alpha}{2}} \cdot S^2_{\text{theor}}}{E^2} = \frac{2^2 \cdot z^2_{1-\frac{\alpha}{2}} \cdot S^2_{\text{emp}}}{E^2} + 1$

Confidence Interval for Proportions

Calculating confidence intervals for proportions: see Fig. 8.10.
Length of a two-sided confidence interval for proportions:$E = 2 \cdot z_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{\bar{x}} = 2 \cdot z_{1-\frac{\alpha}{2}} \frac{S_{\text{theor}}}{\sqrt{n}} = 2 \cdot z_{1-\frac{\alpha}{2}} \frac{S_{\text{emp}}}{\sqrt{n-1}}$

Planning sample size for confidence intervals for proportions: $E^2 = \frac{2^2 \cdot z^2_{1-\frac{\alpha}{2}} \cdot S^2_{\text{theor}}}{n} = \frac{2^2 \cdot z^2_{1-\frac{\alpha}{2}} \cdot \overline{p} \cdot (1-\overline{p})}{n}$

Confidence Interval for Variances

Two-sided:

$$P\left(\frac{(n-1)\cdot S^2_{\text{theor}}}{\chi^2_{1-\frac{\alpha}{2};n-1}} \le \sigma^2 \le \frac{(n-1)\cdot S^2_{\text{theor}}}{\chi^2_{\frac{\alpha}{2};n-1}}\right) = P\left(\frac{n\cdot S^2_{\text{emp}}}{\chi^2_{1-\frac{\alpha}{2};n-1}} \le \sigma^2 \le \frac{n\cdot S^2_{\text{emp}}}{\chi^2_{\frac{\alpha}{2};n-1}}\right) = 1-\alpha$$

One-sided:

$$P\left(\frac{(n-1)\cdot S^2_{\text{theor}}}{\chi^2_{1-\alpha;n-1}} = \frac{n\cdot S^2_{\text{emp}}}{\chi^2_{1-\alpha;n-1}} \le \sigma^2\right) = 1-\alpha \text{ or } P\left(\sigma^2 \le \frac{(n-1)\cdot S^2_{\text{theor}}}{\chi^2_{\alpha;n-1}} = \frac{n\cdot S^2_{\text{emp}}}{\chi^2_{\alpha;n-1}}\right) = 1-\alpha$$

One-Sample *t*-Test

One-sample *Z*-test/one-sample *t*-test: see Fig. 9.6.

For a one-tailed hypothesis test, the *p*-value is calculated by: $p = 2 \cdot \left(1 - P\left(t^{n-1} \le t_{\text{critical}} = \left|\frac{\overline{x}-\mu_o}{\sigma_{\overline{x}}}\right|\right)\right)$

For a two-tailed hypothesis, the *p*-values are calculated by:

$$p_{\text{left}} = P\left(t^{n-1} \le t_{\text{critical}} = \frac{\overline{x}-\mu_0}{\sigma_{\overline{x}}}\right) \text{ for } \mathrm{H}_1 : \mu < \mu_0$$

$$p_{\text{right}} = \left(1 - P\left(t^{n-1} \le t_{\text{critical}} = \frac{\overline{x}-\mu_0}{\sigma_{\overline{x}}}\right)\right) \text{ for } \mathrm{H}_1 : \mu > \mu_0$$

Kruskal–Wallis-Test (H Test)

Hypotheses: $H_0 : E(\overline{R_1}) = E(\overline{R_2}) = \cdots = E(\overline{R_k}); H_1 : E(\overline{R_i}) \neq E(\overline{R_j})$, for at least one $i \neq j$

Test statistic: $H^{\text{finite}}_{\text{corr}} = \frac{H}{C} = \frac{\frac{N-1}{N}\sum_{i=1}^{k}\left(\frac{\left(\overline{R}_j - \frac{N+1}{2}\right)^2}{\frac{N^2-1}{12\cdot n_i}}\right)}{1-\frac{\sum_{j=1}^{m} t_j^3 - t_j}{N^3-N}} \sim \chi^2_{k-1}$

Decision: H_0 has to be rejected if $\chi^2_{k-1} \le H^{\text{finite}}_{\text{corr}}$

Chi-Square Test of Independence

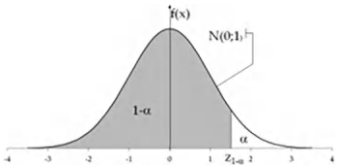
Calculation of Chi-Square: $\chi^2_{\text{emp}} = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(n_{ij} - n^e_{ij}\right)^2}{n^e_{ij}}$

Decision: H_0 (Independence) has to be rejected if: $\chi^2_{\text{emp}} > \chi^2_{1-\alpha;(m-1)\cdot(q-1)}$.

Appendix A: The Standard Normal Distribution

The table contains abscissa values for between 0.00 and 3.29 along with the cumulative probabilities $\Phi(z) = P(Z \leq z)$ For negative z values, $\Phi(-z) = 1 - \Phi(z)$.

Example $\Phi(2.33) = 0.9901$.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015

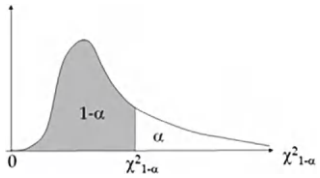
© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-031-78070-7

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995

Appendix B: The Chi-Squared Distribution

The table contains the quantiles corresponding to the cumulative probability $p = (1 - a)$ for a particular number of degrees of freedom.



$p = (1 - q)$						
n	0.001	0.005	0.025	0.005	0.1	0.5
1	0.000	0.000	0.001	0.000	0.016	0.455
2	0.002	0.010	0.051	0.010	0.211	1.386
3	0.024	0.072	0.216	0.072	0.584	2.366
4	0.091	0.207	0.484	0.207	1.064	3.357
5	0.210	0.412	0.831	0.412	1.610	4.351
6	0.381	0.676	1.237	0.676	2.204	5.348
7	0.598	0.989	1.690	0.989	2.833	6.346
8	0.857	1.344	2.180	1.344	3.490	7.344
9	1.152	1.735	2.700	1.735	4.168	8.343
10	1.479	2.156	3.247	2.156	4.865	9.342
11	1.834	2.603	3.816	2.603	5.578	10.341
12	2.214	3.074	4.404	3.074	6.304	11.340
13	2.617	3.565	5.009	3.565	7.042	12.340
14	3.041	4.075	5.629	4.075	7.790	13.339
15	3.483	4.601	6.262	4.601	8.547	14.339
16	3.942	5.142	6.908	5.142	9.312	15.338

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-031-78070-7

$p = (1 - q)$						
n	0.001	0.005	0.025	0.005	0.1	0.5
17	4.416	5.697	7.564	5.697	10.085	16.338
18	4.905	6.265	8.231	6.265	10.865	17.338
19	5.407	6.844	8.907	6.844	11.651	18.338
20	5.921	7.434	9.591	7.434	12.443	19.337
21	6.447	8.034	10.283	8.034	13.240	20.337
22	6.983	8.643	10.982	8.643	14.041	21.337
23	7.529	9.260	11.689	9.260	14.848	22.337
24	8.085	9.886	12.401	9.886	15.659	23.337
25	8.649	10.520	13.120	10.520	16.473	24.337
26	9.222	11.160	13.844	11.160	17.292	25.336
27	9.803	11.808	14.573	11.808	18.114	26.336
28	10.391	12.461	15.308	12.461	18.939	27.336
29	10.986	13.121	16.047	13.121	19.768	28.336
30	11.588	13.787	16.791	13.787	20.599	29.336
40	17.916	20.707	24.433	20.707	29.051	39.335
60	31.738	35.534	40.482	35.534	46.459	59.335
80	46.520	51.172	57.153	51.172	64.278	79.334
100	61.918	67.328	74.222	67.328	82.358	99.334

The Chi-Squared Distribution (Continued)

The table contains the quantiles corresponding to the cumulative probability $p - (1 - a)$ for a particular number of degrees of freedom. Example: $\chi^2_{1-0.1;20} = \chi^2_{90\%;20} = 28.412$.

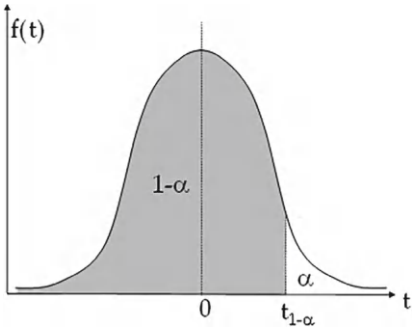
$p = (1 - a)$						
n	0.9	0.95	0.975	0.99	0.995	0.999
1	2.706	3.841	5.024	6.635	7.879	10.828
2	4.605	5.991	7.378	9.210	10.597	13.816
3	6.251	7.815	9.348	11.345	12.838	16.266
4	7.779	9.488	11.143	13.277	14.860	18.467
5	9.236	11.070	12.833	15.086	16.750	20.515
6	10.645	12.592	14.449	16.812	18.548	22.458
7	12.017	14.067	16.013	18.475	20.278	24.322
8	13.362	15.507	17.535	20.090	21.955	26.124

$p = (1 - a)$						
n	0.9	0.95	0.975	0.99	0.995	0.999
9	14.684	16.919	19.023	21.666	23.589	27.877
10	15.987	18.307	20.483	23.209	25.188	29.588
11	17.275	19.675	21.920	24.725	26.757	31.264
12	18.549	21.026	23.337	26.217	28.300	32.909
13	19.812	22.362	24.736	27.688	29.819	34.528
14	21.064	23.685	26.119	29.141	31.319	36.123
15	22.307	24.996	27.488	30.578	32.801	37.697
16	23.542	26.296	28.845	32.000	34.267	39.252
17	24.769	27.587	30.191	33.409	35.718	40.790
18	25.989	28.869	31.526	34.805	37.156	42.312
19	27.204	30.144	32.852	36.191	38.582	43.820
20	28.412	31.410	34.170	37.566	39.997	45.315
21	29.615	32.671	35.479	38.932	41.401	46.797
22	30.813	33.924	36.781	40.289	42.796	48.268
23	32.007	35.172	38.076	41.638	44.181	49.728
24	33.196	36.415	39.364	42.980	45.559	51.179
25	34.382	37.652	40.646	44.314	46.928	52.620
26	35.563	38.885	41.923	45.642	48.290	54.052
27	36.741	40.113	43.195	46.963	49.645	55.476
28	37.916	41.337	44.461	48.278	50.993	56.892
29	39.087	42.557	45.722	49.588	52.336	58.301
30	40.256	43.773	46.979	50.892	53.672	59.703
40	51.805	55.758	59.342	63.691	66.766	73.402
60	74.397	79.082	83.298	88.379	91.952	99.607
80	96.578	101.879	106.629	112.329	116.321	124.839
100	118.498	124.342	129.561	135.807	140.169	149.449

Appendix C: The Student's *t*-Distribution

For a particular number of degrees of freedom, the table contains t-values $\left(t_{1-\alpha;n}\right)$ for the probability $p = (1-\alpha)$ in the lower tail of the t distribution If $n > 30$, the t-distribution will approximate the normal distribution.

Example $t_{1-0.1;30} = t_{90\%;30} = 1.310$.



$t^n_{1-\alpha}$ for $p = (1-\alpha)$ in the lower tail of the t-distribution								
$n =$	0.8	0.9	0.95	0.975	0.99	0.995	0.999	0.9995
1	1.376	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	1.061	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.978	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.941	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.920	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.906	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.896	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.889	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.883	1.383	1.833	2.262	2.821	3.250	4.297	4.781

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-031-78070-7

$t^n_{1-\alpha}$ for $p = (1-\alpha)$ in the lower tail of the *t*-distribution								
$n =$	0.8	0.9	0.95	0.975	0.99	0.995	0.999	0.9995
10	0.879	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.876	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.873	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.870	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.868	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.866	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.865	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.863	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.862	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.861	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.860	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.859	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.858	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.858	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.857	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.856	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.856	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.855	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.855	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.854	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.854	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.851	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.848	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	0.845	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	0.842	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Appendix D: Critical Values for the Wilcoxon Signed-Rank Test

One-tailed test ($\alpha =$)		5%	2.5%	1%	0.5%			5%	2.5%	1%	0.5%
Two-tailed test ($\alpha =$)		10%	5%	2%	1%			10%	5%	2%	1%
$n =$	1	—	—	—	—	$n =$	26	110	98	84	75
	2	—	—	—	—		27	119	107	92	83
	3	—	—	—	—		28	130	116	101	91
	4	—	—	—	—		29	140	126	110	100
	5	0	0	—	—		30	151	137	120	109
	6	2	0	—	—		31	163	147	130	118
	7	3	2	0	—		32	175	159	140	128
	8	5	3	1	0		33	187	170	151	138
	9	8	5	3	1		34	200	182	162	148
	10	10	8	5	3		35	213	195	173	159
	11	13	10	7	5		36	227	208	185	171
	12	17	13	9	7		37	241	221	198	182
	13	21	17	12	9		38	256	235	211	194
	14	25	21	15	12		39	271	249	224	207
	15	30	25	19	15		40	286	264	238	220
	16	35	29	23	19		41	302	279	252	233
	17	41	34	27	23		42	319	294	266	247
	18	47	40	32	27		43	336	310	28	261
	19	53	46	37	32		44	353	327	296	276
	20	60	52	43	37		45	371	343	312	291
	21	67	58	49	42		46	389	361	328	307
	22	75	65	55	48		47	407	378	345	322

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-031-78070-7

One-tailed test ($\alpha =$)		5%	2.5%	1%	0.5%			5%	2.5%	1%	0.5%
Two-tailed test ($\alpha =$)		10%	5%	2%	1%			10%	5%	2%	1%
	23	83	73	62	54		48	426	396	362	339
	24	91	81	69	61		49	446	415	379	355
	25	100	89	76	68		50	466	434	397	373

Appendix E: Getting Started with SPSS

SPSS is organized with a core system that provides basic data management and analysis tools, and users can enhance its functionality by adding specific modules or extensions (so-called add-ins) that cater to their analytical needs.[1] These add-ins can be purchased separately and their combination with the core system makes SPSS a versatile tool for a wide range of basic and advanced statistical analysis tasks. Key components of SPSS are:

- Data Editor (see Fig. E.1): This is where you put in and manage your data. It looks like a spreadsheet, with rows representing cases (observations) and columns representing variables. You can enter data manually or import it from various file formats.
- Variable View (see Fig. E.1): Within the Data Editor, you can toggle between Data View and Variable View. Data View displays the actual data values, while Variable View allows you to define variable properties like labels, data types, missing values, and measurement levels. SPSS data files have the extension **.sav*.
- Menu Bar (see Fig. E.2): The menu bar at the top of the SPSS window provides access to various commands and functions, organized into submenus such as:
  - "Data" and "Transform": SPSS provides a wide range of data preparation and transformation tools, including functions for recoding variables, computing new variables, filtering cases, and handling missing data.
  - "Analyze": It provides access to numerous statistical procedures and tests, including multivariate statistical methods. Advanced procedures are often available through additional modules.
  - "Graphs": It allows you to create various types of graphs and charts to visualize your data, such as histograms, scatterplots, bar charts, and more.

[1] For further information, see https://www.ibm.com/docs/en/spss-statistics.

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-031-78070-7

- Output Viewer: The output viewer is where the results of your analyses are displayed after an analysis command has been issued. It includes tables, charts, and statistical summaries. Users can save and print these outputs for reporting. SPSS output files have the extension **.spv*.
- Syntax Editor: The Syntax Editor is where users can write, edit, and run syntax commands to perform advanced analyses or automate tasks. SPSS syntax programming refers to the practice of using a specific scripting language within the SPSS software to automate, customize, and document data analysis tasks. Instead of relying solely on the graphical user interface ("point-and-click"), users can write scripts in the SPSS syntax language to perform sequences of commands to be executed "automatically" one after another. This can save time and reduce the risk of errors, especially when working with large datasets or when performing the same operations multiple times. At the same time, syntax scripts document the performed data analysis steps and facilitate their comprehension and reproduction for third parties. This makes analysis easily reproducible and is crucial for research and data analysis transparency. Last but not least, keeping track of changes and revisions is easier with syntax files. Different versions of your analysis can be stored, annotated with comments, and be compared over time. SPSS syntax files have the extension **.sps*.

Undoubtedly, for beginners who are new to data analysis or lack programming skills, the point-and-click interface provides an easy entry point without the need to learn a new syntax language first. Beginners can concentrate on the interpretation of statistical analyses. However, it is important to note that over time many users, especially as they progress to more advanced data analysis tasks, often transition to learning and using syntax programming because it provides more control, automation capabilities, and flexibility. In this sense, the so-called "Paste" button, typically found in various SPSS dialogue boxes, plays a valuable role in facilitating syntax programming by allowing users to quickly generate syntax commands based on the use of the point-and-click interface: The "Paste" button allows users, instead of clicking the OK button to execute the command directly, to copy the associated SPSS syntax command straight to the syntax file.

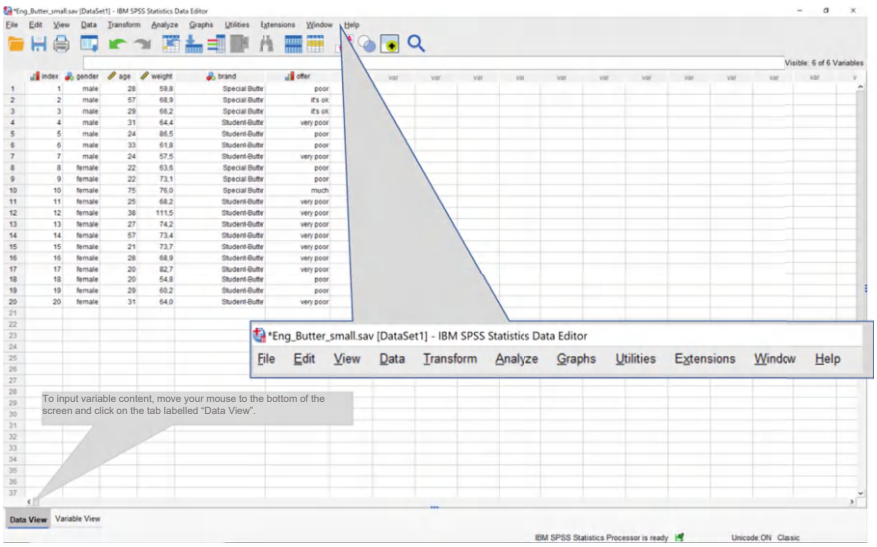


Fig. E.1 SPSS data viewer and SPSS menu bar

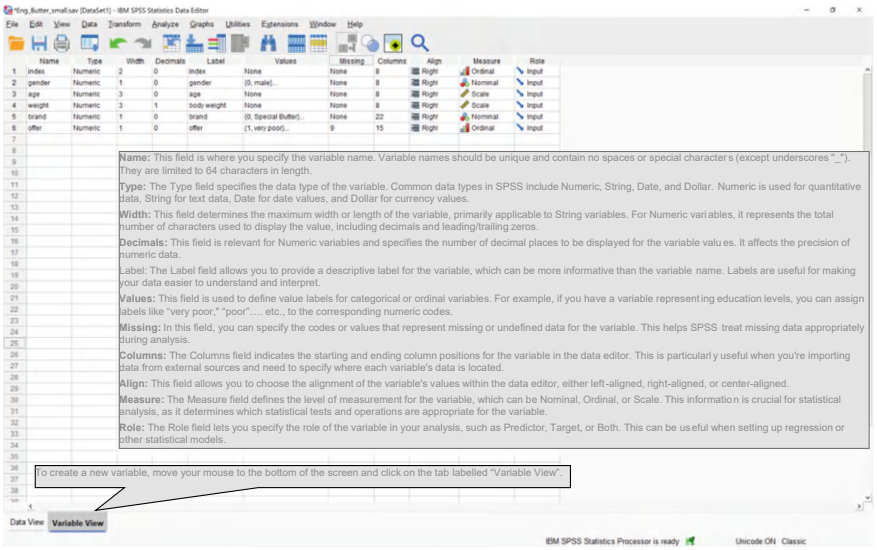


Fig. E.2 SPSS variable view

Appendix F: Getting Started with Stata

Stata provides a wide range of data management and analysis tools. Users can enhance its functionality by adding specific modules or extensions. These so-called "Ado files" are used to add new commands, functions, or features to Stata that are not available by default. They allow users to create custom routines and procedures to perform specialized tasks or automate repetitive data analysis processes.[2] Usually, these ado files can be downloaded free of charge. This makes Stata one of the most cutting-edge tools for statistical data analyses.

The Stata window (see Fig. F.1) typically consists of several key components:

- Menu Bar: The menu bar is located at the top of the Stata window and contains various menus such as File, Edit, Data, Graphics, Statistics, and more. Each menu provides access to a wide range of commands and options for performing tasks within Stata.
- Command Line: The command line is where users can enter Stata commands directly. Users can type commands here and press Enter to execute them. The command line also displays the results of executed commands and any error messages that may occur.
- Review and Results Panel: This panel displays the output and results of executed commands. It includes tables, graphs, summary statistics, and other information generated by Stata.
- Log: Keeps a record of all executed commands and their results. The log is useful for documenting your analysis and troubleshooting issues.
- Variable Panel: This panel provides a list of variables in the currently loaded dataset. Users can view variable names, labels, data types, and properties. It is also where users can define variable properties and create or modify variables.
- Properties Panel: It provides information about the current dataset, such as the number of observations and variables.

[2] For further information, see https://www.stata.com/manuals13/u17.pdf.

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-031-78070-7

In addition, other important windows can be opened by clicking on the Tool Bar (see Fig. F.2):

- Data Editor: The data editor is used for viewing and editing the dataset. It opens when you load a dataset or create a new one. Users can enter and edit data values, recode variables, and perform basic data manipulation tasks within the data editor.
- Do-File Editor: Stata provides a built-in do-file editor that allows users to write and edit script files containing Stata commands. Do-files are often used for automating repetitive tasks and documenting analysis procedures. Stata syntax files have the extension **.do*.

Instead of relying solely on the graphical user interface ("point-and-click"), users can write scripts in the Stata syntax language to perform sequences of commands to be executed "automatically" one after another. Undoubtedly, for beginners who are new to data analysis or lack programming skills, the point-and-click interface provides an easy entry point without the need to learn a new syntax language first. Beginners can concentrate only on the interpretation of statistical analyses. However, it is important to note that over time many users, especially as they progress to more advanced data analysis tasks, often transition to learning and using syntax programming because it provides more control, automation capabilities, and flexibility. In this sense, "copy and paste" from the Log panel or the command line into a do-file plays a valuable role in facilitating syntax programming by allowing users to quickly generate syntax commands based on the use of the point-and-click interface. With Stata syntax, users can automate repetitive or complex data manipulation and analysis tasks. This can save time and reduce the risk of errors, especially when working with large datasets or when performing the same operations multiple times. At the same time, syntax scripts document the performed data analysis steps and facilitate their comprehension and reproduction for third parties. This makes analysis easily reproducible and is crucial for research and data analysis transparency. Last but not least, keeping track of changes and revisions is easier with syntax files. Different versions of your analysis can be stored, annotated with comments, and be compared over time.

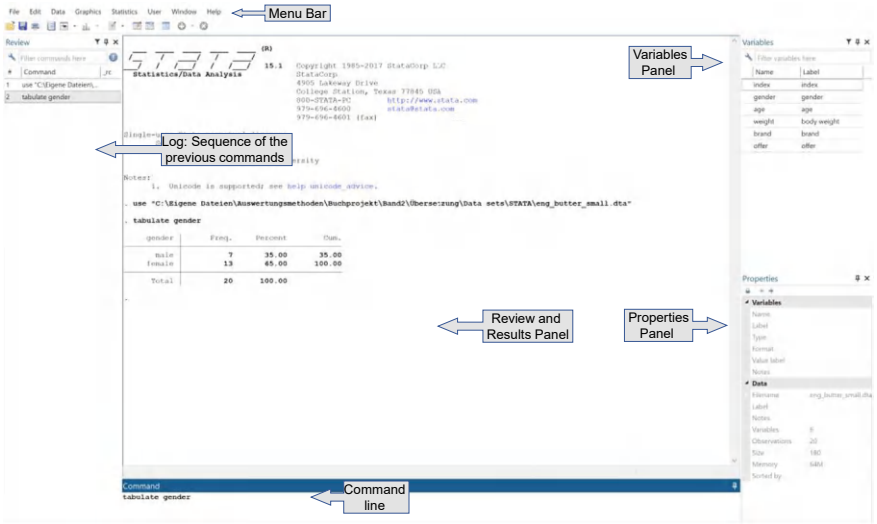


Fig. F.1 The Stata window

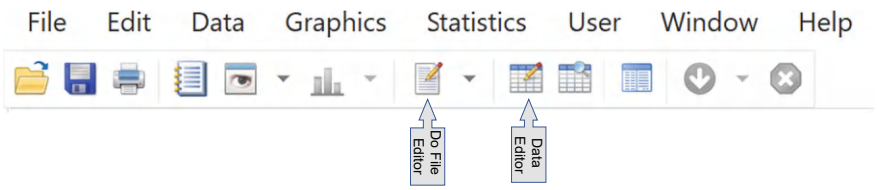


Fig. F.2 The Stata tool bar

Appendix G: Getting Started with R and the R Commander

The R software can be downloaded from the official R Project website at https://cran.r-project.org/. On the website's main page, click on "Mirrors" to see a list of CRAN (Comprehensive R Archive Network) mirrors worldwide. Select a mirror location closest to your geographical location. On the mirror's webpage, you will typically find links to download R for different operating systems (e.g., Windows, macOS, Linux). Once the download is complete, run the installer and follow the installation prompts. After the installation is completed, you should have the R software installed (see Fig. G.1) on your computer and the graphical user interface R Commander can be installed by typing in "install.packages("Rcmdr")" into the R console. As soon as the package is installed, just enter the "library(Rcmdr)" to open the R Commander.

Sometimes it may happen that one or more R packages required during the installation process of the R Commander are not installed so that R Commander cannot be executed. In this case, an informative error message is usually issued about the missing package(s). There are two likely reasons for this problem: (1) the package(s) in question may be temporarily missing from the chosen CRAN mirror or (2) the package(s) may be permanently unavailable for an older version of R. In these cases, install the latest version of R and choose another CRAN mirror.

The language of the R Commander can be changed from the system language set in Windows to English by using the command "Sys.setenv(LANG = 'en')". Please find more information about installation and use of the R Commander on the CRAN website: https://cran.r-project.org/doc/contrib/Karp-Rcommander-intro.pdf.

Unfortunately, R Commander does not open automatically whenever R is launched. What is needed instead is to type the command "library(Rcmdr)" into the R console each time to launch R Commander.

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-031-78070-7

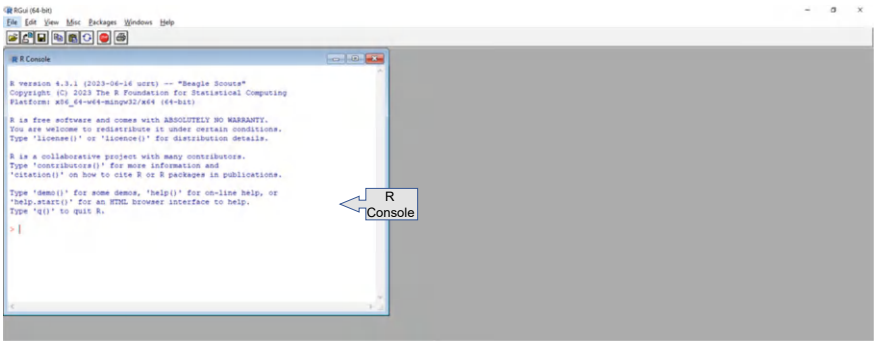


Fig. G.1 The R console

The R Commander consists of the following components (see Fig. G.2):

- The Drop Menu Bar: At the top of the R Commander window, you will find a menu bar with various menu options. These menus provide access to various data analysis functions and commands, allowing you to perform tasks like data import, data manipulation, statistical tests, and creating plots. The Data Menu for example provides options for importing, exporting, and managing datasets. You can also use this menu to perform basic data pre-processing tasks. The Statistics Menu offers a wide range of statistical analyses and tests, including descriptive statistics, hypothesis tests, regression analysis, and more.
- The Toolbar contains additional utilities and options for customizing the R Commander interface and settings.
- The Script Window allows to write, edit, and execute R commands and scripts. You can either type R code directly into this window or use the menus to perform actions, which are translated into R syntax and displayed in the Script Window.
- The Output Window shows any textual output generated by R, such as summary statistics, regression results, or data manipulation outputs.
- The Message Window displays messages, status updates, or information about the current operation or task being performed.

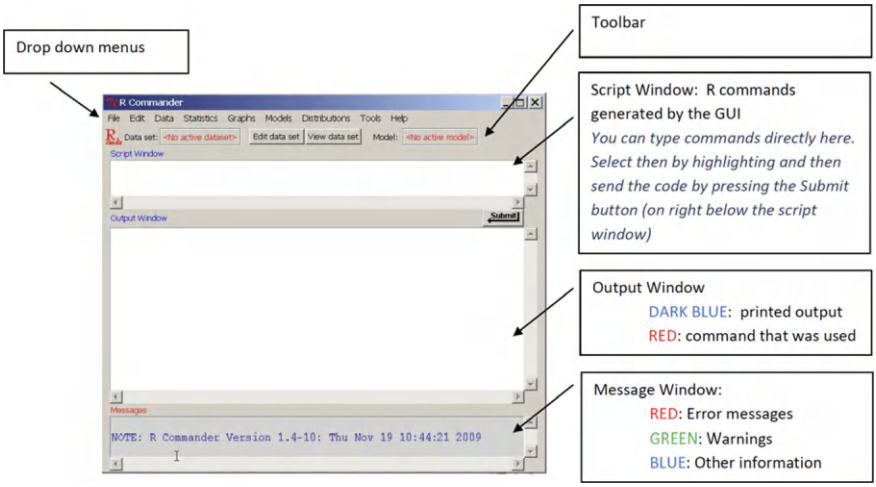


Fig. G.2 The R commander window. *Source* https://cran.r-project.org/doc/contrib/Karp-Rcommander-intro.pdf, p. 5

Index

A
Absolute deviation, 58
Absolute scales, 24
Accuracy, 439
Addition rule
  for disjoint events, 167
  for nondisjoint events, 169
Adjusted R-square, 404, 397–398
Agglomeration schedule, 478, 481, 490
Agglomerative methods, 472
Alpha error, 280
Alternative hypothesis, 279–284
Analysis of covariance (ANCOVA), 335
Analysis of variance (ANOVA), 327–341
  one-way ANOVA, 327–331
  two-way ANOVA, 331–335
  with Excel, 339
  with R, 338
  with SPSS, 338
  with Stata, 338
Anti-image covariance matrix (AIC), 503
Arithmetic, *see mean*
Autocorrelation, 406
Auxiliary regression, 409–411, 441
Average linkage, *see Linkage methods*
Average, *see mean*

B
Bar chart, 35
Bartlett test, 330, 503–504
Base period, 454, 457, 458
Bayes' theorem, 172–173
Bessel's corrected variance, 59
Beta coefficient of a regression, 386
Beta error, 280
Binary regression, 421
Binomial distribution, 191–195
  with Excel, 194
  with Stata, 195
Birthday paradox, 177
Biserial rank correlation, 122
Bivariate association, 79–144
  strength of, 91, 107
Bivariate centroid, 106, 387, 400
Boxplot, 54
Bravais-Pearson, 103

C
Cardinal scale, 23–25, 26, 28, 37
Causality, 383
Central limit theorem, 248, 331, 354
Central tendency, 283, 303, 342
  measures of, 40–54
Centroid linkage, *see Linkage methods*
Chi-square, 82–86
Chi-squared distribution, 218–220
  table of the, 538
  with Excel, 219
  with Stata, 219
Chi-square test of independence, 348–354
  with Excel, 353
  with R, 353
  with SPSS, 351
  with Stata, 351
Classification table, 439
  accuracy, 439
  precision, 440
  sensitivity, 439
  specificity, 440
Cluster analysis, 471–488
  with R, 489
  with SPSS, 488
  with Stata, 489
Coefficient of correlation, 103–107

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

T. Cleff, *Applied Statistics and Multivariate Data Analysis for Business and Economics*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-031-78070-7

Coefficient of determination (adjusted), 393
Coefficient of determination, *see R-square*
Coefficient of variation, 59
Combination
  without repetition, 165
  with repetition, 165
Combinatorics, 163–167
Communalities, 503–505, 515
Complete linkage, *see Linkage methods*
Concentration
  concentration ratio, 64
  Gini coefficient, 67
  Herfindahl index, 65
  Lorenzcurve, 65
  measures of, 64–67
Concentration ratio, *see Concentration*
Conditional frequency, 81
Conditional probability, 169–170
Confidence intervals
  for proportions, 258–260
  for the mean, 250–255
  for variances, 262–263
  with Excel, 263
  with R, 270
  with SPSS, 265
  with Stata, 266
Confidence level, 250, 252, 256, 266, 280, 350
Confusion table, 439
Consistent estimation, 250
Contingency coefficient, 89–90
Contingency table, 80, 349
Correlation
  Pearson, 103–107
  Spearman, *see Spearman's rank correlation*
  spurious, 127
  with Excel, 126
  with R, 123
  with SPSS, 124
  with Stata, 125
Correlation matrix, 502–504
Covariable, 335
Covariance, 104, 106, 125
Covariate, 335
Cox-Snell $R^{2}2$, 436
Cramer's V, 135
Cross-sectional analyses, 150, 453
Crosstab, 80–81

D
Dendrogram, 481
Density, 39, 204, 208, 209
Density function, 204, 206, 218
Dependent t-test, *see t-test for dependent samples*
Descriptive statistics, 4
Dichotomous variable, 80
Dispersion parameters, 56–60
Distance matrix, 476
Distribution function, 35, 37–40
Ducan, 335

E
Eigenvalue, 504–506
Empirical standard deviation, 58
Empirical variance, 58
Equidistance, 24, 44, 109, 114
Error probability, 6
Error sum of squares (ESS), 393
Error term, 405
Euclidian distance, 474, 475
Event, 155–162
  combined, 156
  complementary, 157
  elementary, 156
  intersection of events, 156
Excel
  analysis of variance (ANOVA) with, 339
  binomial distribution with, 194
  chi-squared distribution with, 219
  chi-square test of independence with, 353
  confidence intervals with, 263
  correlation with, 126
  dependent t-test with, 302
  F-distribution with, 225
  hypergeometric distribution with, 199
  independent t-test with, 319
  nominal association with, 96–100
  normal distribution with, 216
  one-sample t-test with, 292
  paired t-test with, 302
  partial correlation with, 133
  Poisson distribution with, 202
  regression with, 396
  t- distribution with, 223
  t-test for independent samples with, 319
  univariate parameters with, 71
  Wilcoxon signed-rank test with, 311
Excel Software, 17
Excess, 63
Expected counts, 82–84, 92, 135, 353
Expected frequency, 83, 92, 350
Expected relative frequency, 83
Expected value
  continuous distribution, 205
  discrete distribution, 190

Extreme values, 54

F
Factor analysis, 501–509
  with R, 510
  with SPSS, 510
  with Stata, 510
Factor matrix, 507
Factor score coefficient matrix, 508
F-distribution, 225–227
  with Excel, 225
  with Stata, 225
Fisher index, *see Index*
Fourth central moment, 63
Frequency distribution, 35, 72, 190, 208
Frequency table, 34, 80
F-test, 314
Full survey, 3, 7

G
Gauß test, *see One-sample Z-test*
Geometric mean, 46
Gini coefficient, *see Concentration*
Goodness of fit, 397

H
Harmonic mean, 47
Herfindahl index, *see Concentration*
Heteroscedastic, 407
Heteroskedasticity, 441
H_1 hypothesis, 279–284
Histogram, 37–40
Homoscedastic, 407
Hosmer-Lemeshow test, 437
H test, *see Kruskal-Wallis H test*
Hypergeometric distribution, 195–200
  with Excel, 199
  with Stata, 199
Hypothesis
  about a relationship, 282
  falsification, 279
  verification, 279

I
Inability error, 152
Inclusion–Exclusion Principle
  for disjoint events, 167
  for nondisjoint events, 169
Independent t-test, *see t-test for independent samples*
Index, 421–468
  Fisher quantity index, 461
  Laspeyres price index, 457
  Paasche price index, 459
  price index, 454
  sales index, 462
  value index, 462
Inductive statistics, 7, 11
Interaction effect, 332–335
Interquartile range, 54
Interval estimation, 250
Interval scales, 24
Item non-response, 256

K
Kaiser-criterion, 506
Kaiser-Meyer-Olkin measure, 503, 510
Kendall's Tau, 114–119
KMO measure, 503, 510
Kolmogorov-Smirnov test, 355
Kruskal-Wallis H test, 341–348
  with R, 347
  with SPSS, 346
  with Stata, 347
Kurtosis, 63

L
Laspeyres price index, *see Index*
Law of total probability, 171
Left-skewed, 60–62
Leptokurtic distribution, 63
Level of measurement, 79
Levene's test, 314
Likelihood ratio test, 433
Linear relationship, 103–107
Linkage methods, 479–481
  average linkage, 479
  centroid linkage, 479
  complete linkage, 479
  single linkage, 479
  Ward's method, 480
Logistic regression, 421
  classification table, 439
  coefficient, 430, 435
  confusion table, 439
  Cox-Snell $R^{2}{}^{2}$, 435
  estimation, 429
  heteroskedasticity, 441
  Hosmer-Lemeshow test, 437
  likelihood ratio test, 433
  McFadden R^2, 435
  multicollinearity, 441

Nagelkerke R^2, 436
Pseudo-R^2, 435
p-value, 435
Wald test, 434
with R, 446
with SPSS, 444
with Stata, 444
Logit, 421, 425
Log-odds, 425, 430
Longitudinal study, 453
Lorenzcurve, *see Concentration*

M
Mann-Whitney U test, 320–326
with R, 326
with SPSS, 324
with Stata, 324
MANOVA, 337
Marginal frequencies, 80, 84
Maximum likelihood, 422, 427
McFadden R^2, 435
Mean
arithmetic, 41
geometric, 46
harmonic, 47
trimmed, 42
Mean rank, 110, 113, 127, 343, 362
Measurement error, 153
Measurement theorie, 147
Measure of sample adequancy, 503
Median, 50
Median absolute deviation, 57
Mesokurtic distribution, 63
Metric variable, 100, 119, 122
Missing values, 27
Modal value, *see Mode*
Mode, 41, 63
Model
symbolic, 11
verbal, 11
Monotonic relationsship, 107, 109–113, 114–119
Monty Hall problem, 173–176
Multicollinearity, 441
Multicollinearity, *see Regression*
Multiple regression, *see Regression*

N
Nagelkerke R^2, 436
Nominal association with
Excel, 96–99
R, 96
SPSS, 91–95
Stata, 95–96
Nominally scaled variables, 80
Nonlinear regression, *see* Regression
Non-opinion, 27, 153
Nonparametric test, 283, 303, 320, 346
Nonresponse error, 152
Nonsampling error, 151–154
No-opinion, 27
Normal distribution, 208–217
tests for, 354–358
with Excel, 216
with Stata, 217
Null hypothesis, 219–284

O
Odds, 425
Odds ratio, 432
One-sample t-test, 287–291
with Excel, 292
with R, 291
with SPSS, 291
with Stata, 291
One-sample Z-test, 285–287
One-way analysis of variance (ANOVA), *see Analysis of variance (ANOVA)*
Ordinal scaled variables, 107
Ordinary least squares method (OLS), 388
Outliers, 29, 42, 52, 54, 56, 59, 63, 107, 247, 401, 481

P
Paasche price index, *see Index*
Paired t-test, *see t-test for dependent samples*
Parametric test, 283
Partial correlation, 131
with Excel, 133
with R, 131
with SPSS, 131
with Stata, 133
Partial sample, 4, 6
Pearson correlation, *see correlation*
Percentile, 52
Permutation
without repetition, 163
with repetition, 164
Phi coefficient, 86–89, 94
Pie chart, 35
Planning the sample size
for the mean, 256–258
for the proportion, 261
Platykurtic distribution, 63

Point-biserial correlation, 104, 119–121
Point estimation, 244–250
Poisson distribution, 200–203
  with Excel, 202
  with Stata, 202
Population, 4, 6
Population variance, 58
Post-hoc methods, 334, 335
Power of a test, 280
Precision, 400
Price index, *see Index*
Pricipal component analysis, 504
Pricipal factor analysis, 504
Principle of indifference, 158
Probability, 157–161
  empirical, 159
  Laplace, 158
  statistical, 159
  subjective, 161
Probability function, 191
Probability tree, 161–163
Prognosis model, 12
Pseudo-R^2, 435
P-value, 282, 290

Q
Quantile, 52
Quartile, 52, 57

R
R
  analysis of variance (ANOVA) with, 338
  chi-square test of independence with, 353
  cluster analysis with, 489
  confidence intervals with, 270
  correlation with, 123
  dependent t-test with, 301
  factor analysis with, 510
  independent t-test with, 317
  Kruskal-Wallis H test with, 347
  logistic regression with, 446
  Mann-Whitney U test with, 326
  nominal association with, 96
  one-sample t-test with, 291
  paired t-test with, 301
  partial correlation with, 131
  regression with, 395
  tests for normal distribution with, 358
  t-test for independent samples with, 317
  univariate parameters with, 67
  Wilcoxon signed-rank test with, 311
Random variable, 189
Range, 57
Rank correlation, *see Spearman's rank correlation*
Rank ties, 113, 116, 312, 322, 343
Ratio scales, 24
R Commander, 19, 545
Regression
  autocorrelation, 406
  beta coefficient, 386
  coefficient of the, 385–389
  diagnostics, 404–411
  heteroscedastic, 406
  homoscedastic, 406
  multicollinearity, 407
  multiple, 389–393
  nonlinear, 401–404
  with Excel, 396
  with R, 395
  with SPSS, 394
  with Stata, 394
Regression sum of squares (TSS), 392
Relative frequencies, 34, 35, 159
Reliability, 148
Reporting period, 454, 459
Representative sample, 4
Reproduced correlation matrix, 505
Reproductivity
  binomial distribution, 194
  chi-squared distribution, 219
  normal distribution, 210
  Poisson distribution, 201
Response error, 152
Retail questionnaire, 22
Rho, *see Spearman's rank correlation*
Right-skewed, 60–62
Robust, 107, 316, 318, 323, 331, 410, 441, 480
Robustness of parameters, 63
Rotated factor matrix, 506
R Software, 17, 545
  starting with R, 545
R-square, 393, 397–398

S
Sales index, *see Index*
Sample space, 156
Sampling
  cluster, 150
  convenience, 150
  jugemental, 150
  nonrandom, 150
  quota, 150
  random, 150
  sequential, 150

simple random, 150
stratified, 150
Sampling error, 149–151
Sampling methods, 150–151
Scatterplot, 101–103, 137, 384, 385, 400, 402
Scheffé, 334
Scree plot, 481–482, 506
S-curve, 429
Sensitivity, 440
Shapiro-Francia test, 355
Shapiro-Wilk test, 355
Significance level, 262, 280, 285
Single linkage, *see Linkage methods*
Skewness, 60–62
Spearman's rank correlation, 109–113
Specificity, 440
Sphericity test, 503–504
SPSS
analysis of variance (ANOVA) with, 338
chi-square test of independence with, 351
cluster Analysis with, 488
confidence intervals with, 265
correlation with, 124
dependent t-test with, 299
factor Analysis with, 510
independent t-test with, 315
Kruskal-Wallis H test with, 346
logistic regression with, 444
Mann-Whitney U test with, 324
nominal association with, 91–96
Normal distribution with, 216
one-sample t-test with, 291
paired t-test with, 299
partial correlation with, 131
regression with, 394
starting with SPSS, 537
tests for normal distribution with, 355
t-test for dependent samples with, 299
t-test for independent samples with, 315
univariate parameters with, 69
Wilcoxon signed-rank test with, 308
SPSS software, 17, 537
Spurious correlation, 127
Standard deviation, 121
Standard error, 62, 246, 248, 249, 252, 286, 289
Standardization, 419, 475, 509
Standard normal distribution
table of the, 527
Starting with
R, 545
SPSS, 537
Stata, 541
Stata
analysis of variance (ANOVA) with, 338
binomial distribution with, 195
chi-squared distribution with, 219
chi-square test of independence with, 351
cluster analysis with, 489
confidence intervals with, 266
correlation with, 125
dependent t-test with, 299
factor analysis with, 510
F-distribution with, 225
hypergeometric distribution with, 199
independent t-test with, 316
Kruskal-Wallis H test with, 347
logistic regression with, 444
Mann-Whitney U test with, 324
nominal association with, 95–96
one-sample t-test with, 291
paired t-test with, 299
partial correlation with, 133
Poisson distribution with, 202
regression with, 394
t-distribution with, 224
tests for normal distribution with, 357
t-test for independent samples with, 316
univariate parameters with, 69
Wilcoxon signed-rank test with, 310
Stata software, 17, 541
starting with Stata, 541
Statistical unit, 22–25
Student distribution, *see t-distribution*
Survey, 3, 79
Systematic bias, 27
Systematic error, 27, 148, 402

T

T-distribution, 221–223
table of the, 538
with Excel, 223
with Stata, 224
Test
difference between dependent and independent, 282
difference between one-sided and two-sided, 285
nonparametric, 283, 303, 320, 346
parametric, 283
Tests for normal distribution, 354–358
with R, 358
with SPSS, 355
with Stata, 357
Third central moment, 61
Tolerance, 409–411, 441
Total sum of squares (TSS), 392, 393, 411

T-test for dependent samples, 294–303
with Excel, 302
with R, 301
with SPSS, 299
with Stata, 300
T-test for independent samples, 312–320
with Excel, 319
with R, 317
with SPSS, 315
with Stata, 316
Tukey HSD test, 335
Tukey LSD test, 335
Two-way analysis of variance (ANOVA), *see Analysis of variance (ANOVA)*
Type I error, 280
Type II error, 280

U
Unbiased sample standard deviation, 59
Unbiased sample variance, 59
Uniform distribution
continuous, 206–208
discrete, 190
Univariate parameters
with R, 67
with SPSS, 69
with Stata, 69
Unrotated factor matrix, 506
U test, *see Mann-Whitney U test*

V
Validity, 148
Value index, *see Index*
Variance Inflation Factor (VIF), 409–411, 441
Variation
without repetition, 166
with repetition, 165
Varimax rotation, 507–508, 510

W
Wald test, 434
Ward's method, *see Linkage methods*
Whiskers, 54–56
Wilcoxon rank-sum test, *see Mann-Whitney U test*
Wilcoxon signed-rank test, 303–312
table of the, 538
with Excel, 311
with R, 311
with SPSS, 308
with Stata, 310

Z
Z-test, *see One-sample Z-test*
Z-transformation, 251, 475, 488, 211–266