Thomas Cleff

# Exploratory Data Analysis in Business and Economics

## An Introduction Using SPSS, Stata, and Excel

Springer

# Exploratory Data Analysis in Business and Economics

Thomas Cleff

# Exploratory Data Analysis in Business and Economics

## An Introduction Using SPSS, Stata, and Excel

Springer

Thomas Cleff
Pforzheim University
Pforzheim, Germany

Chapters 1–6 translated from the German original, Cleff, T. (2011). Deskriptive Statistik und moderne Datenanalyse: Eine computergestützte Einführung mit Excel, PASW (SPSS) und Stata. 2. überarb. u. erw. Auflage 2011 © Gabler Verlag, Springer Fachmedien Wiesbaden GmbH, 2011

ISBN 978-3-319-01516-3　　ISBN 978-3-319-01517-0 (eBook)
DOI 10.1007/978-3-319-01517-0
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013951433

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

# Preface

This textbook, *Exploratory Data Analysis in Business and Economics: An Introduction Using SPSS, Stata, and Excel*, aims to familiarize students of economics and business as well as practitioners in firms with the basic principles, techniques, and applications of descriptive statistics and data analysis. Drawing on practical examples from business settings, it demonstrates the basic descriptive methods of univariate and bivariate analyses. The textbook covers a range of subject matter, from data collection and scaling to the presentation and univariate analysis of quantitative data, and also includes analytic procedures for assessing bivariate relationships. In this way, it addresses all of the topics typically covered in a university course on descriptive statistics.

In writing this book, I have consistently endeavoured to provide readers with an understanding of the thinking processes underlying descriptive statistics. I believe this approach will be particularly valuable to those who might otherwise have difficulty with the formal method of presentation used by many textbooks. 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. Indeed, 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 computers now makes it possible to teach statistics in new ways. Indeed, students now have access to a range of powerful computer applications, from Excel to various statistics programmes. Accordingly, this textbook does not confine itself to presenting descriptive statistics, but also addresses the use of programmes such as Excel, SPSS, and Stata. To aid the learning process, datasets have been made available at springer.com, 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. First and foremost, I would like to thank Lucais Sewell (lucais.sewell@gmail.com) for translating this work from German into English. It is no small feat to render an academic text such as this into precise but readable English. Well-deserved gratitude for their critical review of the manuscript and valuable suggestions goes to Birgit Aschhoff, Christoph Grimpe, Bernd Kuppinger, Bettina Müller, Bettina Peters, Wolfgang Schäfer, Katja Specht, Fritz Wegner, and Kirsten Wüst, as well as many other unnamed individuals. Any errors or shortcomings that remain are entirely my own. I would also like to express my thanks to Alice Blanck at Springer Science + Business Media for her assistance with this project. 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 Thomas Cleff
March 2013

# Contents

# List of Figures

# List of Tables

# List of Formulas

*Measures of Central Tendency:*

Mean (from raw data): $\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$

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: $\bar{x}_{geom} = \sqrt[n]{(x_1 \cdot x_2) \cdot \cdots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n}(1 + x_i)}$

Geometric Mean for Change Rates:

$\bar{p}_{geom} = \sqrt[n]{(1 + p_1) \cdot (1 + p_2) \cdot \cdots \cdot (1 + p_n)} - 1 = \sqrt[n]{\prod_{i=1}^{n}(1 + p_i)} - 1$

Harmonic Mean (unweighted) for k observations: $\bar{x}_{harm} = \frac{k}{\sum_{i=1}^{k} \frac{1}{x_i}}$

Harmonic Mean (weighted) for k observations: $\bar{x}_{harm} = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$

Median (from classed data): $\tilde{x} = x_{0.5} = x_{i-1}^{UP} + \frac{0.5 - F\left(x_{i-1}^{UP}\right)}{f(x_i)}\left(x_i^{UP} - x_i^{LOW}\right)$

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)·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: $(1-f)\cdot x_{(i)}+f\cdot x_{(i+1)}$

Quantile (from classed data): $x_p=x^*_{i-1}+\frac{p-F\left(x^*_{i-1}\right)}{f_i}\Delta x_i$

*Dispersion Parameters:*

Interquartile Range: IQR= $x_{0.75}-x_{0.25}$

Mid-Quartile Range: MQR=0.5·$(x_{0.75}-x_{0.25})$

Range: Range=Max($x_i$)-Min($x_i$)

Median Absolute Deviation: MAD $=\frac{1}{n}\sum_{i=1}^{n}|x_i-\tilde{x}|$

Empirical Variance: $Var(x)_{emp}=S^2_{emp}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2$

Empirical Standard Deviation: $S_{emp}=\sqrt{Var(x)_{emp}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$

(Unbiased Sample) Variance: $Var(x)=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

(Unbiased Sample) Standard Deviation: $S=\sqrt{Var(x)}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$

Coefficient of Variation: $V=\frac{S}{|\bar{x}|},\bar{x}\neq 0$

*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: 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: GINI $=\frac{2\sum_{i=1}^{n}i\cdot f_i-(n+1)}{n}$

Normalized Gini Coefficient ($GINI_{norm.}$): Normalized by Multiplying Each of the Above Formulas by $\frac{n}{n-1}$

*Skewness and Kurtosis:*

Skewness (Yule & Pearson): Skew= $\frac{3\cdot(\bar{x}-\tilde{x})}{S}$

Skewness (Third Central Moment): Skew$= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^3}{S^3}$

Kurtosis: Kurt $= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^4}{S^4}$

*Bivariate Association:*

Chi-Square: $\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij}-n_{ij}^e)^2}{n_{ij}^e}$

Phi: $PHI = \sqrt{\frac{\chi^2}{n}}$

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: $\mathrm{cov}(x;y) = S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) = \frac{1}{n}\sum_{i=1}^{n}x_iy_i - \overline{xy}$

Bravais-Pearson Correlation: $r = \frac{S_{xy}}{S_xS_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2\right)}}$

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

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

Spearman's Rank Correlation (Short-Hand Version):

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)} mit\ d_i = (R(x_i) - R(y_i))$$

Spearman's Rank Correlation (Short-Hand Version With Rank Ties):

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

- T as the length of b tied ranks among x variables: $T = \frac{\sum_{i=1}^{b} \left(t_i^3 - t_i\right)}{12}$, 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: $U = \frac{\sum_{i=1}^{c} \left(u_i^3 - u_i\right)}{12}$, 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 $\tau_a$ (Without Rank Ties): $\tau_a = \frac{P - I}{n \cdot (n-1)/2}$

Kendall's $\tau_b$ (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 i$^{th}$ 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 i$^{th}$ of c groups of tied ranks for the y variables.

Biserial Rank Correlation (Without Rank Ties) $r_{bisR} = \frac{2}{n} \cdot \left(\overline{R(y_1)} - \overline{R(y_0)}\right)$

*Regression Analysis:*

Intercept of a Bivariate Regression Line: $\alpha = \overline{y} - \beta \cdot \overline{x}$

Slope of a Bivariate Regression Line:

$$\beta = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{\mathrm{cov}(x, y)}{S_x^2} = \frac{r \cdot S_y}{S_x} = \frac{n \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Coefficients of a Multiple Regression: $\beta = (X'X)^{-\ 1}X'y$

R$^2$/Coefficient of Determination:

$$R^2 = \frac{RSS}{TSS} = \frac{SS_{\widehat{Y}}}{SS_Y} = \frac{\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = 1 - \frac{ESS}{TSS} = 1 - \frac{SS_\epsilon}{SS_Y} = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

Adjusted R$^2$/Coefficient of Determination:

$$R^2_{adj} = R^2 - \frac{(1 - R^2)(k - 1)}{(n - k)} = 1 - (1 - R^2)\frac{n - 1}{n - k}$$

*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 (Sales Index): $W_{0,t} = \frac{\sum_{i=1}^{n} p_{i,t} \cdot q_{i,t}}{\sum_{i=1}^{n} p_{i,0} \cdot q_{i,0}} = Q^F_{0,t} \cdot P^F_{0,t} = Q^L_{0,t} \cdot P^P_{0,t} = Q^P_{0,t} \cdot P^L_{0,t}$

Deflating Time Series by Price Index: $L^{real}_t = \frac{L^{no\min al}_t}{P^L_{0,t}}$

Base Shift of an Index: $I^{new}_{\tau,t} = \frac{I^{old}_{0,t}}{I^{old}_{0,\tau}}$

Chaining an Index (Forward Extrapolation): $\tilde{I}_{0,t} = \begin{cases} I^1_{0,t} & \text{für t} \leq \tau \\ I^1_{0,\tau} \cdot I^2_{\tau,t} & \text{für t} > \tau \end{cases}$

Chaining an Index (Backward Extrapolation): $\tilde{I}_{0,t} = \begin{cases} \frac{I^1_{0,\tau}}{I^2_{\tau,t}} & \text{für t} < \tau \\ I^2_{\tau,t} & \text{für t} \geq \tau \end{cases}$

# 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 Disreali, 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.

We are thus caught in the middle of a seeming contradiction. Why do statistics in some contexts attract aspersion, yet in others emanate an aura of authority, a nearly mystical precision? If statistics are indeed the superlative of all lies – as claimed by Disreali – then why do individuals and firms rely on them to plan their activities? Swoboda (1971, p. 16) has identified two reasons for this ambivalence with regard to statistical procedures:

Chapter 1 Translated from the German original, Cleff, T. (2011). 1 Statistik und empirische Forschung. In *Deskriptive Statistik und moderne Datenanalyse* (pp. 1–14) © Gabler Verlag, Springer Fachmedien Wiesbaden GmbH, 2011.

T. Cleff, *Exploratory Data Analysis in Business and Economics*,
DOI 10.1007/978-3-319-01517-0_1, © Springer International Publishing Switzerland 2014

- 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 journalists or readers. Yet journalists 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. Thus, we find that statistics can be an aid to understanding phenomena, but they may also be based on the false application of statistical methods, whether intentional or unintentional.

Krämer (2005, p. 10) 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 Goethe, who famously said in a conversation with Eckermann, “That I know, the numbers instruct us” (*Das aber weiß ich, dass die Zahlen uns belehren*). Statistical models and methods are one of the most important tools in microeconomic analysis, 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 Two 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 6th 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.

All early statistical record keeping took the form of a *full survey* in the sense that an attempt was made to literally count every person, animal, and object. At the beginning of the 20th century, employment was a key area of interest; the effort to track unemployment was difficult, however, due to the large numbers involved. 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 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.

It was not until the beginning of the 20th century that the now common form of *inductive data analysis* 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

[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.

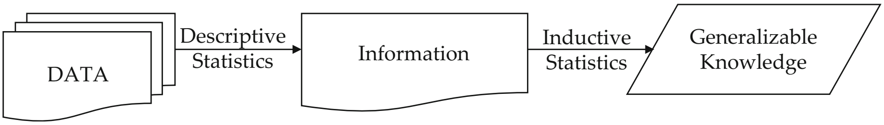


**Fig. 1.1** Data begets information, which in turn begets knowledge

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.

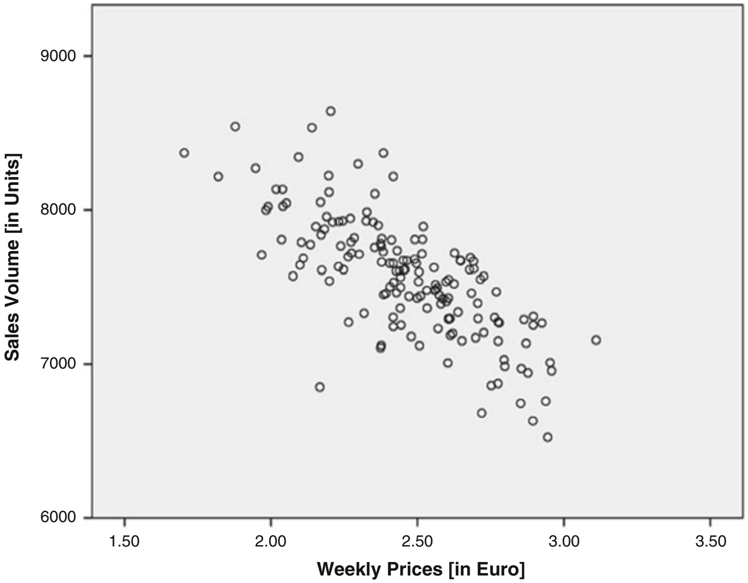
Descriptive and inductive statistics are a scientific discipline used in business, economics, the natural sciences, and the social sciences. It is a discipline that encompasses methods for the description and analysis of mass phenomena with the aid of numbers and data. The analytical goal is to draw 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 analyzed 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.

## 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 (Fig. 1.2). The researcher would first attempt to gain an understanding of the market by gathering individual pieces of information. He could, for example,



**Note**: The figure shows the average weekly prices and associated sales volumes over a 3 year period. Each point represents the amount of units sold at a certain price within a given week

**Fig. 1.2** Price and demand function for sensitive toothpaste

analyze weekly toothpaste prices and sales over the last 3 years. 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.

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. 1.3. 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 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 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.

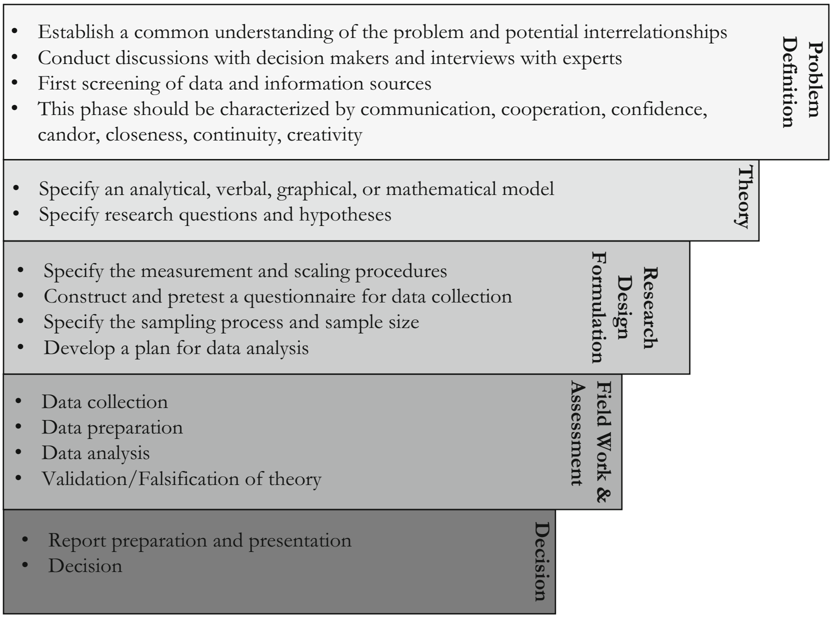

Establish a common understanding of the problem and potential interrelationships
Conduct discussions with decision makers and interviews with experts
First screening of data and information sources
This phase should be characterized by communication, cooperation, confidence, candor, closeness, continuity, creativity
Problem Definition
Specify an analytical, verbal, graphical, or mathematical model
Specify research questions and hypotheses
Theory
Specify the measurement and scaling procedures
Construct and pretest a questionnaire for data collection
Specify the sampling process and sample size
Develop a plan for data analysis
Research Design Formulation
Data collection
Data preparation
Data analysis
Validation/Falsification of theory
Field Work & Assessment
Report preparation and presentation
Decision
Decision

**Fig. 1.3** The phases of empirical research

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

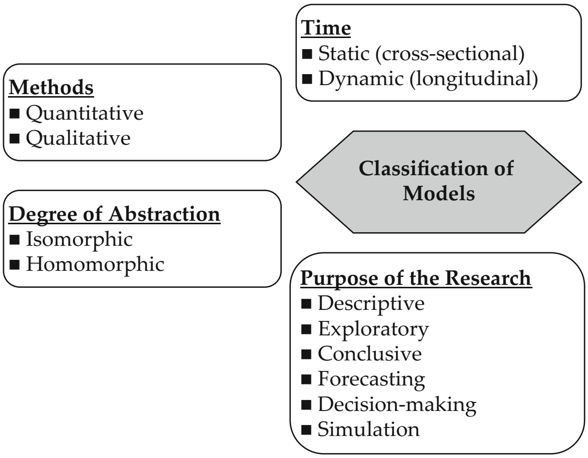


**Fig. 1.4** A systematic overview of model variants

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 (Fig. 1.4). 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 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 analyze, 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 serve 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 necessary 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: Price at point in time i; $w_i$: marketing expenditures at point in time I; $\alpha$ refers to the effectiveness of each variable; $\beta$ 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

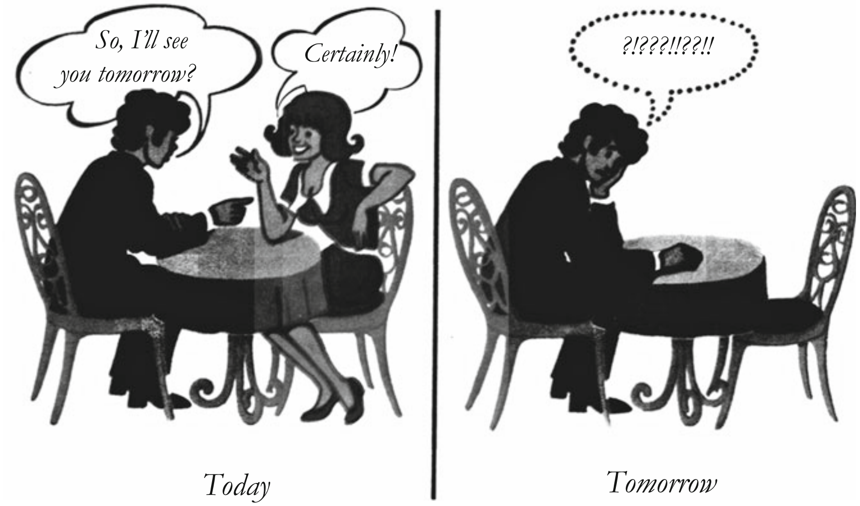


**Fig. 1.5** What is certain? (Source: Swoboda 1971, p. 31)

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. 1.5 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 the image above 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 (cf. 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

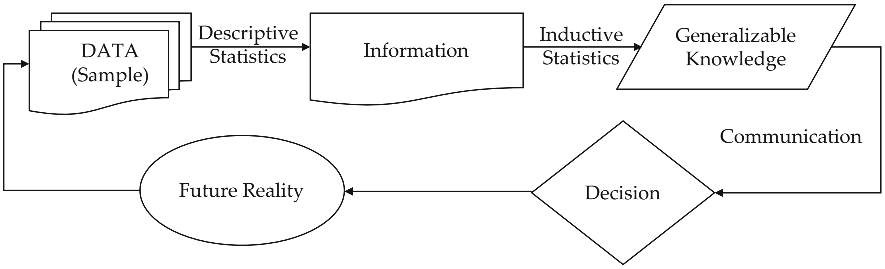


**Fig. 1.6** The intelligence cycle (Source: Own graphic, adapted from Harkleroad 1996, p. 45)

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 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) (Fig. 1.6).

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 the above figure, 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, analyzed, 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. 1.3). 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.

# Disarray to Dataset

# 2

## 2.1 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. The following 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

Chapter 2 Translated from the German original, Cleff, T. (2011). 2 Vom Zahlenwust zum Datensatz. In *Deskriptive Statistik und moderne Datenanalyse* (pp. 15–29) © Gabler Verlag, Springer Fachmedien Wiesbaden GmbH, 2011.

T. Cleff, *Exploratory Data Analysis in Business and Economics*,
DOI 10.1007/978-3-319-01517-0_2, © Springer International Publishing Switzerland 2014

**Table 2.1** External data sources at international institutions

| | | |
|---|---|---|
| German federal statistical office | destatis.de | Offers links to diverse international data bases |
| Eurostat | epp.eurostat.ec.europa.eu | Various databases |
| OECD | oecd.org | Various databases |
| Worldbank | worldbank.org | World & 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. |

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 delivery capacity. So the procurement (or inventory) department decides to consult the most recent Ifo Business Climate Index.[1] Using this information, the department staff can create a valid forecast of the end-user industry for the next 6 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.[2]

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 one 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 provides, 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 reflects only a household, but not its members.

[1] The Ifo Business Climate Index is released each month by Germany's Ifo Institute. It is based on a monthly survey that queries some 7,000 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.

[2] For more, see the method described in Chap. 5.

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.

## 2.2 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 2010). Consequently, we focus below on the criteria for choosing a specific quantitative assessment method.

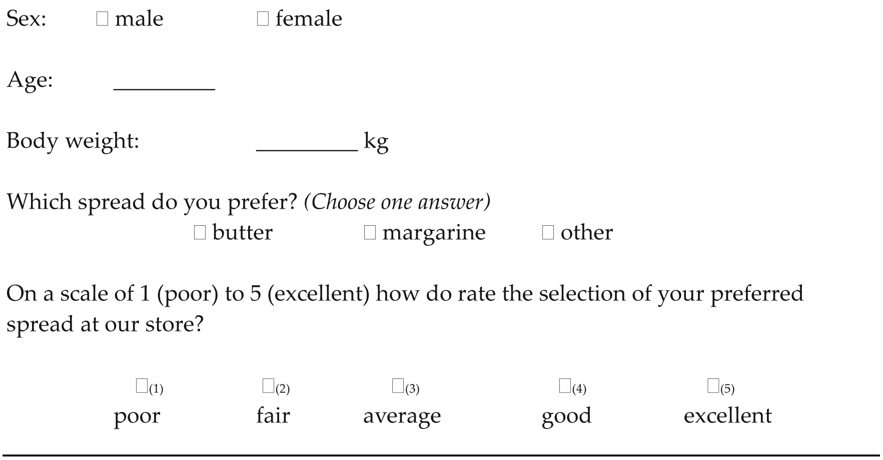
Let us begin with an example. Imagine you own a little grocery store in a small town. Several customers have requested that you expand your selection of butter and margarine. Because you have limited space for display and storage, you want to know whether this request is *representative* of the preferences 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 *sex*, *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 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



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) poor □(2) fair □(3) average □(4) good □(5) excellent

**Fig. 2.1** Retail questionnaire

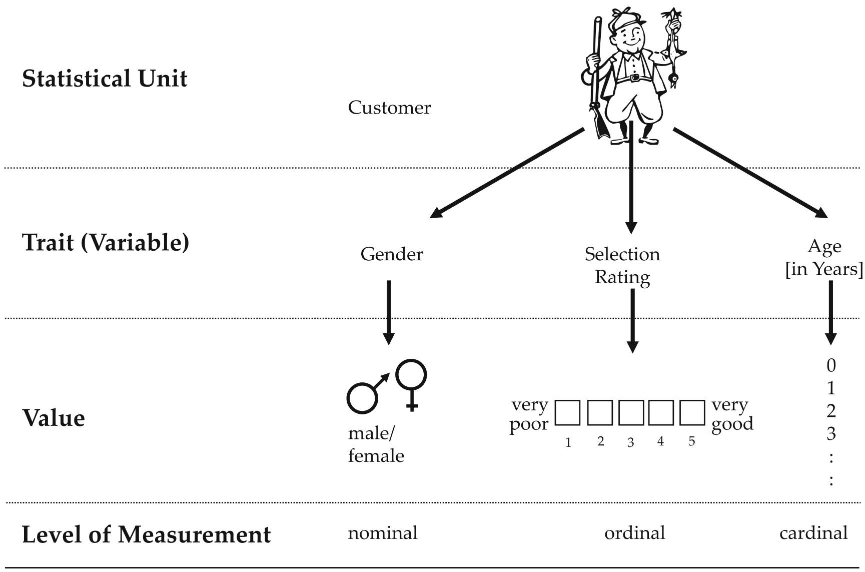


**Fig. 2.2** Statistical units/Traits/Trait values/Level of measurement

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 *sex*, a *1* for *male* is no better or worse than a *2* 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 1 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 necessary 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 2 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,[3] ratio scales,[4] and interval scales.[5] 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 scale 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,

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

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

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

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 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. 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 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.

## 2.3 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 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, p. 24). 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 "*selectio*".

```
------------------------------------------------------------------------
value label: selectio
------------------------------------------------------------------------
  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: sex
------------------------------------------------------------------------
  definition
           0   male
           1   female
   variables:  gender
```

**Fig. 2.3** Label book

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 – e.g. "Compute graphic for the variable selectio" – 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 scale 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.

## 2.4 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*. Even with complex techniques such as regression analysis, it is usually possible to interpret missing values to some extent. We'll address this issue again in later chapters.

(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. 5.

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.

## 2.5 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. 2.4.

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.

## 2.6 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 Stata, SPSS, or Excel for the questionnaire in Fig. 2.1 (p. 16) and enter the data from Fig. 3.1 (p. 24). Allow for missing values in the dataset.

# 3 Univariate Data Analysis

## 3.1 First Steps 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. In the first step of data assessment, they investigate each variable – for example, average respondent age – separately. This is called *univariate analysis* (Fig. 3.1). By contrast, when researchers analyze the relationship between two variables – for example, between gender and choice of spread – this is called *bivariate analysis* (see Sect. 4). With relationships between more than two variables, one speaks of *multivariate analysis* (see Sect. 5.3).

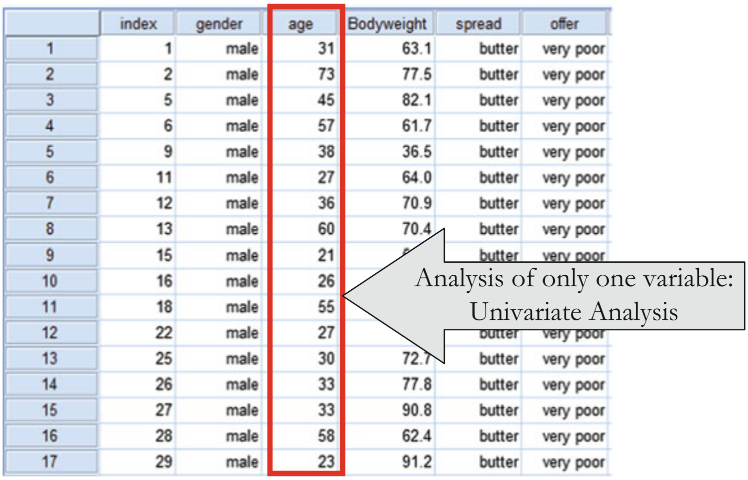
How can the results of 850 responses be "distilled" to create a realistic and accurate impression of the surveyed attributes and their relationships? Here the importance of statistics becomes apparent. Recall the professor who was asked about the results of the last final exam. The students expect distilled 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]

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.

Chapter 3 Translated from the German original, Cleff, T. (2011). 3 Vom Datensatz zur Information. In *Deskriptive Statistik und moderne Datenanalyse* (pp. 31–77) © Gabler Verlag, Springer Fachmedien Wiesbaden GmbH, 2011.

[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 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.

T. Cleff, *Exploratory Data Analysis in Business and Economics*,
DOI 10.1007/978-3-319-01517-0_3, © Springer International Publishing Switzerland 2014



| | 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

**Note**: 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

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 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:

$$F(x_p) = f(x_1) + f(x_2) + \cdots + f(x_p) = \sum_{i=1}^{p \le n} f(x_i) \qquad (3.1)$$

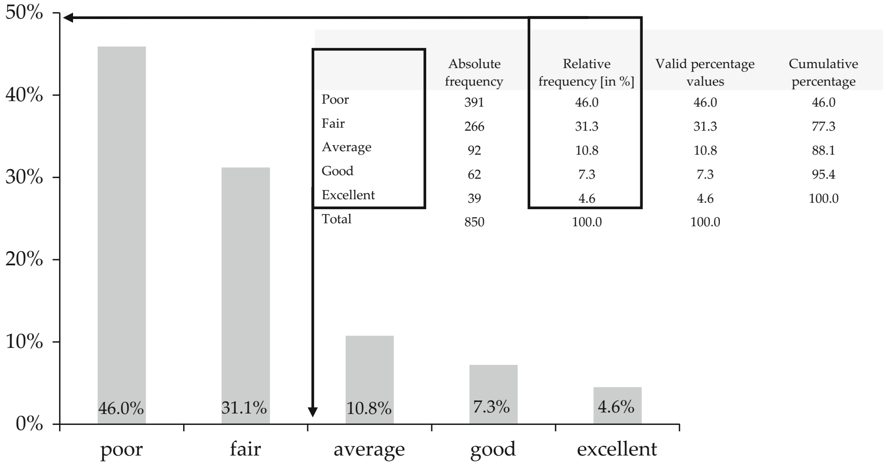
These results can also be represented graphically as a *pie chart*, a *horizontal bar chart*, or a *vertical bar chart*. All three diagram forms can be used with nominal and ordinal variables, though pie charts are used mostly for nominal variables.

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

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

| | 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



| | 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

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 (Fig. 3.4).

In many publications, the scaling on the y-axis of a vertical bar chart begins not with 0 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).

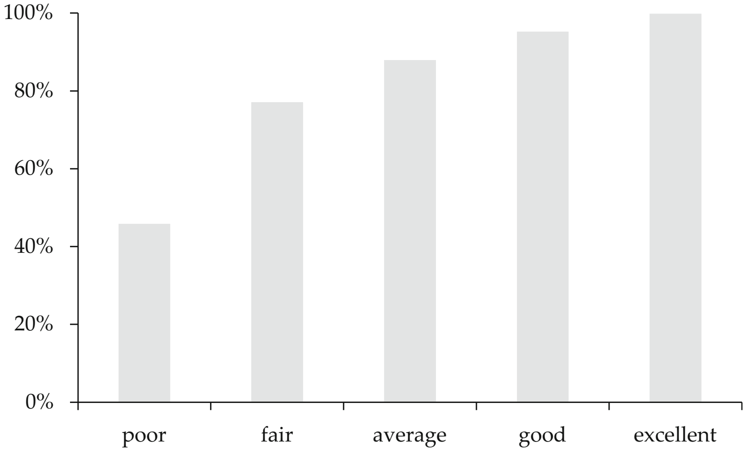
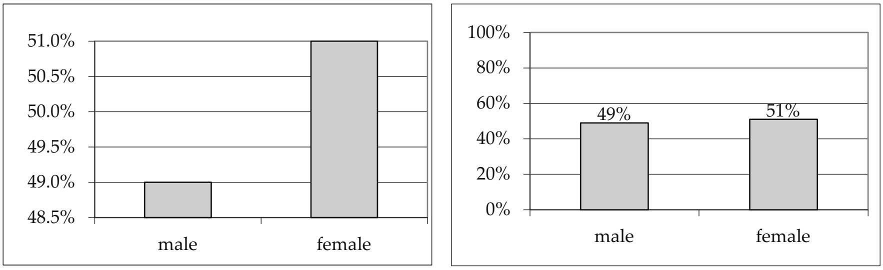


**Fig. 3.4** Distribution function for the selection variable



Part 1

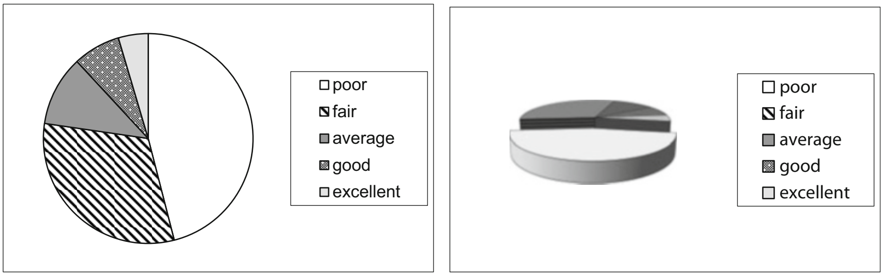
Part 2

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

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 2 % points seems larger than it actually is. For this reason, the second graph in Fig. 3.5 is the preferable form of representation.

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 $\alpha_i = f(x_\mathrm{i}) \cdot 360°$.

Since most viewers read pie charts clockwise from the top, the traits to be emphasized should be placed in the 12 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.



Part 1

Part 2

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

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 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.

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

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

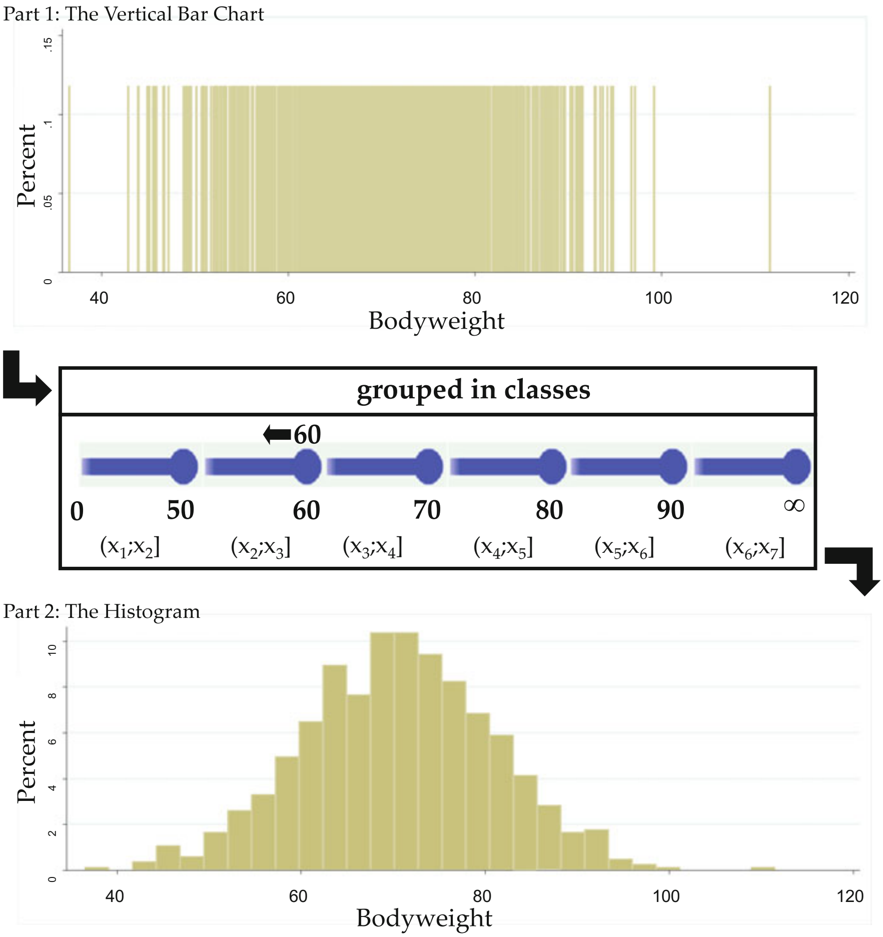
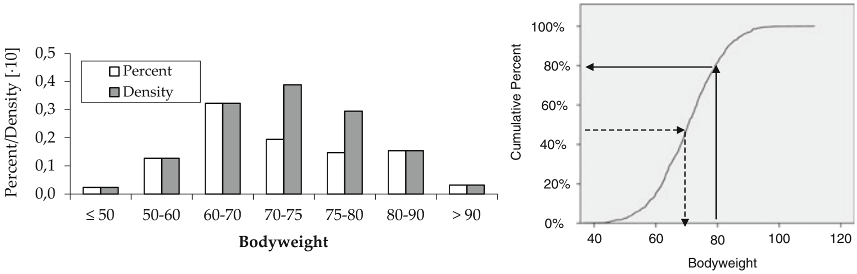


**Fig. 3.7** Using a histogram to classify data

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 5 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.



Part 1

Part 2

**Fig. 3.8** Distorting interval selection with a distribution function

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.

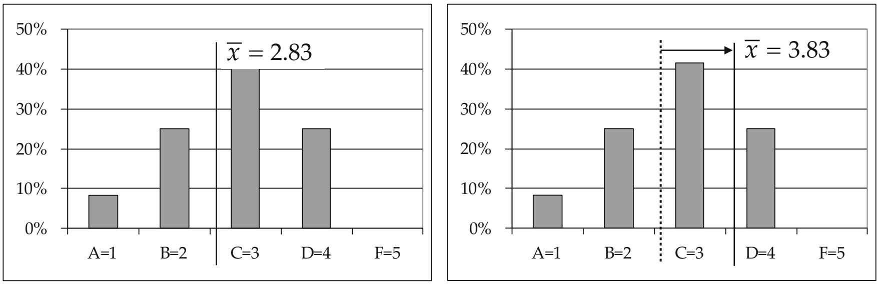
The distribution function of a cardinal variable can be represented as unclassed. Here too, the frequencies are cumulative as one moves along the x-axis. The values of the distribution function rise evenly and remain between 0 and 1. The distribution function for the bodyweight variable is represented 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.



Part 1 Part 2

**Fig. 3.9** Grade averages for two final exams

### 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 appears in a raw data table that includes all the individual trait values. For raw data tables, the mean is derived from the formula:

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

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

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 3 g 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

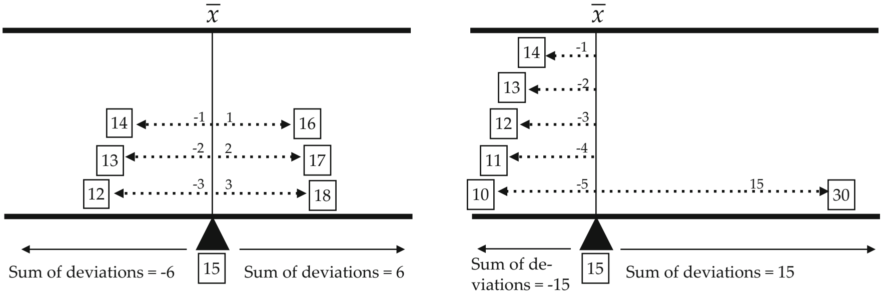


**Fig. 3.10** Mean expressed as a balanced scale

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:

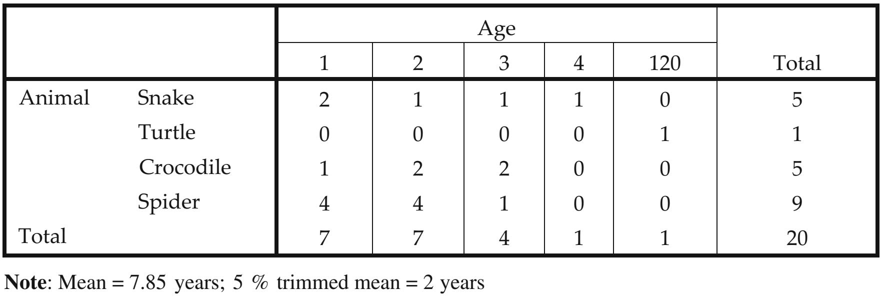
$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0 \tag{3.3}$$

In real life, if a heavy weight is on 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 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. 2.5; in such cases, an outlier value is usually responsible for distorting the results. Assume you want to calculate the average age of animals in a zoo terrarium containing five snakes, nine spiders, five crocodiles, and one turtle. The last animal – the turtle – is 120 years old, while all the others are no older than four (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 2 years, a more realistic description of the age distribution. We should 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:

$$\bar{x} = \frac{1}{n} \sum_{v=1}^{k} x_v \cdot n_v = \sum_{v=1}^{k} x_v \cdot f_v \tag{3.4}$$



| | | Age | | | | | Total |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 120 | |
| 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 |

**Note**: Mean = 7.85 years; 5 % trimmed mean = 2 years

**Fig. 3.11** Mean or trimmed mean using the zoo example

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:

$$\bar{x} = \frac{1}{850} \cdot (391 \cdot 1 + 266 \cdot 2 + 92 \cdot 3 + 62 \cdot 4 + 39 \cdot 5) = 1.93 \qquad (3.5)$$

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 $f_v$:

$$\bar{\text{x}} = (0.46 \cdot 1 + 0.313 \cdot 2 + 0.108 \cdot 3 + 0.073 \cdot 4 + 0.046 \cdot 5) = 1.93 \qquad (3.6)$$

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

$$\bar{x} = \frac{1}{n} \sum_{v=1}^{k} n_v m_v = \sum_{v=1}^{k} f_v m_v, \qquad (3.7)$$

where $m_v$ is the mean of class number $v$.

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.

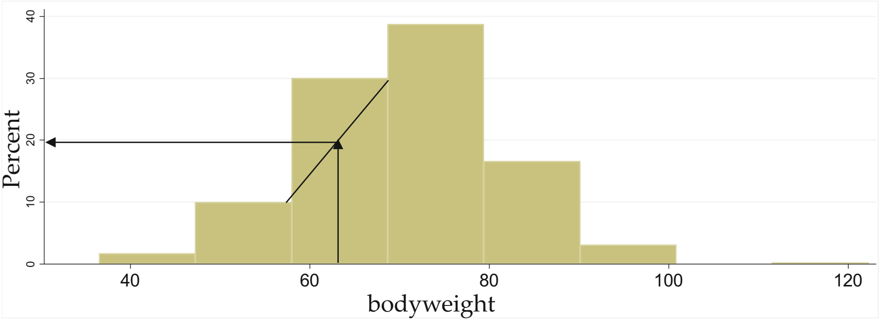


**Fig. 3.12** Calculating the mean from classed data

**Table 3.1** Example of mean calculation from classed data

| Water use [in l] | 0–200 | 200–400 | 400–600 | 600–1,000 |
|---|---|---|---|---|
| Rel. frequency | 0.2 | 0.5 | 0.2 | 0.1 |

Source: Schwarze (2008, p. 16), translated from the German

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.

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

The water-use average can be calculated as follows:

$$\bar{x} = \sum_{v=1}^{k} f_v m_v = \sum_{v=1}^{4} f_v m_v = 0.2 \cdot 100 + 0.5 \cdot 300 + 0.2 \cdot 500 + 0.1 \cdot 800 = 350 \quad (3.8)$$

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 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 (2005, p. 61). 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 necessary 25 MUs. If males constitute 70 % of the workforce, the average salary will be: $0.7 \cdot 30$ MU + $0.3 \cdot 20$ MU = 27 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 10 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 (2005, p. 70) 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

| | | | Changes in sales when using | |
|---|---|---|---|---|
| Year | Sales [mio.] | Rate of change [in %] | arithm. mean | geom. mean |
| 2002 | €20,000.00 | | €20,000.00 | €20,000.00 |
| 2003 | €22,000.00 | 1.000% | €20,250.00 | €20,170.56 |
| 2004 | €20,900.00 | -5.000% | €20,503.13 | €20,342.57 |
| 2005 | €18,810.00 | -10.000% | €20,759.41 | €20,516.04 |
| 2006 | €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

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 5 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 2002 should have increased to €21,018.91 by 2006, but actual sales in 2006 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 2002 with the subsequent rates of growth each year until 2006. The result is:

$$\begin{aligned} \mathrm{U}_6 &= \mathrm{U}_5 \cdot (1+0.1) = (\mathrm{U}_4 \cdot (1-0.1)) \cdot (1+0.1) = \ldots \\ &= (\mathrm{U}_2 \cdot (1+0.1)) \cdot (1-0.05) \cdot (1-0.1) \cdot (1+0.1). \end{aligned} \tag{3.9}$$

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

$$\left(1+\overline{p}_{geom}\right)\cdot\left(1+\overline{p}_{geom}\right)\cdot\left(1+\overline{p}_{geom}\right)\cdot\left(1+\overline{p}_{geom}\right) = \left(1+\overline{p}_{geom}\right)^4 \tag{3.10}$$

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

$$\overline{p}_{geom} = \sqrt[4]{(1+0.1)\,(1-0.05)\,(1-0.1)\,(1+0.1)} - 1 = 0.853 \tag{3.11}$$

The last column in Fig. 3.13 shows that this value correctly describes the sales growth between 2002 and 2006. Generally, the following formula applies for identifying *average rates of change*:

$$\overline{p}_{geom} = \sqrt[n]{(1+p_1)\cdot(1+p_2)\cdot\cdot(1+p_n)} - 1 = \sqrt[n]{\prod_{i=1}^{n}(1+p_i)} - 1 \tag{3.12}$$

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

$$\bar{x}_{geom} = \sqrt[n]{x_1 \cdot x_2 \cdot \ \ldots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i} \tag{3.13}$$

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:

$$\bar{x} = \frac{1}{3}\left(\frac{S_1}{E_1} + \frac{S_2}{E_2} + \frac{S_3}{E_3}\right) = €433.33 \tag{3.14}$$

If this value were equally applicable to all employees, the firms – which have 16 employees together – would have sales totalling 16·€433.33 ≈ €6,933, but the above table shows that actual total sales are only €3,000. When calculating company sales, it must be taken into account that the firms employ varying numbers 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,. . ., 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:

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

**Table 3.2** Harmonic mean

| | Sales | Employees | Sales per employee (SP) | Formula in Excel |
|---|---|---|---|---|
| | €1,000 | 10 | €100.00 | |
| | €1,000 | 5 | €200.00 | |
| | €1,000 | 1 | €1,000.00 | |
| Sum | €3,000 | 16 | €1,300.00 | SUM(D3:D5) |
| Arithmetic mean | | | €433.33 | AVERAGE(D3:D5) |
| Harmonic mean | | | €187.50 | HARMEAN(D3:D5) |

$$\frac{n_1 \cdot SP_1 + n_2 \cdot SP_2 + n_3 \cdot SP_3}{n} = \frac{10 \cdot €100}{16} + \frac{5 \cdot €200}{16} + \frac{1 \cdot €1,000}{16} \approx €187.50 \quad (3.15)$$

Using this formula, the 16 employees generate the real total sales figure of €3,000. 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*:

$$\bar{x}_{harm} = \frac{k}{\sum_{i=1}^{k} \frac{1}{x_i}} = \frac{k}{\sum_{i=1}^{k} \frac{1}{SP_i}} = \frac{3}{\frac{1}{€100} + \frac{1}{€200} + \frac{1}{€1,000}} = \frac{€187.50}{\text{Employee}} \quad (3.16)$$

Let's look at another example that illustrates the harmonic mean. A student must walk 3 km 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:

$$\bar{x} = \frac{1}{3}\left(2\frac{km}{h} + 3\frac{km}{h} + 4\frac{km}{h}\right) = 3\frac{km}{h}, \text{or 1 hour} \quad (3.17)$$

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. 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:

$$\bar{x}_{harm} = \frac{k}{\sum_{i=1}^{k} \frac{1}{x_i}} = \frac{3}{\frac{1}{2\,\frac{\text{km}}{\text{h}}} + \frac{1}{3\frac{\text{km}}{\text{h}}} + \frac{1}{4\frac{\text{km}}{\text{h}}}} = 2.77\frac{\text{km}}{\text{h}} \quad (3.18)$$

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

In our previous examples the values in the numerator were identical for every observation. In the first example, all three companies had sales of €1,000 and in the second example all route segments were 1 km. 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 $n_1$ = €1,000, $n_2$ = €2,000, and $n_3$ = €5,000, we would use the following calculation:

$$\bar{x}_{harm} = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}} = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{SP_i}} = \frac{€1,000 + €2,000 + €5,000}{\frac{€1,000}{€100} + \frac{€2,000}{€200} + \frac{€5,000}{€1,000}} = \frac{€500}{\text{Employee}} \tag{3.19}$$

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 2 km/h, a second hour at 3 km/h, and the last hour at 4 km/h, the arithmetic mean yields the correct the 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):

$$\bar{x} = \frac{1}{3}\left(2\frac{km}{h} + 3\frac{km}{h} + 4\frac{km}{h}\right) = 3\frac{km}{h} \tag{3.20}$$

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 0. For unequally sized observations, the following applies:

$$\bar{x}_{harm} < \bar{x}_{geom} < \bar{x} \tag{3.21}$$

### 3.2.5 The Median

As the mean is sometimes not "representative" 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 following data (Table 3.3):

Based on what we learned above about calculating the mean using the class midpoint of classed data, we get: $\bar{x} =$ 0.3·0.5 + 0.15·1.5 + 0.25·3.5 + 0.04·8

**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 |

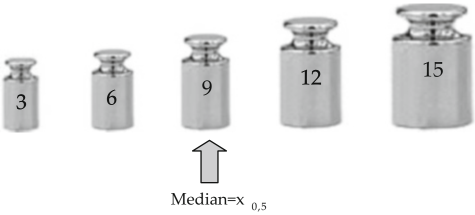


**Fig. 3.14** The median: The central value of unclassed data

$+0.03{\cdot}36 + 0.23{\cdot}81 \approx 21$ years.[8] 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 average age of diaper users? Surely one way would be to find the modal value of the most important age group: 0–1. This value, the so-called *median,* 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 3.

There are several formula for calculating the median. When working with a raw data table – i.e. with unclassed data – most statistics textbooks suggest these formula:

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)} \text{ for an odd number of observations (n)} \qquad (3.22)$$

and

[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.

$$\tilde{x} = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) \text{ for an even number of observations.} \tag{3.23}$$

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

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{5+1}{2}\right)} = x_{(3)} = 9 \tag{3.24}$$

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:

$$\tilde{x} = x_{0.5} = x_{i-1}^{UP} + \frac{0.5 - F\left(x_{i-1}^{UP}\right)}{f(x_i)}\left(x_i^{UP} - x_i^{LOW}\right) \tag{3.25}$$

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

$$0.5 - F\left(x_{i-1}^{UP}\right) = 0.5/0.45 = 0.05 \tag{3.26}$$

This 5 % points must be accounted for from the next largest (ith) class, as it must contain the median. The 5 % points are then set in relation to the relative frequency of the entire class:

$$\frac{0.5 - F\left(x_{i-1}^{UP}\right)}{f(x_i)} = \frac{0.5 - 0.45}{0.25} = 0.2 \tag{3.27}$$

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 3 years, as the class contains all persons who are 2, 3, and 4 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. 2.5). Figure 3.15 traces the steps for us once more.

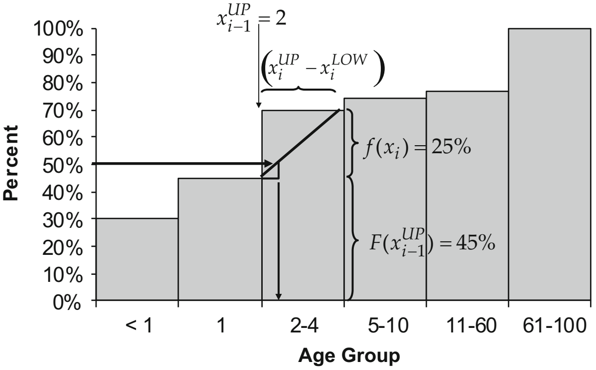


**Fig. 3.15** The median: The middle value of classed data

### 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)·p. In our example, (850 + 1)·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

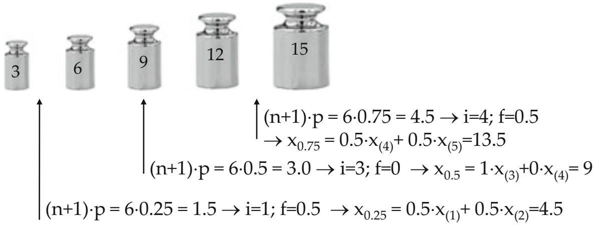

3
6
9
12
15
$(n+1) \cdot p = 6 \cdot 0.75 = 4.5 \rightarrow i=4; f=0.5$
$\rightarrow x_{0.75} = 0.5 \cdot x_{(4)} + 0.5 \cdot x_{(5)} = 13.5$
$(n+1) \cdot p = 6 \cdot 0.5 = 3.0 \rightarrow i=3; f=0 \rightarrow x_{0.5} = 1 \cdot x_{(3)} + 0 \cdot x_{(4)} = 9$
$(n+1) \cdot p = 6 \cdot 0.25 = 1.5 \rightarrow i=1; f=0.5 \rightarrow x_{0.25} = 0.5 \cdot x_{(1)} + 0.5 \cdot x_{(2)} = 4.5$

**Fig. 3.16** Calculating quantiles with five weights

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:

$$(1-f)\cdot x_{(i)} + f \cdot x_{(i+1)} \tag{3.28}$$

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

$$(1-0.75)\cdot x_{(212)} + 0.75 \cdot x_{(213)} = 0.25 \cdot 63.38 + 0.75 \cdot 63.44 = 63.43\ \text{kg} \tag{3.29}$$

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

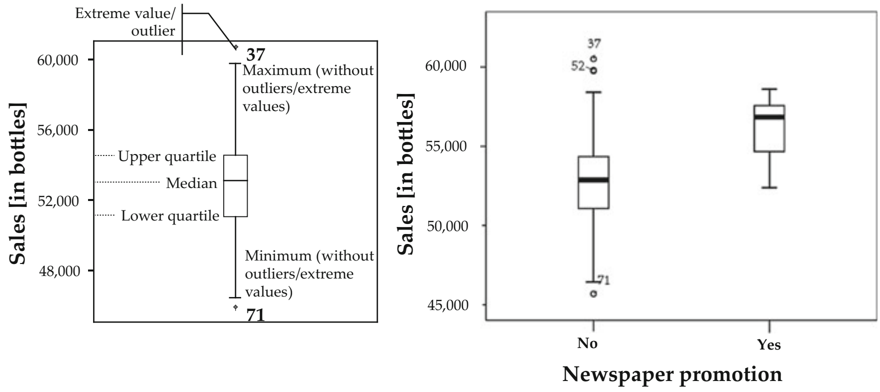
It should be noted here that the weighted average method cannot be used with extreme quantiles. For example, to determine the 99 % quantile for the five weights in Fig. 3.16 a sixth weight is needed, since (n + 1)·p = (5 + 1)· 0.99 = 5.94. This weight does not actually exist. It is fictitious, just like a weight of 0 for determining the 1 % quantile ((n + 1)·p = (5 + 1)·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: $x_{0.99} = 15$ and $x_{0.01} = 3$.

## 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 3 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

[10] The data can be found in the file *salad_dressing.sav* at springer.com.



Part 1 Part 2

**Fig. 3.17** Boxplot of weekly sales

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. 2.5).

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.
- 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 %.

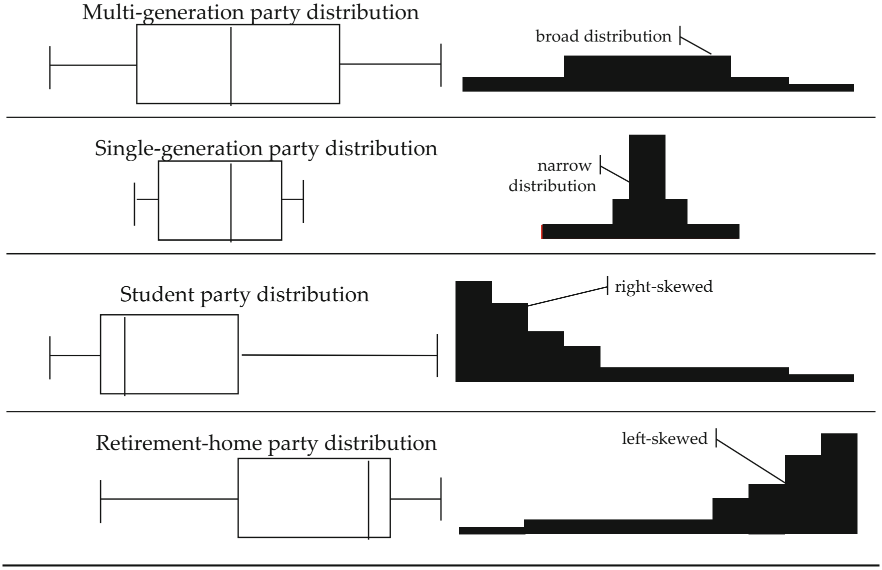


**Fig. 3.18** Interpretation of different boxplot types

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 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 is 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

$$\mathrm{IQR} = (x_{0.75} - x_{0.25}) \tag{3.30}$$

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

$$\mathrm{MQR} = 0.5 \cdot (x_{0.75} - x_{0.25}) \tag{3.31}$$

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:

$$\mathrm{Range} = \mathrm{Max}(\mathrm{x_i}) - \mathrm{Min}(\mathrm{x_i}) \tag{3.32}$$

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 external 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:

$$\mathrm{MAD} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \tilde{x}| \tag{3.33}$$

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:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}) \tag{3.34}$$

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 0. 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:

$$\frac{1}{n}\sum_{i=1}^{n}|x_i-\bar{x}| \tag{3.35}$$

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:

$$Var(x)_{emp}=S_{emp}^{2}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{3.36}$$

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

$$S_{emp}=\sqrt{Var(x)_{emp}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^{2}} \tag{3.37}$$

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$\bar{x} = 3$. Their distribution has four deviations of one unit each. The squared sum of the deviations is:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 1^2 + 1^2 + 1^2 + 1^2 = 4 \text{ units} \tag{3.38}$$

Another distribution contains the observations 2, 4, 4, and 6. Their mean is four, or$\bar{x} = 4$, and the total sum of deviations again is $2 + 2 = 4$ units. Here, two observations have a deviation of 2 and two observations have a deviation of 0. But the sum of the squared deviation is larger:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 2^2 + 0^2 + 0^2 + 2^2 = 8 \text{ units} \tag{3.39}$$

Although the sum of the deviations is identical in each case, a few large deviations lead to a larger empirical variance than many small deviations with the same quantity ($S_{emp}^2 = 1$ versus $S_{emp}^2 = 2$). 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 1 (n−1). Here one speaks of *unbiased sample variance*, or of *Bessel's corrected variance*:

$$Var(x) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3.40}$$

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

$$S = \sqrt{Var(x)} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{3.41}$$

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 analyzed, 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 $S_{emp} = 5.1$ cm – or $S_{emp} = 2.0$ 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:

$$V = \frac{S}{|\bar{x}|}, \text{provided the mean does not have the value } \bar{x} = 0 \tag{3.42}$$

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 grammes;
- . . .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.

| | | 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

## 3.5 Skewness and Kurtosis

The boxplot in Fig. 3.18 not only provides information about central tendency and dispersion, but also describes the symmetry of the distribution. 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 & Pearson* uses the difference between median and mean in asymmetric distributions. Look again 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 $\bar{x} > \tilde{x}$. 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 $(\bar{x} < \tilde{x})$. *Yule & Pearson* express the difference between median and mean as a degree of deviation from symmetry:

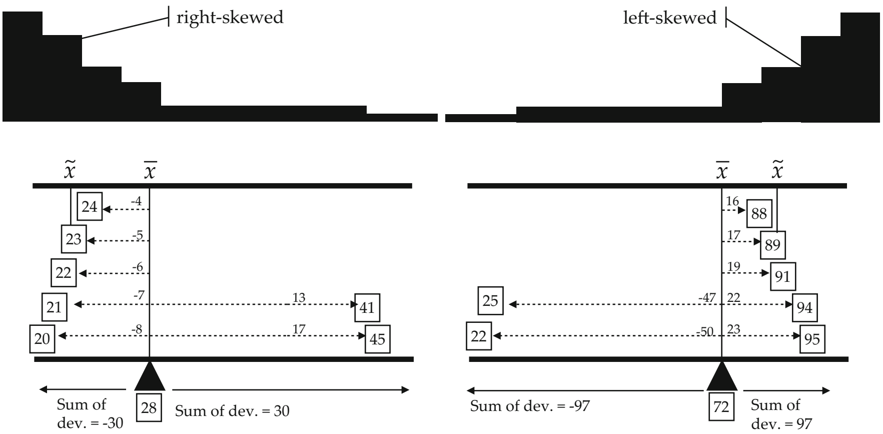
$$\text{Skew} = \frac{3 \cdot (\bar{x} - \tilde{x})}{S} \tag{3.43}$$

Values larger than 0 indicate a right-skewed distribution, values less than 0 indicate a left-skewed distribution, and values that are 0 indicate a symmetric distribution.

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

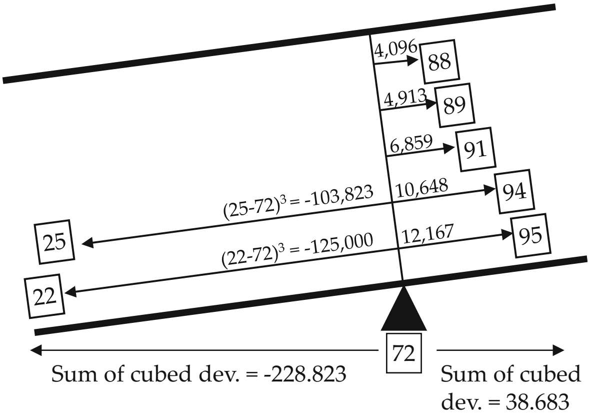
$$\text{Skew} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{S^3} \tag{3.44}$$

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 years to 72 years. Nevertheless, the sum of deviations on the left and right must be identical. The residents of the



**Note**: 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.

**Fig. 3.20** Skewness



**Note**: 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** The third central moment

retirement home 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 $\left((x_i - \bar{x})^3\right)$ before being summed? Cubing produces a value for caregiver ages of

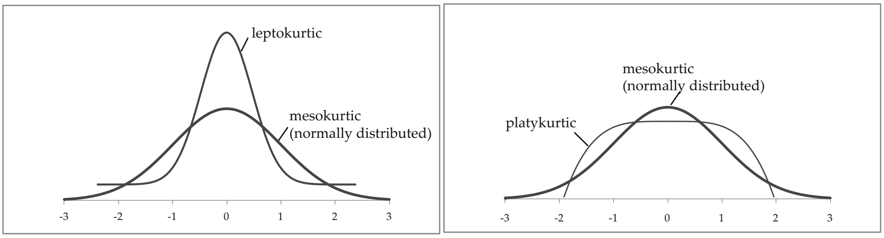


**Fig. 3.22** Kurtosis distributions

−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.

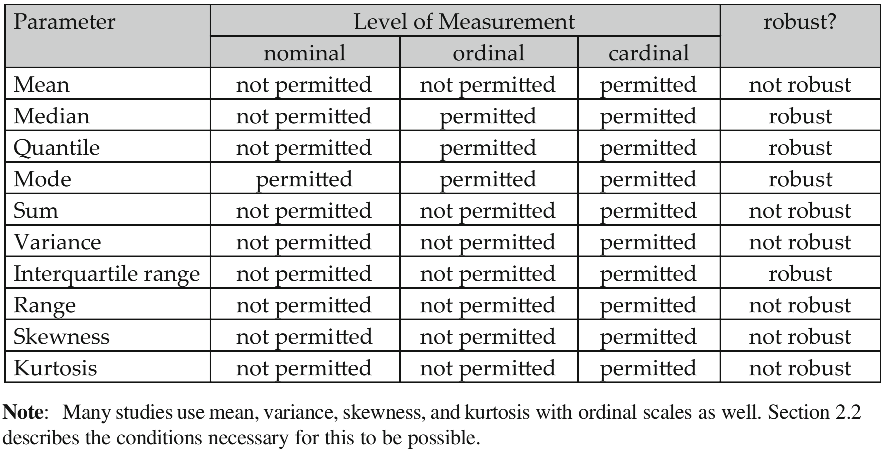
The total sum of the values from the left and right hand sides results in a negative value of −190,140 (= −228,823 + 38,683) for the left-skewed distribution. For a right-skewed distribution, the result is positive, and for symmetric distributions the result is close to 0. A value is considered different than 0 when the absolute value of the skewness is more than twice as large as the *standard error* of the skew. This means that a skewness of 0.01 is not necessary different than 0. The standard error is always indicated in statistics programmes and does not need to be discussed further here.

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:

$$\mathrm{Kurt} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{S^4} \tag{3.45}$$

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 0, not 3. The user must be aware of which formula is being used when calculating kurtosis.



| 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 |

**Note**: Many studies use mean, variance, skewness, and kurtosis with ordinal scales as well. Section 2.2 describes the conditions necessary for this to be possible.

**Fig. 3.23** Robustness of parameters

## 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.

## 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 rates 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.

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

$$\mathrm{H} = \sum_{i=1}^{n} f(x_i)^2 \tag{3.46}$$

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:

$$\mathrm{H} = \sum_{i=1}^{n} f(x_i)^2 = 0.2^2 + 0.2^2 + \ 0.2^2 + 0.2^2 + 0.2^2 = 0.2 \tag{3.47}$$

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

$$\mathrm{H} = \sum_{i=1}^{n} f(x_i)^2 = 1^2 + 0^2 + 0^2 + 0^2 + 0^2 = 1 \tag{3.48}$$

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 1 (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 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 % 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 0 (uniform distribution, since

| | 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

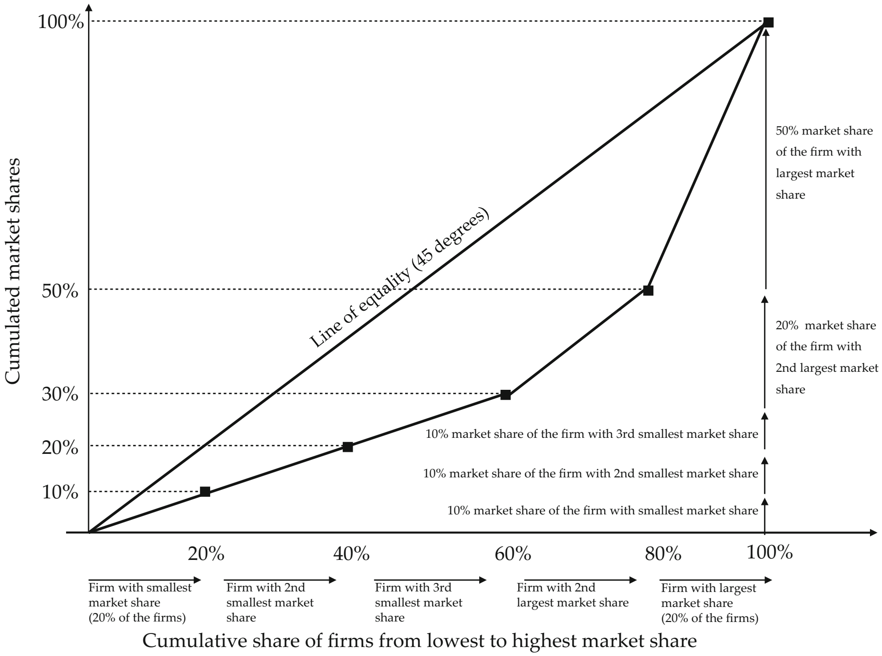


**Fig. 3.25** Lorenz curve

otherwise the area between the bisector and the Lorenz curve would be 0) and (n−1)/n (full possession of all shares by a statistical unit):

$$\text{GINI} = \frac{\text{Area between bisector and the Lorenz curve}}{\text{Entire area below the bisector}} \tag{3.49}$$

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

(a) For unclassed ordered raw data:

$$\mathrm{GINI} = \frac{2\sum_{i=1}^{n} i \cdot x_i - (n+1)\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i} \tag{3.50}$$

(b) For unclassed ordered relative frequencies:

$$\mathrm{GINI} = \frac{2\sum_{i=1}^{n} i \cdot f_i - (n+1)}{n} \tag{3.51}$$

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

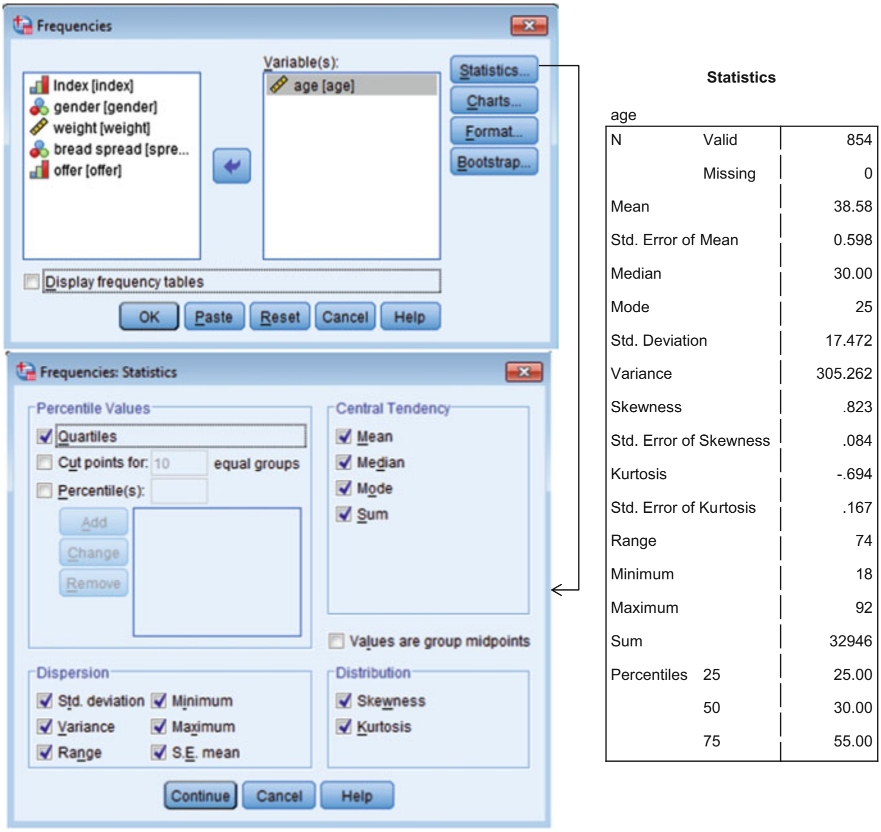
$$\begin{aligned}\mathrm{GINI} &= \frac{2\sum_{i=1}^{n} i \cdot f_i - (n+1)}{n} \\ &= \frac{2 \cdot (1 \cdot 0.1 + 2 \cdot 0.1 + 3 \cdot 0.1 + 4 \cdot 0.2 + 5 \cdot 0.5) - (5+1)}{5} \\ &= 0.36\end{aligned} \tag{3.52}$$

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 $\mathrm{GINI}_{\mathrm{norm.}} = 1$.

## 3.8 Using the Computer to Calculate Univariate Parameters

### 3.8.1 Calculating Univariate Parameters with SPSS

This section uses the sample dataset *spread.sav*. 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.



Frequencies

Index [index]
gender [gender]
weight [weight]
bread spread [spre...
offer [offer]

Variable(s):
age [age]

□ Display frequency tables

Frequencies: Statistics

Percentile Values
☑ Quartiles
□ Cut points for: 10 equal groups
□ Percentile(s):

Central Tendency
☑ Mean
☑ Median
☑ Mode
☑ Sum

□ Values are group midpoints

Dispersion
☑ Std. deviation ☑ Minimum
☑ Variance ☑ Maximum
☑ Range ☑ S.E. mean

Distribution
☑ Skewness
☑ Kurtosis

**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 |

**Note**: Applicable syntax commands: Frequencies; Descriptives

**Fig. 3.26** Univariate parameters with SPSS

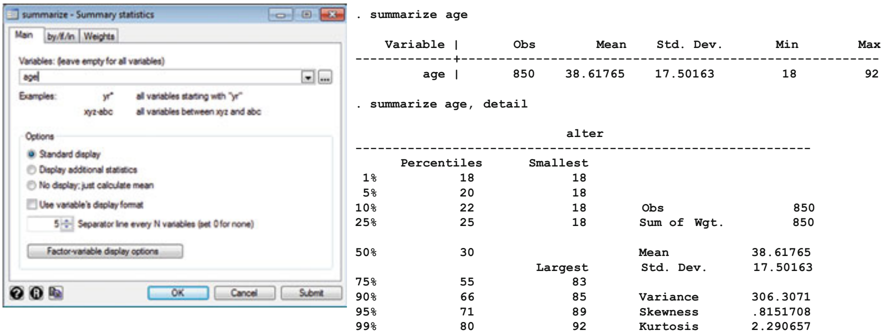
SPSS uses a standard kurtosis of 0, not 3. Figure 3.26 shows the menu and the output for the age variable from the sample dataset.

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.

## 3.8.2 Calculating Univariate Parameters with Stata

Let's return again to the file *spread.dta*. 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


summarize - Summary statistics
Main
by/if/in
Weights
Factor-variable display options
OK
Cancel
Submit

Variables: (leave empty for all variables)
age

Examples: yr* all variables starting with "yr"
xyz-abc all variables between xyz and abc

Options
● Standard display
○ Display additional statistics
○ No display; just calculate mean
□ Use variable's display format
5 Separator line every N variables (set 0 for none)

```
. 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
```

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

**Fig. 3.27** Univariate parameters with Stata

descriptive statistics, make sure to select *Display additional statistics*, as otherwise only the mean, variance, and smallest and greatest values will be displayed. Figure 3.27 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.

### 3.8.3 Calculating Univariate Parameters with Excel 2010

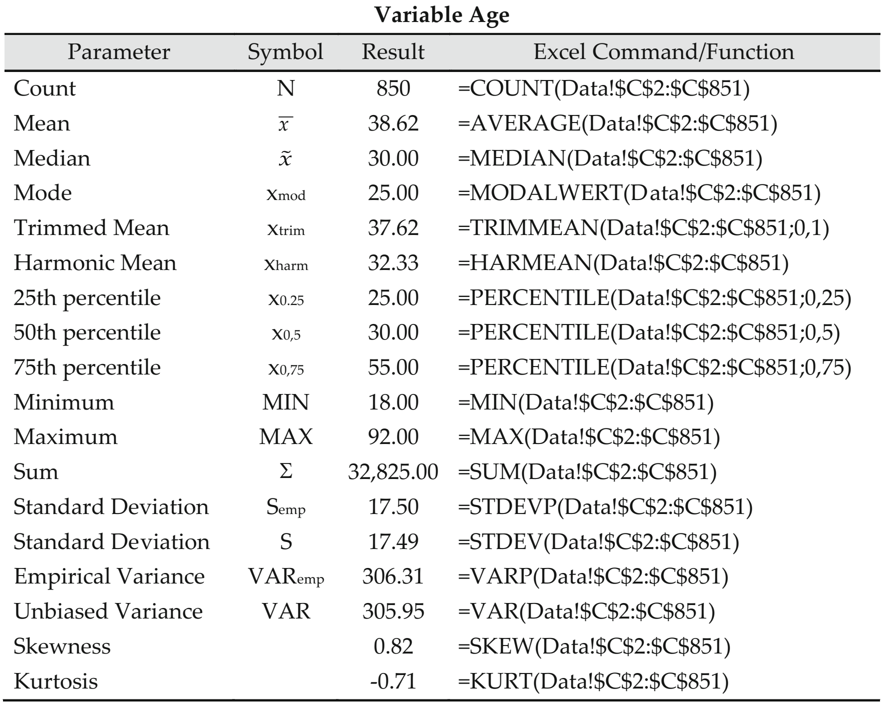
Excel contains a number of preprogrammed statistical functions. These functions can be found under *Formulas* → *Insert Function*. Select the category *Statistical* to set the constraints. Figure 3.28 shows the Excel functions applied to the dataset *spread.xls*. It is also possible to use the Add-in Manager[11] to permanently activate the *Analysis ToolPak* and the *Analysis ToolPak VBA* for Excel 2010. 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.

Go to http://www.reading.ac.uk/ssc/n/software.htm 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 0 and not 3, which equals a subtraction of 3.

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

Example: Calculation of univariate parameters of the dataset *spread.xls*



**Variable Age**

| Parameter | Symbol | Result | Excel Command/Function |
|---|---|---|---|
| Count | N | 850 | =COUNT(Data!$C$2:$C$851) |
| Mean | $\bar{x}$ | 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) |
| Standard Deviation | $S_{emp}$ | 17.50 | =STDEVP(Data!$C$2:$C$851) |
| Standard Deviation | S | 17.49 | =STDEV(Data!$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. 3.28** Univariate parameters with Excel

## 3.9 Chapter Exercises

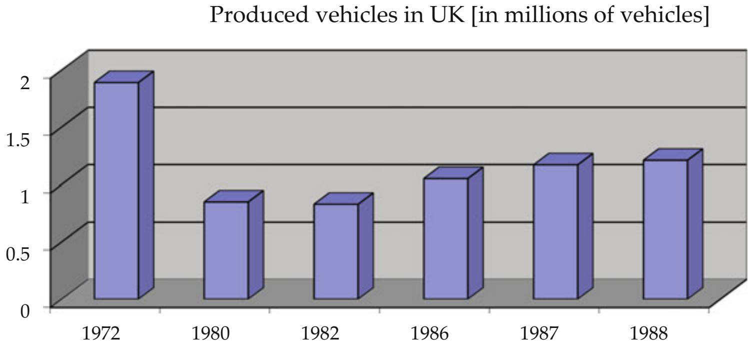
### Exercise 4:

A spa resort in the German town of Waldbronn 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 5:**
Supposed the following figure appears in a market research study. What can be said about it?



**Exercise 6**:
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 7:**
The arithmetic mean $\bar{x} = 10$ and the empirical standard deviation $S_{emp} = 2$ 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 8:**
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 1,000 s of units] | Number of dealerships | Unit price [in 1,000 s of MUs] | Advertising budget [in 100,000 s of MUs] |
|---|---|---|---|---|
| 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 1,000 s of units)?
(b) What is 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: 1998: +2 %; 1999: +4 %; 2000: +1 %. What is the average growth in sales for this period?

**Exercise 9:**

(a) A used car market contains 200 vehicles across the following price categories:

| Car price (in €) | Number |
|---|---|
| Up to 2,500 | 2 |
| Between 2,500 and 5,000 | 8 |
| Between 5,000 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 10:**

Unions and employers sign a 4-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 11:**

A company has sold €30 m worth of goods over the last 3 years. In the first year they sold €8 m, in the second year €7 m, in the third year €15 m. What is the concentration of sales over the last 3 years? Use any indicator to solve the problem.

# 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.

## 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.

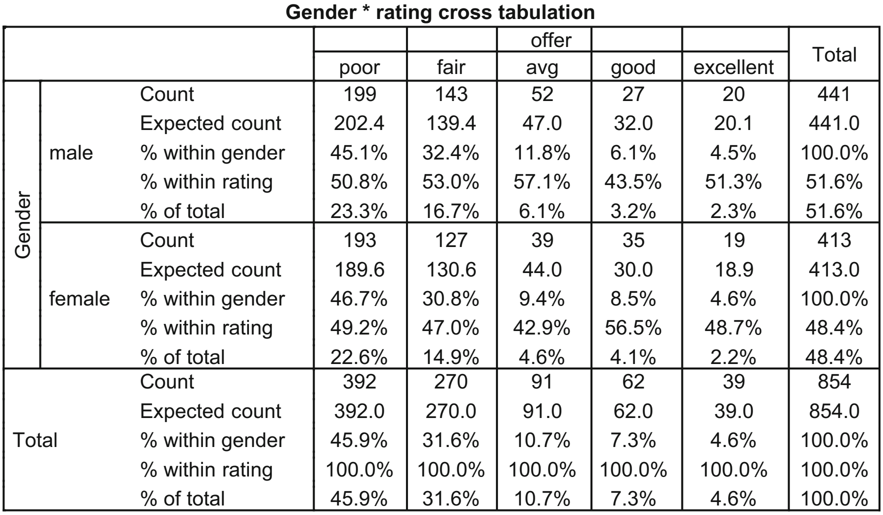
Chapter 4 Translated from the German original, Cleff, T. (2011). 4 Bivariate Zusammenhänge. In *Deskriptive Statistik und moderne Datenanalyse* (pp. 79–146) © Gabler Verlag, Springer Fachmedien Wiesbaden GmbH, 2011.

T. Cleff, *Exploratory Data Analysis in Business and Economics*,
DOI 10.1007/978-3-319-01517-0_4, © Springer International Publishing Switzerland 2014

**Table 4.1** Scale combinations and their measures of association

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



**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)

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 (51.9 %) of the 850 respondents are male and 409 (48.1 %) 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 850 respondents, 391 (46.0 %) find the selection poor, 266 (31.3 %) fair, etc. In the interior of the contingency table we find additional information. For instance, 198 respondents (23.3 %) were *male* and found the selection *poor*.

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 391 respondents, 193 are female, so the answer must be 49.4 %. The formal representation of these conditional relative frequencies is as follows:

$$f\left(\text{gender} = \textit{female} \,\middle|\, \text{selection} = \textit{poor}\right) = 193/391 = 49.4\% \tag{4.1}$$

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 409. This results in the following conditional frequency:

$$f\left(\text{selection rating} = \textit{good} \,\middle|\, \text{gender} = \textit{female}\right) = 35/409 = 8.6\% \tag{4.2}$$

The formula $f(x = 1 | y = 0)$ describes the relative frequency of the value 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 10 male respondents and all 12 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

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

| | | 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. 4.3** Contingency table (crosstab) (2nd)

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

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 0 (no association) to 1 (perfect association). To calculate this parameter, we must first 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

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

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

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

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:

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

From this we can easily determine the expected counts under independence: 6 persons, or 27.3 % of the 22 respondents, are female and make a purchase:

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

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)*:

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

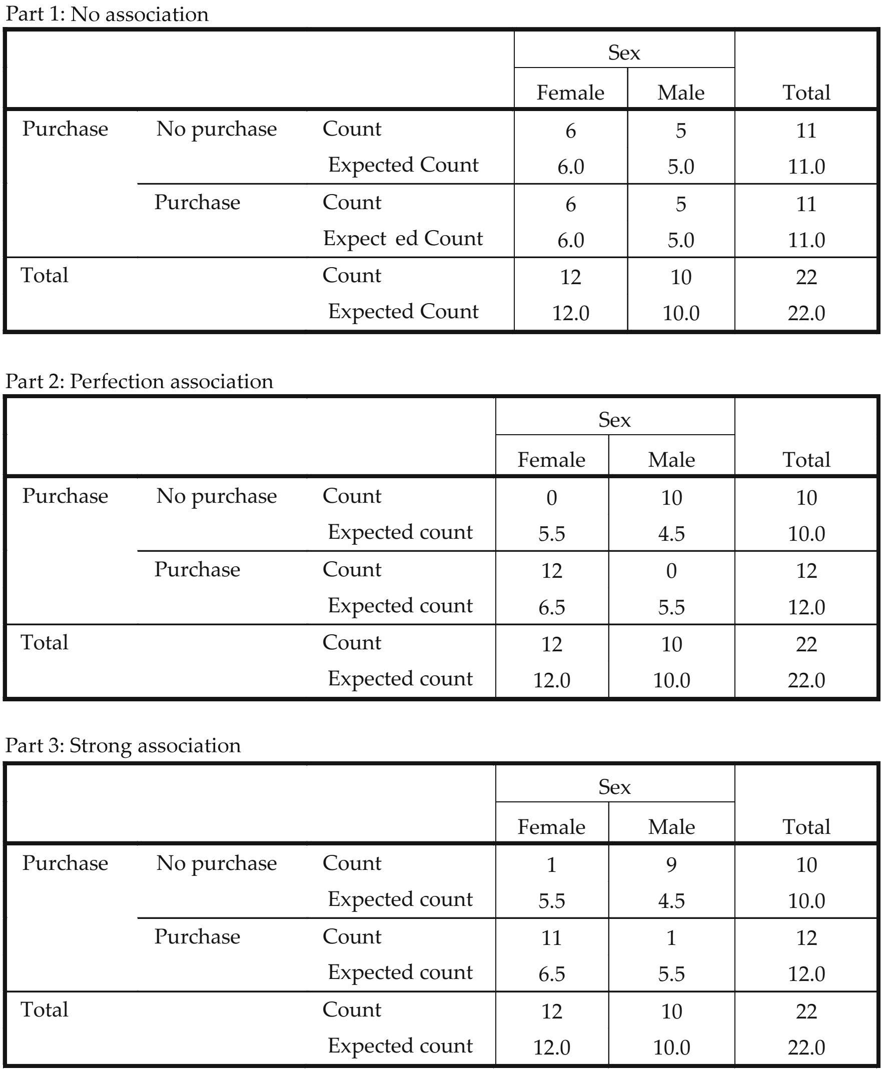
The sum of expected counts in each row or column must equal the absolute frequencies of the row or column. The idea is that a statistical association is not signified by different marginal frequencies but by different distributions of the sums of the marginal frequencies across columns or rows.

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 0, 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: $\frac{\left(n_{12}-n_{12}^{e}\right)^2}{n_{12}^{e}} = \frac{(6-6)^2}{6} = 0$. 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 ($\chi^2$-square):

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{\left(n_{ij} - n_{ij}^{e}\right)^2}{n_{ij}^{e}} = \frac{(6-6)^2}{6} + \frac{(6-6)^2}{6} + \frac{(5-5)^2}{5} + \frac{(5-5)^2}{5} = 0 \tag{4.8}$$

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 0, 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.



Part 1: No association

| | | | Sex | | Total |
|---|---|---|---|---|---|
| | | | Female | Male | |
| Purchase | No purchase | Count | 6 | 5 | 11 |
| | | Expected Count | 6.0 | 5.0 | 11.0 |
| | Purchase | Count | 6 | 5 | 11 |
| | | Expect ed 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 | | Total |
|---|---|---|---|---|---|
| | | | Female | Male | |
| 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 | | Total |
|---|---|---|---|---|---|
| | | | Female | Male | |
| 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

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

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:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{\left(n_{ij} - n_{ij}^e\right)^2}{n_{ij}^e} = \frac{(0-5.5)^2}{5.5} + \frac{(12-6.5)^2}{6.5} + \frac{(10-4.5)^2}{4.5} + \frac{(0-5.5)^2}{5.5} = 22 \tag{4.9}$$

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:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{\left(n_{ij} - n_{ij}^e\right)^2}{n_{ij}^e} = \frac{(1-5.5)^2}{5.5} + \frac{(11-6.5)^2}{6.5} + \frac{(9-4.5)^2}{4.5} + \frac{(1-5.5)^2}{5.5} = 14.7 \tag{4.10}$$

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 below 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 × k) or two columns (m × 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:

$$PHI = \phi = \sqrt{\frac{\chi^2}{n}} \tag{4.11}$$

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.

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 (Fig. 4.6).

[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.

Part 1: Perfect association in a 3x3 contingency table

| | | | Purchase | | | Total |
|---|---|---|---|---|---|---|
| | | | No Purchase | Frequent Purchase | Constant Purchase | |
| Customer | 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 | | | | |
|---|---|---|---|---|---|---|---|
| | | | No | Infrequent | Frequent | Constant | Total |
| Customer Group | | Count | 0 | 0 | 10 | 0 | 10 |
| | Customer | Expected Count | 4.0 | 2.0 | 2.0 | 2.0 | 10.0 |
| | B | Count | 0 | 10 | 0 | 0 | 10 |
| | Customer | Expected Count | 4.0 | 2.0 | 2.0 | 2.0 | 10.0 |
| | C | Count | 10 | 0 | 0 | 0 | 10 |
| | Customer | Expected Count | 4.0 | 2.0 | 2.0 | 2.0 | 10.0 |
| | D | Count | 10 | 0 | 0 | 0 | 10 |
| | Customer | Expected Count | 4.0 | 2.0 | 2.0 | 2.0 | 10.0 |
| | E | Count | 0 | 0 | 0 | 10 | 10 |
| | Customer | Expected Count | 4.0 | 2.0 | 2.0 | 2.0 | 10.0 |
| | Total | Count | 20 | 10 | 10 | 10 | 50 |
| | Customer | 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:

$$\varphi_{\max} = \sqrt{\min(\text{Number of rows}, \text{Number of columns}) - 1} \geq 1 \tag{4.12}$$

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:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \in [0; 1] \tag{4.13}$$

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:

$$C_{\max} = \sqrt{\frac{\min(k, m) - 1}{\min(k, m)}} = \sqrt{1 - \frac{1}{\min(k, m)}} \tag{4.14}$$

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

$$\mathrm{C_{korr}} = \sqrt{\frac{\chi^2}{\chi^2 + \mathrm{n}}} \cdot \sqrt{\frac{\min(\mathrm{k}, \mathrm{m})}{\min(\mathrm{k}, \mathrm{m}) - 1}} = \sqrt{\frac{\chi^2}{\chi^2 + \mathrm{n}}} \cdot \frac{1}{\sqrt{1 - \frac{1}{\min(\mathrm{k},\mathrm{m})}}} \in [0; 1] \tag{4.15}$$

### 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

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 | 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

between two nominal or ordinal variables. Its calculation is an extension of the phi coefficient:

$$\text{Cramer's V} = \sqrt{\frac{\chi^2}{n \cdot (\min(k, m) - 1)}} = \phi * \sqrt{\frac{1}{\min(k, m) - 1}} \in [0; 1] \qquad (4.16)$$

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:

$$1.\ \text{Cramer's V} = \sqrt{\frac{\chi^2}{n \cdot (\min(k,m) - 1)}} = \sqrt{\frac{22}{22 \cdot (2-1)}} = 1 \tag{4.17}$$

$$2.\ \text{Cramer's V} = \sqrt{\frac{\chi^2}{n \cdot (\min(k,m) - 1)}} = \sqrt{\frac{60}{30 \cdot (3-1)}} = 1 \tag{4.18}$$

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

Cramer's V ∈ [0.00; 0.10[ → no association
Cramer's V ∈ [0.10; 0.30[ → weak association
Cramer's V ∈ [0.30; 0.60[ → moderate association
Cramer's V ∈ [0.60; 1.00] → strong association

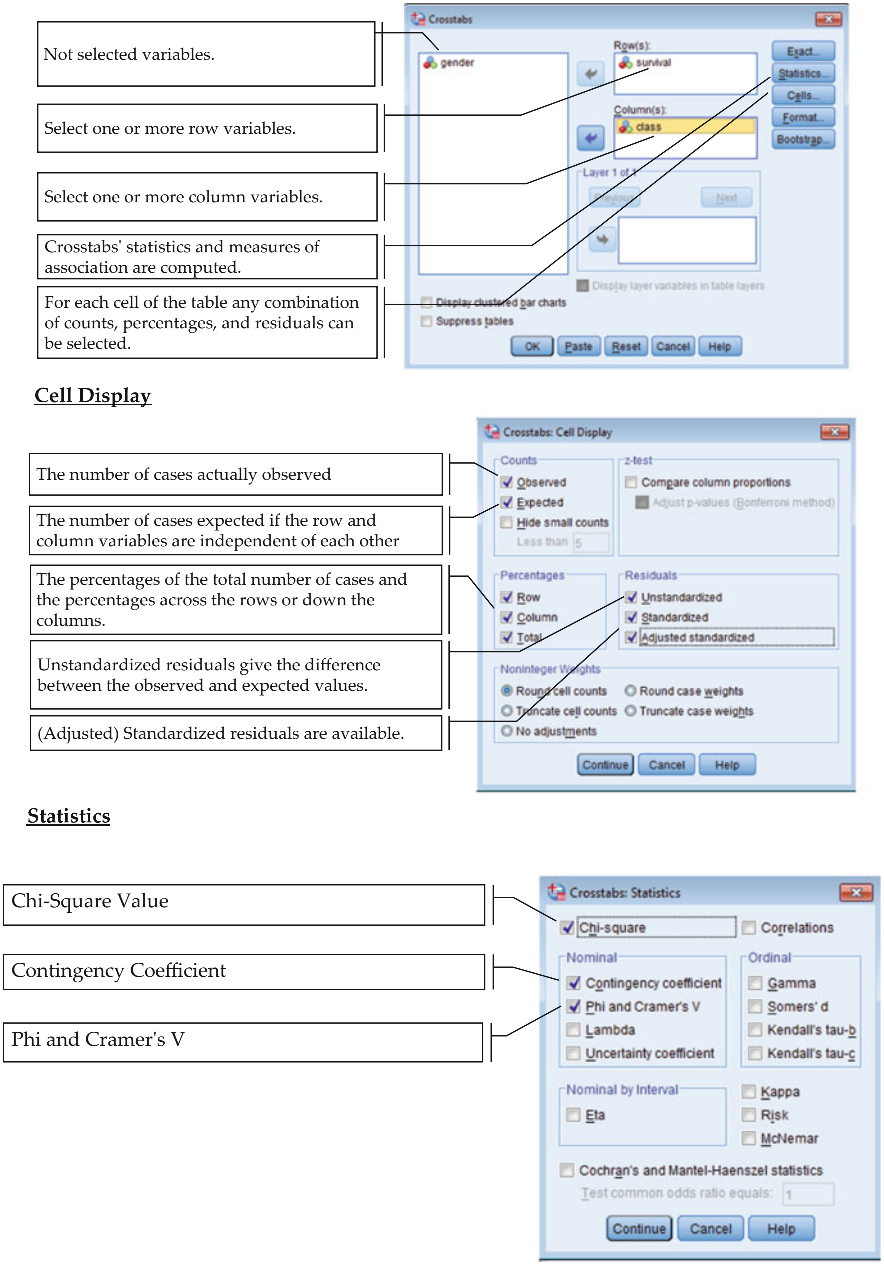
### 4.2.6 Nominal Associations with SPSS

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 2,201 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, 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 analyze. 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. 4.8: The cell display). The association measure can be selected under *statistics…*. Click *OK* to generate the tables in Figs. 4.9 and 4.10.

[2] The data in Titanic.sav (SPSS), Titanic.dta (Stata), and Titanic.xls (Excel) contain figures on the number of persons on board and the number of victims. The data is taken from the British Board of Trade Inquiry Report (1990), Report on the Loss of the Titanic' (S.S.), Gloucester (reprint).



Not selected variables.

Select one or more row variables.

Select one or more column variables.

Crosstabs' statistics and measures of association are computed.

For each cell of the table any combination of counts, percentages, and residuals can be selected.

Crosstabs

gender

Row(s): survival

Column(s): class

Layer 1 of 1

□ Display layer variables in table layers

□ Display clustered bar charts

□ Suppress tables

**Cell Display**

The number of cases actually observed

The number of cases expected if the row and column variables are independent of each other

The percentages of the total number of cases and the percentages across the rows or down the columns.

Unstandardized residuals give the difference between the observed and expected values.

(Adjusted) Standardized residuals are available.

Crosstabs: Cell Display

Counts
- ☑ Observed
- ☑ Expected
- □ Hide small counts
- Less than 5

z-test
- □ Compare column proportions
- □ Adjust p-values (Bonferroni method)

Percentages
- ☑ Row
- ☑ Column
- ☑ Total

Residuals
- ☑ Unstandardized
- ☑ Standardized
- ☑ Adjusted standardized

Noninteger Weights
- ● Round cell counts
- ○ Round case weights
- ○ Truncate cell counts
- ○ Truncate case weights
- ○ No adjustments

**Statistics**

Chi-Square Value

Contingency Coefficient

Phi and Cramer's V

Crosstabs: Statistics

- ☑ Chi-square
- □ Correlations

Nominal
- ☑ Contingency coefficient
- ☑ Phi and Cramer's V
- □ Lambda
- □ Uncertainty coefficient

Ordinal
- □ Gamma
- □ Somers' d
- □ Kendall's tau-b
- □ Kendall's tau-c

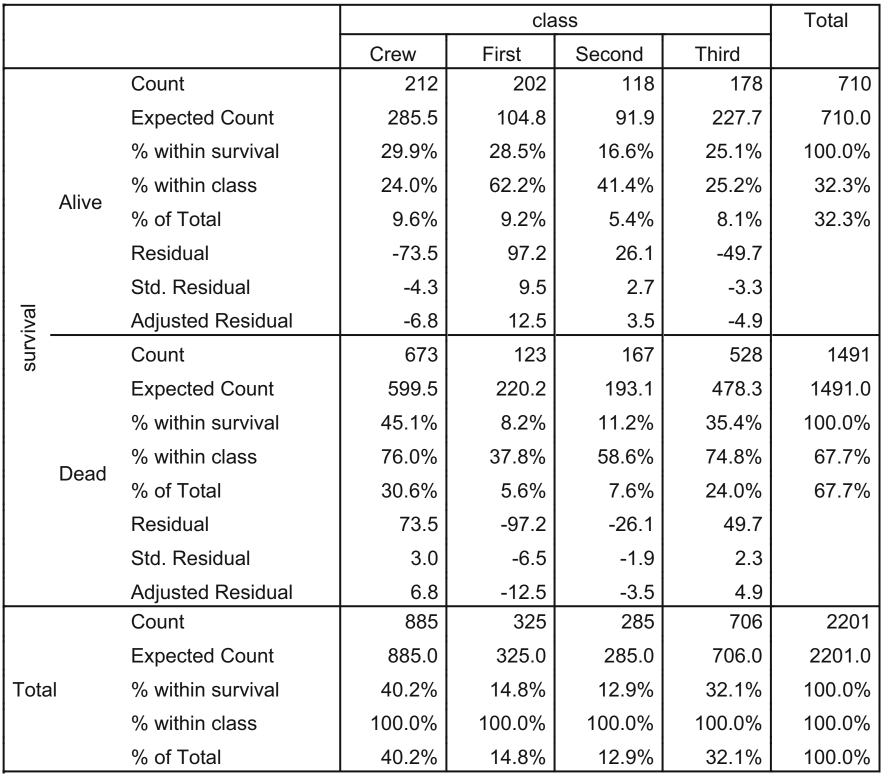
Nominal by Interval
- □ Eta

- □ Kappa
- □ Risk
- □ McNemar

□ Cochran's and Mantel-Haenszel statistics

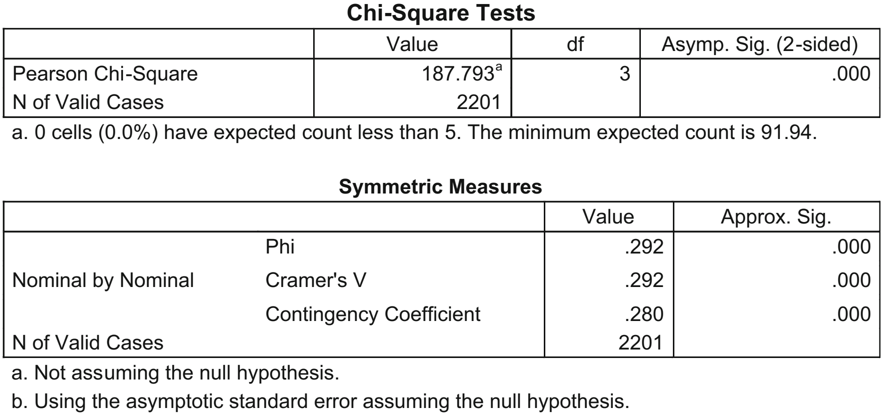
Test common odds ratio equals: 1

**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)



**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

Consider the contingency table in Fig. 4.9 below, 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 versus 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 25.3 % 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 $66.7\ \%\cdot 706 \approx 478$ would have died, not 528.

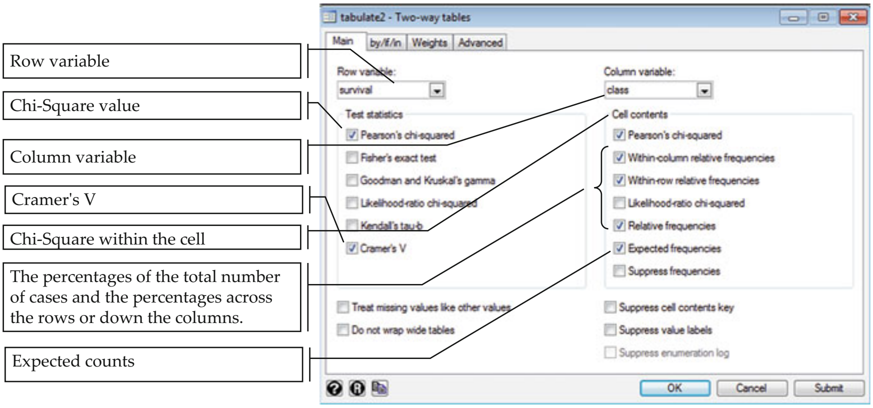
As we saw in the previous sections, the differences between the expected counts and the actual absolute frequency give us a general idea about the relationship between the variables. For closer analysis, however, the differences must be standardized by dividing them by the root of the expected counts (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.

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 × k or m × 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 – must be verified by comparing standardized residuals between actual and the expected frequencies.

### 4.2.7 Nominal Associations with Stata

To analyze nominal associations with Stata, follow a similar approach. Select *Statistics → Summaries, tables, and tests → Tables → Two-way tables with measures of association* to open the following window (Fig. 4.11):

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



tabulate2 - Two-way tables

Main | by/if/in | Weights | Advanced

Row variable: survival

Column variable: class

Test statistics
- ☑ Pearson's chi-squared
- □ Fisher's exact test
- □ Goodman and Kruskal's gamma
- □ Likelihood-ratio chi-squared
- □ Kendall's tau-b
- ☑ Cramer's V

Cell contents
- ☑ Pearson's chi-squared
- ☑ Within-column relative frequencies
- ☑ Within-row relative frequencies
- □ Likelihood-ratio chi-squared
- ☑ Relative frequencies
- ☑ Expected frequencies
- □ Suppress frequencies

□ Treat missing values like other values

□ Do not wrap wide tables

□ Suppress cell contents key

□ Suppress value labels

□ Suppress enumeration log

Row variable
Chi-Square value
Column variable
Cramer's V
Chi-Square within the cell
The percentages of the total number of cases and the percentages across the rows or down the columns.
Expected counts
OK
Cancel
Submit

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

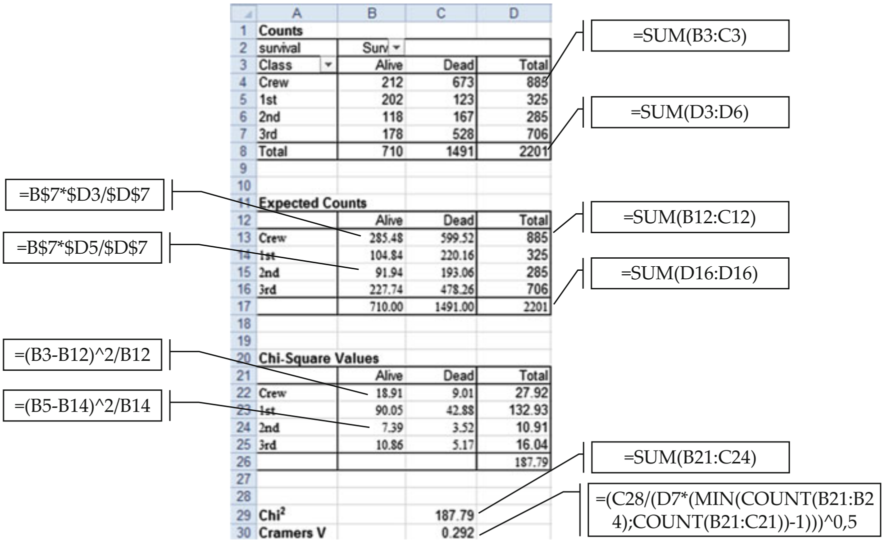
the contingency table. Click on *OK* or *Submit* to perform the Stata calculation.[3] The results can now be interpreted as in the SPSS example.

### 4.2.8 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 preprogrammed functions for contingency tables.

Here is a brief sketch of how to perform these functions in Excel if needed. First select the (conditional) actual frequencies for each cell as in Fig. 4.12. 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 ∑ *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 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 programme 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. 4.12). In a new table we

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



|  | 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 |  |

=SUM(B3:C3)

=SUM(D3:D6)

=B$7*$D3/$D$7

=B$7*$D5/$D$7

=SUM(B12:C12)

=SUM(D16:D16)

=(B3-B12)^2/B12

=(B5-B14)^2/B14

=SUM(B21:C24)

=(C28/(D7*(MIN(COUNT(B21:B24);COUNT(B21:C21))-1)))^0,5

**Fig. 4.12** Crosstabs and nominal measures of association with Excel (Titanic)

can calculate the individual chi-squares for each cell (see the third table in Fig. 4.12). The sum of these chi-squares equals the total chi-square value. From this, we can calculate Cramer’s V. The formulas in Fig. 4.12 provide an example.

### 4.2.9 Chapter Exercises

**Exercise 12:**

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|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<br>(y = 1) | Moderate<br>(y = 2) | Low<br>(y = 3) | Sum<br>(X) |
|---|---|---|---|---|---|
| With Music<br>(x = 1) | Count<br>(Expected Count) | 130 (____) | 30 (____) | 50 (____) | |
| Without Music<br>(x = 2) | Count<br>(Expected Count) | 40 (____) | 20 (____) | 130 (____) | |
| Sum (Y) | Count | | | | |

(c) Determine the chi-square value.
(d) Calculate Cramer's V.

**Exercise 13:**
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 $f(x = 4|y = 2) = 1/18$ and the absolute frequency is $h(x = 2|y = 3) = 30$. 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 14:**
A company measures customer satisfaction in three regions, producing the following crosstab:

| | | | Region | | | |
|---|---|---|---|---|---|---|
| | | | region 1 | region 2 | region 3 | Total |
| 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 % |
| | | % 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 | .000 |
| Likelihood Ratio | 48.519 | 4 | .000 |
| Linear-by-Linear Association | .569 | 1 | .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 | .818 | .000 |
| | Cramer's V | .578 | .000 |
| | Contingency Coefficient | .633 | .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.

## 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

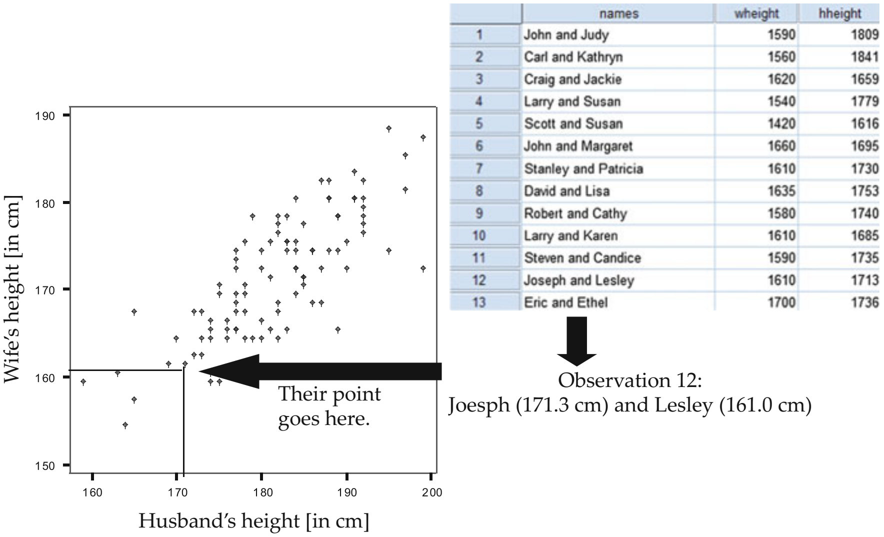
Officials performing civil marriage ceremonies frequently observe that brides and grooms tend to be of similar height. Taller men generally marry taller women and vice versa. One official decides to verify this impression by recording the heights of 100 couples. How can he tell whether there’s an actual 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 1: “Peter and Petra”) is represented by the coordinate with the values 171 for the groom and 161 for the bride. Plotting all the observed pairs results in a cloud of points, or scatterplot (see Fig. 4.13).

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 4.14 provides some examples.

1. *The direction of the relationship*. Relationships can be positive, negative, or nonexistent. 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.



| | 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 |

**Fig. 4.13** The scatterplot

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.

There are many software tools that make creating scatterplots easy.[4] But their interpretation requires care. Figure 4.15 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 first diagram in Fig. 4.15 the y-axis is scaled between 140 and 200 cm and the x-axis between 10 and 70. In the second diagram height is scaled between 0 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 first 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 first 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.

[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.

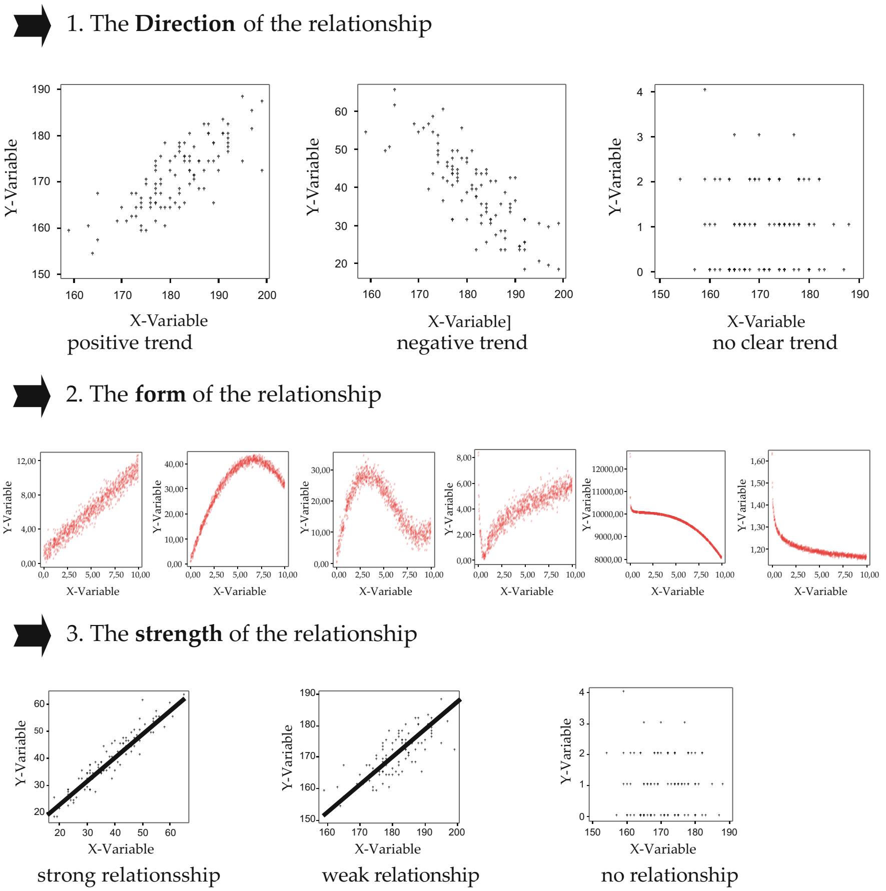


**Fig. 4.14** Aspects of association expressed by the scatterplot

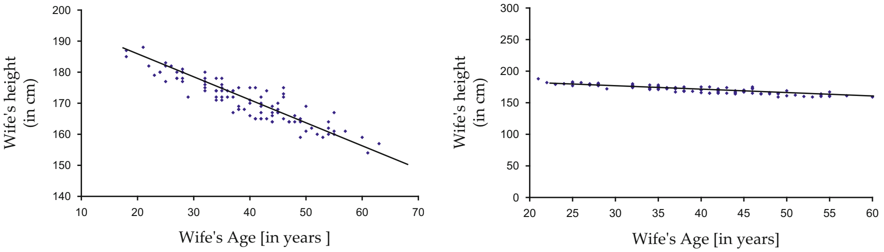


**Fig. 4.15** Different representations of the same data (3)....

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 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 0, 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. 4.16.

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 average height. The value pair of the bivariate centroid then becomes the centre of a new coordinate system with four quadrants (see Fig. 4.17).

All points in quadrant 1 involve marriages between men and women of above-average heights. When the values of quadrant 1 are entered into the equation $(x_i - \bar{x}) \cdot (y_i - \bar{y})$, the results are always positive. All points in quadrant 3 involve marriages between men and women of below-average heights. Here too, values fed into the equation$(x_i - \bar{x}) \cdot (y_i - \bar{y})$ produce positive results, as the product of two negative values is always positive.

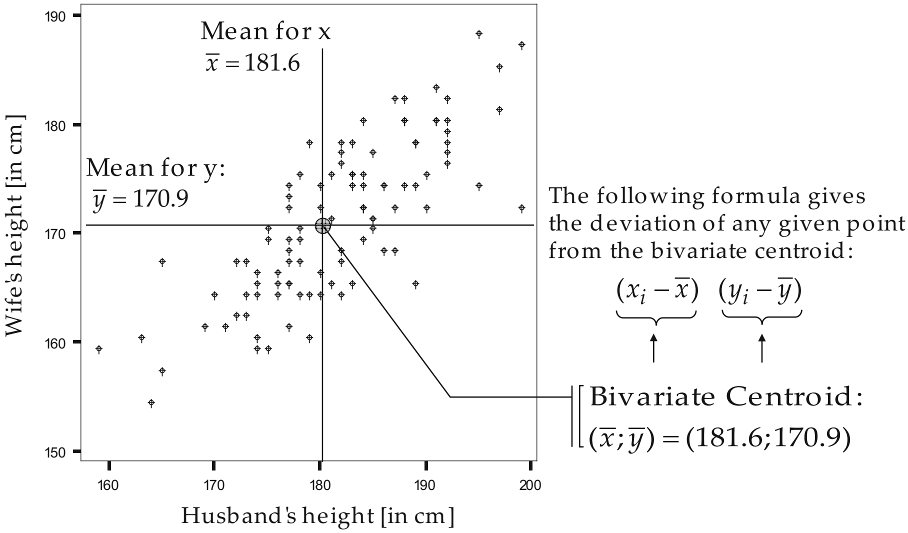


**Fig. 4.16** Relationship of heights in married couples

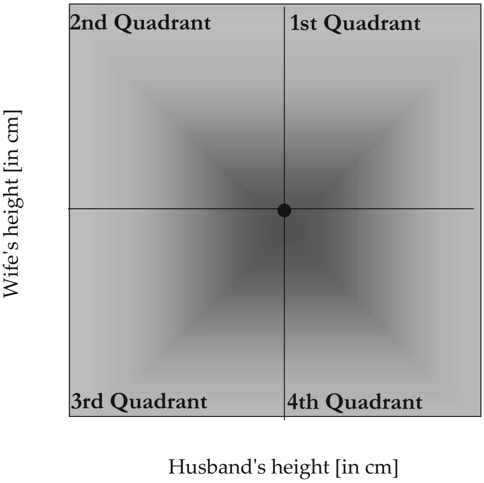


**Fig. 4.17** Four-quadrant system

All the data points in quadrant 1 and 3 are located at positive intervals to the bivariate centroid, with intervals being measured by the product $(x_i - \bar{x}) \cdot (y_i - \bar{y})$. This makes sense: the cloud of points formed by the data has a positive slope.

Quadrant 2 contains data from taller-than-average women who married shorter-than-average men, while quadrant 4 contains data from shorter-than-average women who married taller-than-average men. For these observations, the product of $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ 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 1 and 3 compared with the sum of the negative intervals in quadrants 2 and 4. The larger the sum of the intervals in quadrants 1 and 3, 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 1 and 3 are similar to those in quadrants 2 and 4, 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 2 and 4. 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:

$$\mathrm{cov}(x;y) = S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}x_i y_i - \bar{x}\,\bar{y} \tag{4.19}$$

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 $\chi^2$ 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 $\chi^2$ 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.

$$r = \frac{S_{xy}}{S_x S_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right) \cdot \left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}} \text{ with} - 1 \leq r \leq +1 \tag{4.20}$$

The values of Pearson's correlation coefficient always lie between r = (−1) and r = (+1). The closer the correlation coefficient is to 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 downwards 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 ≤ |r| < 0.8 moderate linear association
|r| ≥ 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 4.18 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.
- Figure 4.19 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. 4.20. 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

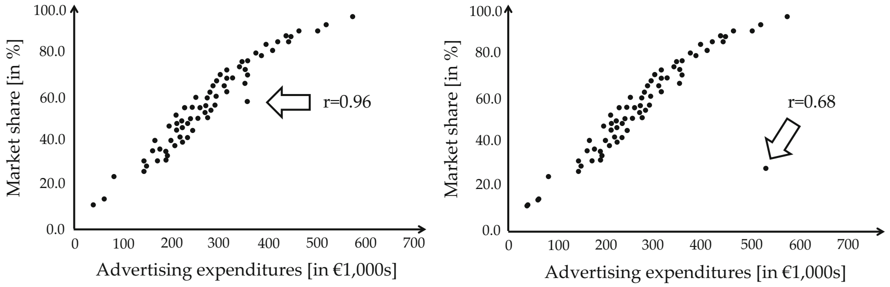


**Fig. 4.18** Pearson's correlation coefficient with outliers

**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. 4.19** Wine bottle design survey

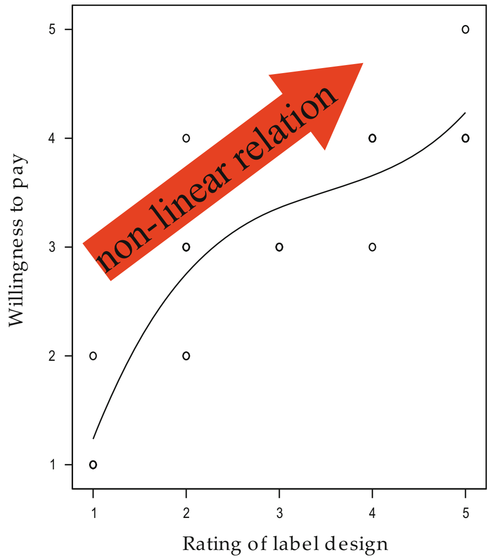


**Fig. 4.20** Nonlinear relationship between two variables

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.

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. 4.20. 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 0, the less the value pairs share a perfect monotonic relationship.

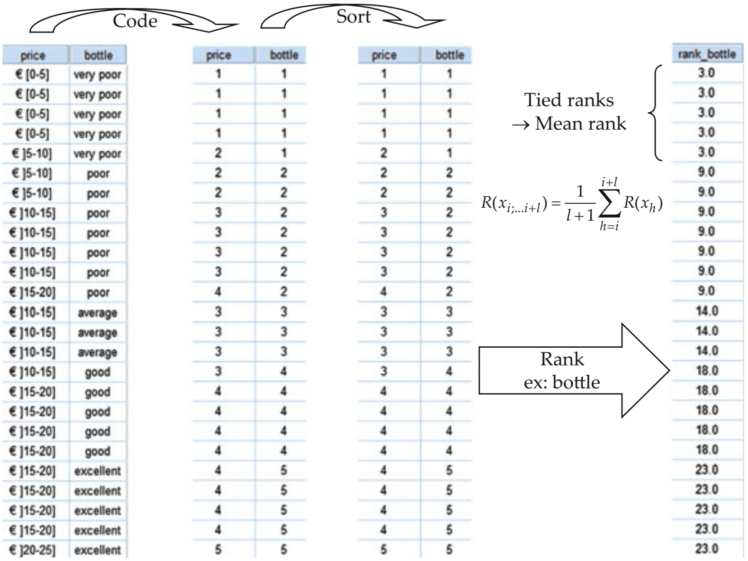
The basic idea of Spearman's rho is to create a ranking order for each dataset and then to measure the difference between the ranks of each observation. Spearman's rho treats the ranking orders as cardinal scales by assuming that the distances between them are equidistant. From a theoretical perspective, 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. 4.19. 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 follows:

First the dataset is sorted by the value size of one of both variables. In Fig. 4.21 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 1 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



$$R(x_{i;\dots i+l})=\frac{1}{l+1}\sum_{h=i}^{i+l}R(x_h)$$

| price | bottle | price | bottle | price | bottle | rank_bottle |
|---|---|---|---|---|---|---|
| € [0-5] | very poor | 1 | 1 | 1 | 1 | 3.0 |
| € [0-5] | very poor | 1 | 1 | 1 | 1 | 3.0 |
| € [0-5] | very poor | 1 | 1 | 1 | 1 | 3.0 |
| € [0-5] | very poor | 1 | 1 | 1 | 1 | 3.0 |
| € ]5-10] | very poor | 2 | 1 | 2 | 1 | 3.0 |
| € ]5-10] | poor | 2 | 2 | 2 | 2 | 9.0 |
| € ]5-10] | poor | 2 | 2 | 2 | 2 | 9.0 |
| € ]10-15] | poor | 3 | 2 | 3 | 2 | 9.0 |
| € ]10-15] | poor | 3 | 2 | 3 | 2 | 9.0 |
| € ]10-15] | poor | 3 | 2 | 3 | 2 | 9.0 |
| € ]10-15] | poor | 3 | 2 | 3 | 2 | 9.0 |
| € ]15-20] | poor | 4 | 2 | 4 | 2 | 9.0 |
| € ]10-15] | average | 3 | 3 | 3 | 3 | 14.0 |
| € ]10-15] | average | 3 | 3 | 3 | 3 | 14.0 |
| € ]10-15] | average | 3 | 3 | 3 | 3 | 14.0 |
| € ]10-15] | good | 3 | 4 | 3 | 4 | 18.0 |
| € ]15-20] | good | 4 | 4 | 4 | 4 | 18.0 |
| € ]15-20] | good | 4 | 4 | 4 | 4 | 18.0 |
| € ]15-20] | good | 4 | 4 | 4 | 4 | 18.0 |
| € ]15-20] | good | 4 | 4 | 4 | 4 | 18.0 |
| € ]15-20] | excellent | 4 | 5 | 4 | 5 | 23.0 |
| € ]15-20] | excellent | 4 | 5 | 4 | 5 | 23.0 |
| € ]15-20] | excellent | 4 | 5 | 4 | 5 | 23.0 |
| € ]15-20] | excellent | 4 | 5 | 4 | 5 | 23.0 |
| € ]20-25] | excellent | 5 | 5 | 5 | 5 | 23.0 |

**Fig. 4.21** 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):

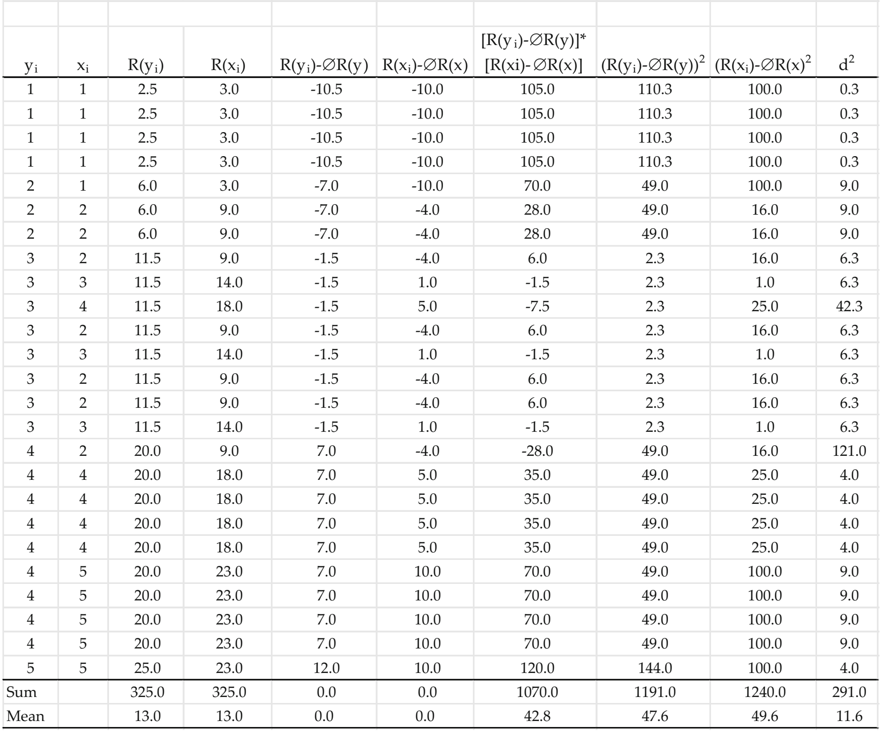
$$1/3\cdot 1+1/3\cdot 2+1/3\cdot 3=1/3\cdot(1+2+3)=2 \qquad (4.21)$$

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) = 36.

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 3, as $1/5\cdot(1+2+3+4+5)=3$. 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 $1/7\cdot(6+7+8+9+10+11+12)=9$.

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: $1/5\cdot(21+22+23+24+25)=23$



| $y_i$ | $x_i$ | $R(y_i)$ | $R(x_i)$ | $R(y_i)$-∅R(y) | $R(x_i)$-∅R(x) | [R(y i)-∅R(y)]*<br>[R(xi)- ∅R(x)] | $(R(y_i)$-∅R(y)$)^2$ | $(R(x_i)$-∅R(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. 4.22** Rankings from the wine bottle design survey

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. 4.22, 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 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 $\overline{R(y)}$:

$$\rho = \frac{S_{xy}}{S_x S_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i)-\overline{R(x)}\right)\cdot\left(R(y_i)-\overline{R(y)}\right)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i)-\overline{R(x)}\right)^2\right)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(y_i)-\overline{R(y)}\right)^2\right)}} \tag{4.22}$$

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

- $\overline{\mathrm{R(x)}} = \overline{\mathrm{R(y)}} = \frac{1}{25}(1+2+3+\cdots+25) = \frac{1}{\mathrm{n}}\cdot\frac{\mathrm{n}\cdot(\mathrm{n}+1)}{2} = \frac{1}{25}\cdot\frac{25\cdot(25+1)}{2} = 13$
- $\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i)-\overline{R(x)}\right)^2 = \frac{1}{25}\left((3-13)^2+\cdots+(23-13)^2\right) = \frac{1{,}240}{25} = 49.6$

- $\frac{1}{n}\sum_{i=1}^{n}\left(R(y_i)-\overline{R(y)}\right)^2=\frac{1}{25}\left((2.5-13)^2+\cdots+(25-13)^2\right)=\frac{1{,}191}{25}=47.6$
- $\sum_{i=1}^{n}\left(R(x_i)-\overline{R(x)})(R(y_i)-\overline{R(y)}\right)=((3-13)(2.5-13))+\cdots$ $+((23-13)(20-13))=42.8$

These, in turn, produce:

$$\rho=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i)-\overline{R(x)})\cdot(R(y_i)-\overline{R(y)}\right)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i)-\overline{R(x)}\right)^2\right)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(y_i)-\overline{R(y)}\right)^2\right)}}$$
$$=\frac{42.8}{\sqrt{49.6\cdot 47.6}}=0.880 \tag{4.23}$$

Calculating this formula by hand is time consuming. As computers are widely available today, a short-hand version of the formula is frequently used:

$$\rho=1-\frac{6\cdot\sum_{i=1}^{n}d_i^2}{n\cdot(n^2-1)}\text{with } \mathrm{d_i}=(R(x_i)-R(y_i)) \tag{4.24}$$

We first calculate the difference between ranks for each value pair. In our wine bottle survey, the first row contains $d_1=(2.5\text{–}3.0)=(-0.5)$. We then square and sum all the differences (see column $d^2$ in Fig. 4.22). This produces the following:

$$\rho=1-\frac{6\cdot\sum_{i=1}^{n}d_i^2}{n\cdot(n^2-1)}=1-\frac{6\cdot 291}{25\cdot\left(25^2-1\right)}=\frac{1,746}{15,600}=0.888 \tag{4.25}$$

There is a slight deviation between this result (ρ = 0.888) and that of the full-length version of the formula (ρ = 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).

$$\rho_{korr} = \frac{2 \cdot \left(\frac{N^3-N}{12} - N\right) - T - U - \sum_{i=1}^{n} d_i^2}{2 \cdot \sqrt{\left(\frac{N^3-N}{12} - T\right) \cdot \left(\frac{N^3-N}{12} - U\right)}} \text{ with} \tag{4.26}$$

- T as the length of b tied ranks among x variables: $T = \frac{\sum_{i=1}^{b} \left(t_i^3 - t_i\right)}{12}$, 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: $U = \frac{\sum_{i=1}^{c} \left(u_i^3 - u_i\right)}{12}$, 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, new 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.

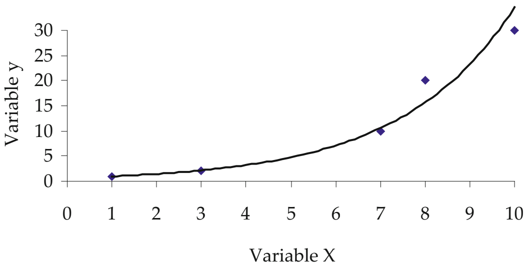
Two short examples serve to illustrate the basic idea of Kendall's τ. Let us assume a perfect positive monotonic relationship between variables x and y, as shown in Fig. 4.23.

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. 4.23 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.

Variable x: 7 1 10 3 8
Variable y: 10 1 30 2 20

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

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



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 2 3 4 5

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

| | | | |
|---|---|---|---|
| $R(y_1)$- $R(y_2)$ ⇨(+)<br>$R(y_1)$- $R(y_3)$ ⇨(+)<br>$R(y_1)$- $R(y_4)$ ⇨(+)<br>$R(y_1)$- $R(y_5)$ ⇨(+) | $R(y_2)$- $R(y_3)$ ⇨(+)<br>$R(y_2)$- $R(y_4)$ ⇨(+)<br>$R(y_2)$- $R(y_5)$ ⇨(+) | $R(y_3)$- $R(y_4)$ ⇨(+)<br>$R(y_3)$- $R(y_5)$ ⇨(+) | $R(y_4)$- $R(y_5)$ ⇨(+) |

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

**Fig. 4.23** Kendall's τ and a perfect positive 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.

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 – in our example there are 10; $\frac{n\cdot(n-1)}{2}=\frac{5\cdot(5-1)}{2}=10$ – we determine the surplus of concordant pairs (P) to discordant pairs (I). In our example the surplus is 10: (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 $\tau_a$:

$$\tau_a=\frac{No.of\ concordant\ pairs-No.of\ discordant\ pairs}{n\cdot(n-1)/2}=\frac{P-I}{n\cdot(n-1)/2}$$
$$=\frac{10-0}{10}=1 \qquad (4.27)$$

If the association was perfectly negative and monotonic, there would have been 10 discordant pairs and no concordant pairs. For Kendall's $\tau_a$ we arrive at the following:

$$\tau_a = \frac{P - I}{n \cdot (n-1)/2} = \frac{0 - 10}{10} = -1 \tag{4.28}$$

As with Spearman's rank correlation coefficient, the values of Kendall's $\tau_a$ lie between $\tau_a = (-1)$ and $\tau_a = (+1)$. 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 $\tau_a = (+1)$. 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 $\tau_a = (-1)$. The more the value of the coefficient approaches $\tau_a = 0$, 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 $\tau_a$ assumes a value of $\tau_a = 0$, as shown in Fig. 4.24:

$$\tau_a = \frac{P - I}{n \cdot (n-1)/2} = \frac{5 - 5}{10} = 0 \tag{4.29}$$

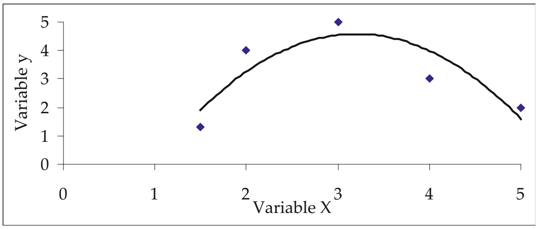
The simple formula Kendall's $\tau_a$ assumes that no tied ranks are present. If tied ranks are present, the corrected formula *Kendall's* $\tau_b$ should be used:

$$\tau_b = \frac{P - I}{\sqrt{\left(\frac{n \cdot (n-1)}{2} - T\right)\left(\frac{n \cdot (n-1)}{2} - U\right)}} \text{ where} \tag{4.30}$$

- 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.
- 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.

The more tied ranks that are present in a dataset, the smaller the value of Kendall's $\tau_a$ compared with Kendall's $\tau_b$. The practical application of this very complicated formula can be illustrated using our wine bottle survey (see Fig. 4.25).

| | | | | | |
|---|---|---|---|---|---|
| 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. 4.24** Kendall's τ for a non-existent monotonic association

| i | y | x | R(y) | R(x) | concordant pairs | discordant pairs |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 2.5 | 3.0 | 0 | 0 |
| 2 | 1 | 1 | 2.5 | 3.0 | 0 | 0 |
| 3 | 1 | 1 | 2.5 | 3.0 | 0 | 0 |
| 4 | 1 | 1 | 2.5 | 3.0 | 0 | 0 |
| 5 | 2 | 1 | 6.0 | 3.0 | 0 | 0 |
| 6 | 2 | 2 | 6.0 | 9.0 | 0 | 0 |
| 7 | 2 | 2 | 6.0 | 9.0 | 0 | 0 |
| 8 | 3 | 2 | 11.5 | 9.0 | 0 | 0 |
| 9 | 3 | 3 | 11.5 | 14.0 | 0 | 0 |
| 10 | 3 | 4 | 11.5 | 18.0 | 0 | 0 |
| 11 | 3 | 2 | 11.5 | 9.0 | 0 | 0 |
| 12 | 3 | 3 | 11.5 | 14.0 | 0 | 0 |
| 13 | 3 | 2 | 11.5 | 9.0 | 0 | 0 |
| 14 | 3 | 2 | 11.5 | 9.0 | 0 | 0 |
| 15 | 3 | 3 | 11.5 | 14.0 | 0 | 0 |
| 16 | 4 | 2 | 20.0 | 9.0 | 0 | 0 |
| 17 | 4 | 4 | 20.0 | 18.0 | 0 | 0 |
| 18 | 4 | 4 | 20.0 | 18.0 | 0 | 0 |
| 19 | 4 | 4 | 20.0 | 18.0 | 0 | 0 |
| 20 | 4 | 4 | 20.0 | 18.0 | 0 | 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 $\tau_b$ 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. 4.25** Kendall's τ for tied ranks

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 – 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 3. 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 5 to 7 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 $\tau_b$ assumes that tied ranks in the anchor column can lead concordant pairs and discordant pairs in the reference column to be overlooked. For observation 5 there are only 18 concordant pairs – all observation values between 8 and 25. We proceed the same way with observation 8. For observations 8 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 to 25. For observation 9 there are 9 concordant pairs and 1 discordant pair.

This results in 197 concordant pairs and only 4 discordant pairs, so that:

$$\tau_b = \frac{197-4}{\sqrt{\left(\frac{25\cdot(25-1)}{2}-73\right)\left(\frac{25\cdot(n-1)}{2}-54\right)}} = 0.817 \tag{4.31}$$

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 $\tau_b$ 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. 4.26. 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:

$$\begin{aligned} \mathrm{P} = {} & 4\cdot(2+4+1+3+1+4+4+1)+1\cdot(4+1+3+1+4+4+1) \\ & +2\cdot(3+1+4+4+1)+4\cdot(4+4+1)+3\cdot(4+4+1) \\ & +1\cdot(4+1)+1\cdot1+4\cdot1 = 197 \end{aligned} \tag{4.32}$$

R(y)

|  |  | 2.5 | 6.0 | 11.5 | 20.0 | 25.0 | Total |
|---|---|---|---|---|---|---|---|
| R(x) | 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. 4.26** Deriving Kendall's $\tau_b$ from a contingency table

For discordant pairs, the reverse applies:

$$\mathrm{I} = 4 \cdot 0 + 1 \cdot 0 + 2 \cdot 0 + 4 \cdot 0 + 3 \cdot 0 + 1 \cdot 0 + 1 \cdot (3 + 1) + 4 \cdot 0 = 4 \qquad (4.33)$$

Kendall's $\tau_b$ can now be derived from the above formula. Kendall's $\tau$ can also be applied to contingency tables. The scale of both variables must be ordinal; otherwise, the relationships between larger and smaller values cannot be interpreted. If Kendall's $\tau_b$ is derived from a *non-square contingency table*, the values $\tau_b = (+1)$ and $\sum_{i=1}^{n} y_i = -309$ can never be reached, even if the association is perfectly monotonic. Instead we must calculate *Kendall's $\tau_c$:*

$$\tau_c = \frac{2 \cdot \min[\#rows; \#columns] \cdot (P - I)}{(\min[\#rows; \#columns] - 1) \cdot n^2} \qquad (4.34)$$

The example from Fig. 4.26 yields the following calculation:

$$\tau_c = \frac{2 \cdot \min[5; 5] \cdot (197 - 4)}{(\min[5; 5] - 1) \cdot 25^2} = \frac{2 \cdot 5 \cdot (193)}{(5 - 1) \cdot 25^2} = 0.772 \qquad (4.35)$$

## 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:

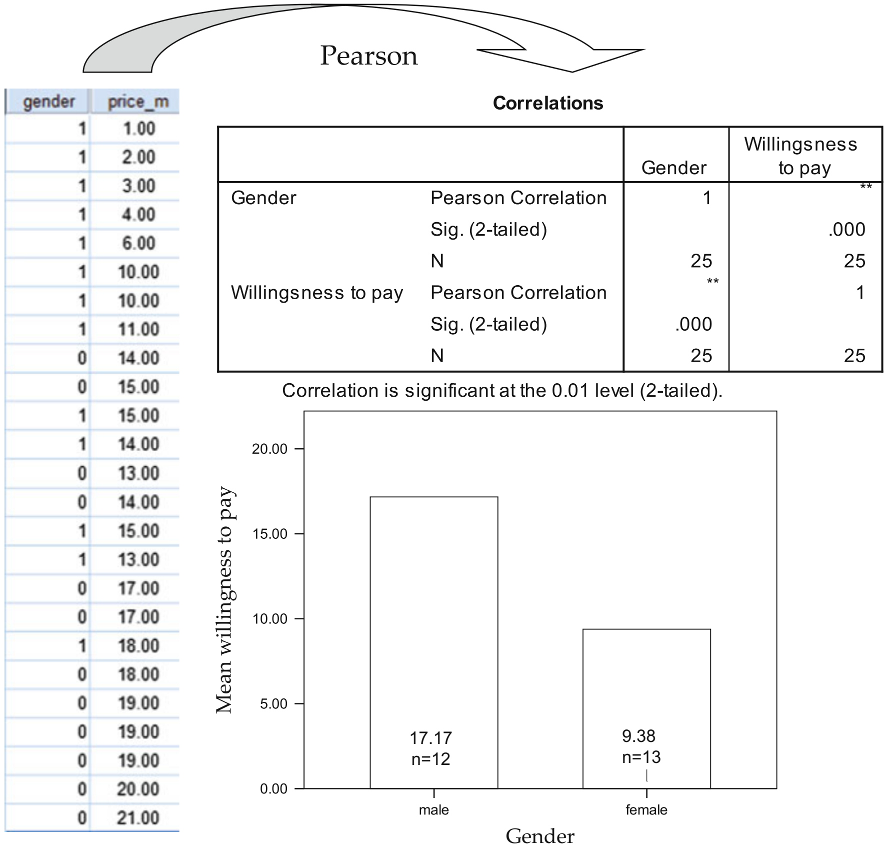
- In practice, statisticians usually apply *statistical tests* (*t*-test or variance analysis) to assess differences between nominal groups with regard to metric variables. These tests belong to inductive statistics and require knowledge of probability theory, which lies outside the scope of this book.
- 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 values) 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*) – 0 for *male* and 1 for *female*. The results are shown in Fig. 4.27.

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. 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 0 or 1 can be regarded metrically. Applied to our case: If the value of the variable *gender* is 1, the more female the respondent is. If the value of the variable *gender* is 0, 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 0 (male) have higher values with the metric variable (willingness-to-pay) than respondents coded as 1 (female). By contrast, a point-biserial correlation of $r_{pb} = (+1)$ means that all respondents coded as 0 (male) have lower values with metric variables (willingness-to-pay) than respondents coded as 1 (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 $r_{pb} = 0$.



| 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**

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

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

**Fig. 4.27** Point-biserial correlation

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

$$r_{pb} = \frac{\bar{y}_1 - \bar{y}_0}{S_y} \sqrt{\frac{n_0 \cdot n_1}{n^2}}, \text{ where} \tag{4.36}$$

- $n_0$: 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:

$$r_{pb} = \frac{\overline{y}_1 - \overline{y}_0}{S_y} \sqrt{\frac{n_0 \cdot n_1}{n^2}} = \frac{9.38 - 17.17}{5.8} \sqrt{\frac{12 \cdot 13}{25^2}} = (-0.67) \qquad (4.37)$$

The negative point-biserial correlation indicates that the respondents whose dichotomous variable value is 1 (female) show a lower willingness-to-pay than the respondents whose dichotomous variable value is 0 (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 values for the ordinal variable is not too large. Statistical tests (Mann–Whitney or Kruskal-Wallis) 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. But these procedures belong to inductive statistics, which lie outside the scope of this book.

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):

$$r_{bisR} = \frac{2}{n} \cdot \left( \overline{R(y_1)} - \overline{R(y_0)} \right), \text{where :} \qquad (4.38)$$

- $n$: 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. 4.28.

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

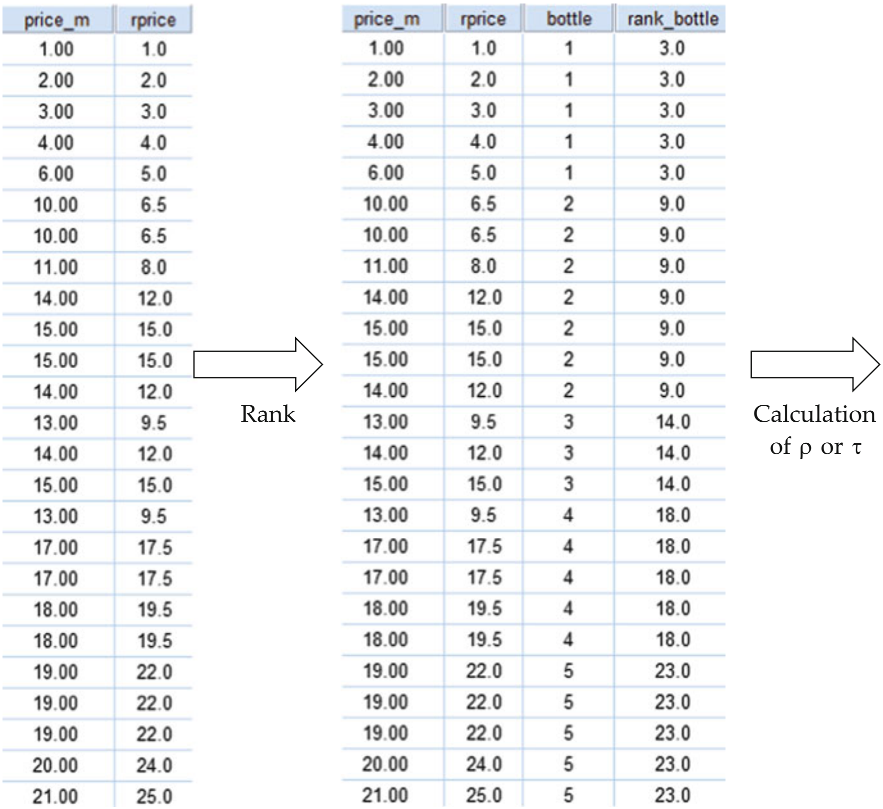
## 4.6 Calculating Correlation with a Computer

When using 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 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 $\tau_b$, 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 $\tau_b$ or Spearman's correlation when determining the monotonic 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*. 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. 4.29. Pearson's correlation has the value $r = 0.789$, Kendall's $\tau_b$ has the value $\tau_b = 0.603$, and Spearman's correlation has the value $\rho = 0.783$.



| price_m | rprice |
|---|---|
| 1.00 | 1.0 |
| 2.00 | 2.0 |
| 3.00 | 3.0 |
| 4.00 | 4.0 |
| 6.00 | 5.0 |
| 10.00 | 6.5 |
| 10.00 | 6.5 |
| 11.00 | 8.0 |
| 14.00 | 12.0 |
| 15.00 | 15.0 |
| 15.00 | 15.0 |
| 14.00 | 12.0 |
| 13.00 | 9.5 |
| 14.00 | 12.0 |
| 15.00 | 15.0 |
| 13.00 | 9.5 |
| 17.00 | 17.5 |
| 17.00 | 17.5 |
| 18.00 | 19.5 |
| 18.00 | 19.5 |
| 19.00 | 22.0 |
| 19.00 | 22.0 |
| 19.00 | 22.0 |
| 20.00 | 24.0 |
| 21.00 | 25.0 |

| price_m | rprice | bottle | rank_bottle |
|---|---|---|---|
| 1.00 | 1.0 | 1 | 3.0 |
| 2.00 | 2.0 | 1 | 3.0 |
| 3.00 | 3.0 | 1 | 3.0 |
| 4.00 | 4.0 | 1 | 3.0 |
| 6.00 | 5.0 | 1 | 3.0 |
| 10.00 | 6.5 | 2 | 9.0 |
| 10.00 | 6.5 | 2 | 9.0 |
| 11.00 | 8.0 | 2 | 9.0 |
| 14.00 | 12.0 | 2 | 9.0 |
| 15.00 | 15.0 | 2 | 9.0 |
| 15.00 | 15.0 | 2 | 9.0 |
| 14.00 | 12.0 | 2 | 9.0 |
| 13.00 | 9.5 | 3 | 14.0 |
| 14.00 | 12.0 | 3 | 14.0 |
| 15.00 | 15.0 | 3 | 14.0 |
| 13.00 | 9.5 | 4 | 18.0 |
| 17.00 | 17.5 | 4 | 18.0 |
| 17.00 | 17.5 | 4 | 18.0 |
| 18.00 | 19.5 | 4 | 18.0 |
| 18.00 | 19.5 | 4 | 18.0 |
| 19.00 | 22.0 | 5 | 23.0 |
| 19.00 | 22.0 | 5 | 23.0 |
| 19.00 | 22.0 | 5 | 23.0 |
| 20.00 | 24.0 | 5 | 23.0 |
| 21.00 | 25.0 | 5 | 23.0 |

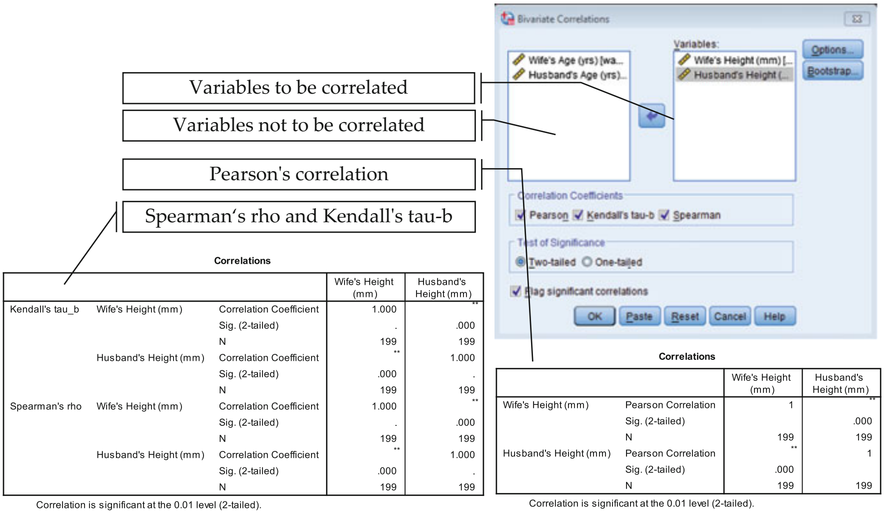
**Fig. 4.28** Association between two ordinal and metric variables

### 4.6.2 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 (Fig. 4.30).[5]

[5] Syntax command: *correlate hheight wheight*.


Variables to be correlated
Variables not to be correlated
Pearson's correlation
Spearman‘s rho and Kendall's tau-b
Bivariate Correlations
Options...
Bootstrap...
OK
Paste
Reset
Cancel
Help

Variables: Wife's Height (mm) [..., Husband's Height (...

Wife's Age (yrs) [wa..., Husband's Age (yrs)...

Correlation Coefficients: ☑ Pearson ☑ Kendall's tau-b ☑ Spearman

Test of Significance: ● Two-tailed ○ One-tailed

☑ Flag significant correlations

**Correlations**

| | | | Wife's Height (mm) | Husband's Height (mm) |
|---|---|---|---|---|
| Kendall's tau_b | Wife's Height (mm) | Correlation Coefficient | 1.000 | ** |
| | | Sig. (2-tailed) | . | .000 |
| | | N | 199 | 199 |
| | Husband's Height (mm) | Correlation Coefficient | ** | 1.000 |
| | | Sig. (2-tailed) | .000 | . |
| | | N | 199 | 199 |
| Spearman's rho | Wife's Height (mm) | Correlation Coefficient | 1.000 | ** |
| | | Sig. (2-tailed) | . | .000 |
| | | N | 199 | 199 |
| | Husband's Height (mm) | Correlation Coefficient | ** | 1.000 |
| | | Sig. (2-tailed) | .000 | . |
| | | N | 199 | 199 |

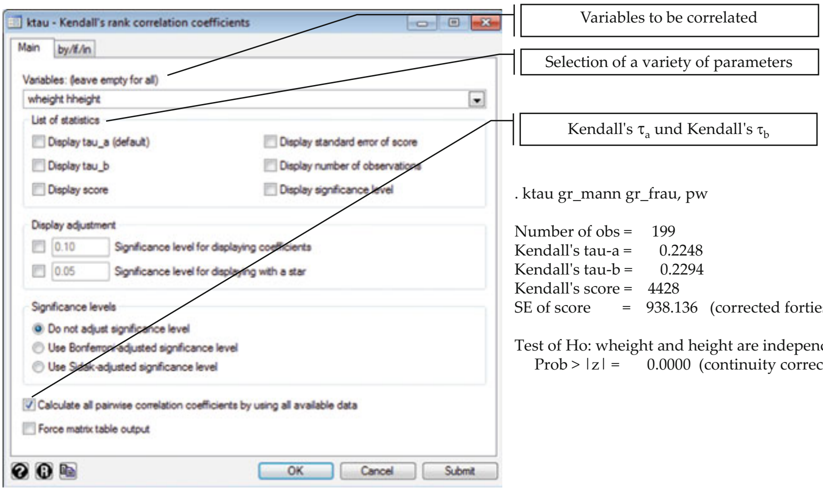
Correlation is significant at the 0.01 level (2-tailed).

**Correlations**

| | | Wife's Height (mm) | Husband's Height (mm) |
|---|---|---|---|
| Wife's Height (mm) | Pearson Correlation | 1 | ** |
| | Sig. (2-tailed) | | .000 |
| | N | 199 | 199 |
| Husband's Height (mm) | Pearson Correlation | ** | 1 |
| | Sig. (2-tailed) | .000 | |
| | N | 199 | 199 |

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

**Fig. 4.29** Calculating correlation with SPSS


Variables to be correlated
Selection of a variety of parameters
Kendall's $\tau_a$ und Kendall's $\tau_b$
ktau - Kendall's rank correlation coefficients
Main
by/if/in
OK
Cancel
Submit

Variables: (leave empty for all) wheight hheight

List of statistics
□ Display tau_a (default) □ Display standard error of score
□ Display tau_b □ Display number of observations
□ Display score □ Display significance level

Display adjustment
□ 0.10 Significance level for displaying coefficients
□ 0.05 Significance level for displaying with a star

Significance levels
● Do not adjust significance level
○ Use Bonferroni-adjusted significance level
○ Use Sidak-adjusted significance level

☑ Calculate all pairwise correlation coefficients by using all available data
□ Force matrix table output

```
. ktau gr_mann gr_frau, pw

Number of obs =     199
Kendall's tau-a =     0.2248
Kendall's tau-b =     0.2294
Kendall's score =   4428
SE of score      =   938.136   (corrected forties)

Test of Ho: wheight and height are independent
    Prob > |z| =      0.0000  (continuity corrected)
```

**Fig. 4.30** 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 $\tau$, we get the values $\tau_a = 0.581$ and $\tau_b = 0.603$. Spearman's correlation can be calculated in the same manner.

### 4.6.3 Calculating Correlation with Excel

Excel has a preprogramed 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. 4.1 we learned that Spearman's correlation 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:

$$[\text{Count (Field)} + 1 - \text{RANK (Cell; Field; 0)} - \text{RANK (Count; Field; 1)}]/2 \quad (4.39)$$

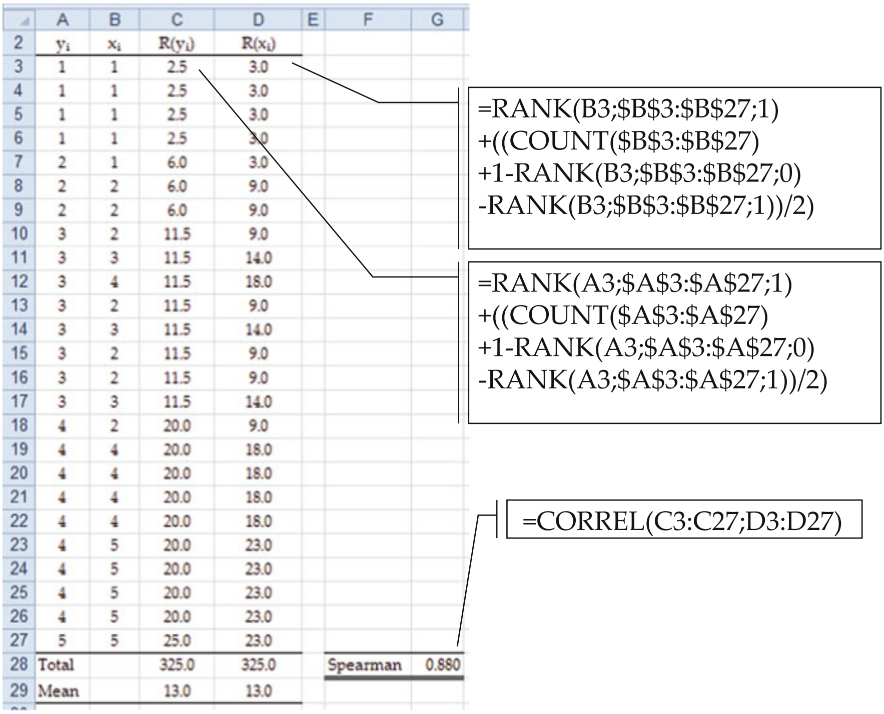
*Field* describes the arrays containing the values of the two variables (e.g. A2:B12). This correction factor must be added to every tied rank:

$$\text{RANK (Cell; Field; 1)} + \text{Correction factor} \quad (4.40)$$

The Excel formula for the correlation coefficient can be applied to the corrected ranks *Correl(Array1; Array2)*. Figure 4.31 shows once again how to calculate Spearman's correlation with Excel.

Calculating Kendall's $\tau$ 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 $\tau$. Fortunately, add-ins can be purchased for Excel that make Kendall's $\tau$ easier to calculate.

[6] Syntax command for Kendall's tau: *ktau hheight wheight, pw*. Syntax command for Spearman's rho: *ktau hheight wheight, pw*.



| | 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 | | | |

```
=RANK(B3;$B$3:$B$27;1)
+((COUNT($B$3:$B$27)
+1-RANK(B3;$B$3:$B$27;0)
-RANK(B3;$B$3:$B$27;1))/2)
```

```
=RANK(A3;$A$3:$A$27;1)
+((COUNT($A$3:$A$27)
+1-RANK(A3;$A$3:$A$27;0)
-RANK(A3;$A$3:$A$27;1))/2)
```

```
=CORREL(C3:C27;D3:D27)
```

**Fig. 4.31** Spearman's correlation with Excel

## 4.7 Spurious Correlations

Correlation is a statistical method that provides information about the 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 is many times higher in smokers than in non-smokers. Is smoking good for your health?

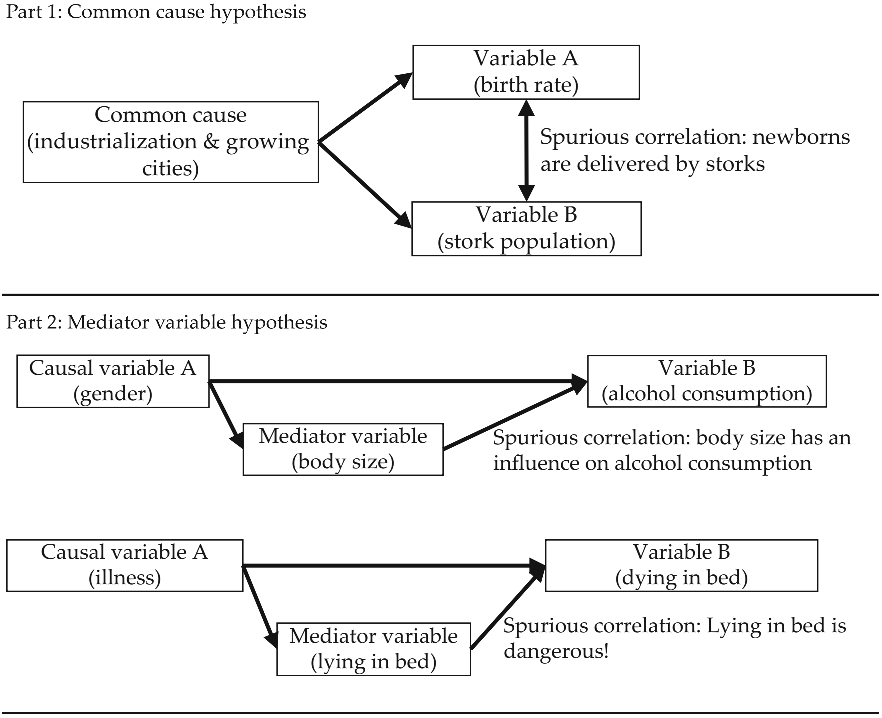

Part 1: Common cause hypothesis
Common cause (industrialization & growing cities)
Variable A (birth rate)
Variable B (stork population)
Spurious correlation: newborns are delivered by storks
Part 2: Mediator variable hypothesis
Causal variable A (gender)
Variable B (alcohol consumption)
Mediator variable (body size)
Spurious correlation: body size has an influence on alcohol consumption
Causal variable A (illness)
Variable B (dying in bed)
Mediator variable (lying in bed)
Spurious correlation: Lying in bed is dangerous!

**Fig. 4.32** Reasons for spurious correlations

- In postwar Germany there was a strong correlation between orange imports and deaths. Are oranges bad for your health?
- 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 (Fig. 4.32). 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 20th 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 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 analyzing 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?

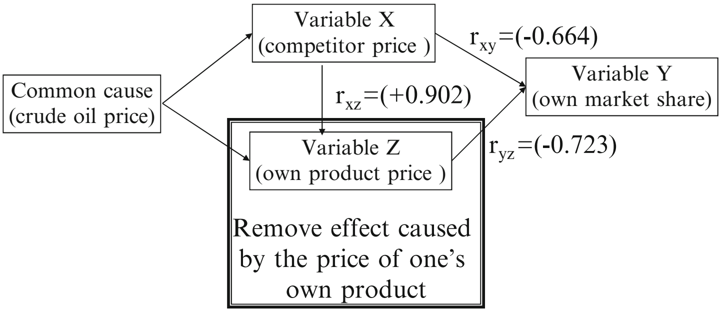


**Fig. 4.33** High-octane fuel and market share: An example of spurious correlation

Now, petrol prices are mainly shaped by crude oil prices (in addition to oligopic 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.

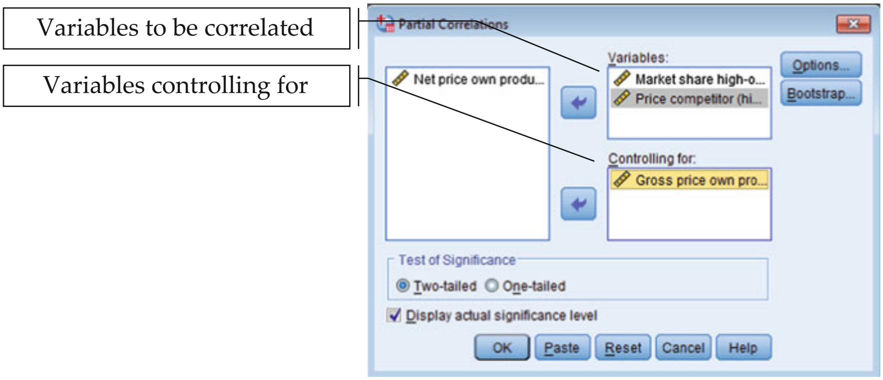
In our example, the crude oil market serves as the common cause for the price association between JETY and SPARAL. This applies both for the correlations described above and for the strong correlation coefficient – $r_{xz} = 0.902$ – between high-octane fuels at JETY and SPARAL. Both petrol stations increase (or sink) their prices almost simultaneously based on the crude oil market. The correlations are represented graphically in Fig. 4.33.

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?

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:

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{\left(1 - r_{xz}^2\right) \cdot \left(1 - r_{yz}^2\right)}} = \frac{-0.664 - (0.902 \cdot (-0.723))}{\sqrt{\left(\left(1 - 0.902^2\right) \cdot \left(1 - (-0.723)^2\right)\right)}} = -0.04 \quad (4.41)$$

This equations 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.



Variables: Market share high-o..., Price competitor (hi...

Controlling for: Gross price own pro...

Test of Significance: ● Two-tailed ○ One-tailed

☑ Display actual significance level

**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. 4.34** Partial correlation with SPSS (high-octane petrol)

### 4.7.2 Partial Correlations with SPSS

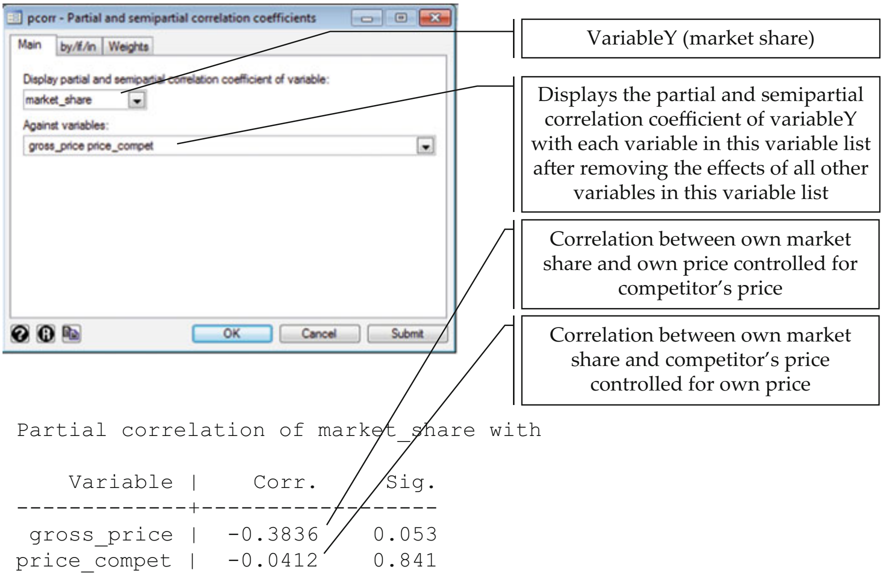
To calculate a partial correlation with SPSS, select *Analyze* → *Correlate* → *Partial*. This opens the *Partial Correlations* dialogue box. Enter the variable to be checked (SPARAL price for high-octane and SPARAL market share) under *Variables*. This produces the following partial correlation coefficient (Fig. 4.34).

### 4.7.3 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

[7] Syntax: pcorr market_share gross_price price_compet



Display partial and semipartial correlation coefficient of variable: market_share

Against variables: gross_price price_compet

VariableY (market share)

Displays the partial and semipartial correlation coefficient of variableY with each variable in this variable list after removing the effects of all other variables in this variable list

Correlation between own market share and own price controlled for competitor's price

Correlation between own market share and competitor's price controlled for own price

```
Partial correlation of market_share with

    Variable |    Corr.     Sig.
-------------+------------------
 gross_price |  -0.3836    0.053
price_compet |  -0.0412    0.841
```

**Fig. 4.35** Partial correlation with Stata (high-octane petrol)

$r_{yz.x} = -0.3836$. With the effect of SPARAL's price removed, the correlation coefficient for the association between JETY and the market share of SPARAL is $r_{xy.z} = -0.041$ (Fig. 4.35).

### 4.7.4 Partial Correlation with Excel

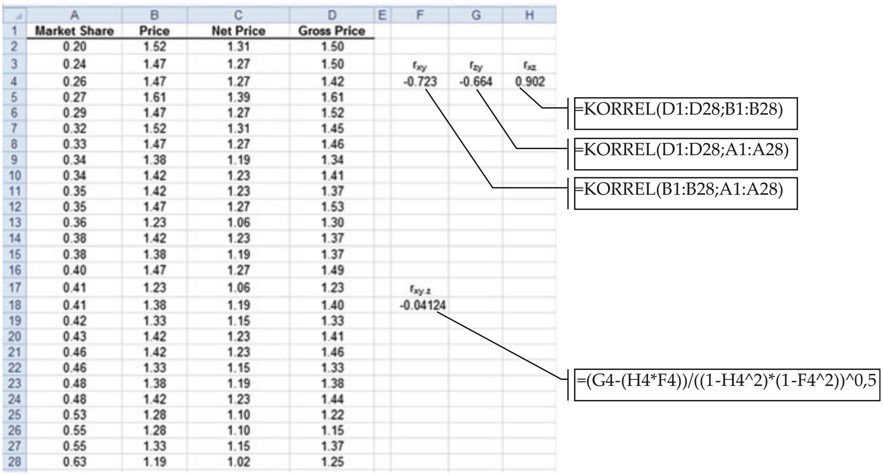
Excel has no preprogramed functions for calculating partial correlations. To perform them, you have to programme them yourself. First calculate the correlations between all variables ($r_{xy}$, $r_{xz}$, $r_{yz}$) with the CORREL command. Then use the formula $r_{xy.z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{\left(1 - r_{xz}^2\right) \cdot \left(1 - r_{yz}^2\right)}}$ to programme the partial correlation coefficient. Figure 4.36 provides some examples.

## 4.8 Chapter Exercises

**Exercise 15:**

(a) Based on the data in Exercise 8 (p. 70) 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.



| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | Market Share | Price | Net Price | Gross Price | | | | |
| 2 | 0.20 | 1.52 | 1.31 | 1.50 | | | | |
| 3 | 0.24 | 1.47 | 1.27 | 1.50 | | $r_{xy}$ | $r_{zy}$ | $r_{xz}$ |
| 4 | 0.26 | 1.47 | 1.27 | 1.42 | | -0.723 | -0.664 | 0.902 |
| 5 | 0.27 | 1.61 | 1.39 | 1.61 | | | | |
| 6 | 0.29 | 1.47 | 1.27 | 1.52 | | | | |
| 7 | 0.32 | 1.52 | 1.31 | 1.45 | | | | |
| 8 | 0.33 | 1.47 | 1.27 | 1.46 | | | | |
| 9 | 0.34 | 1.38 | 1.19 | 1.34 | | | | |
| 10 | 0.34 | 1.42 | 1.23 | 1.41 | | | | |
| 11 | 0.35 | 1.42 | 1.23 | 1.37 | | | | |
| 12 | 0.35 | 1.47 | 1.27 | 1.53 | | | | |
| 13 | 0.36 | 1.23 | 1.06 | 1.30 | | | | |
| 14 | 0.38 | 1.42 | 1.23 | 1.37 | | | | |
| 15 | 0.38 | 1.38 | 1.19 | 1.37 | | | | |
| 16 | 0.40 | 1.47 | 1.27 | 1.49 | | | | |
| 17 | 0.41 | 1.23 | 1.06 | 1.23 | | $r_{xy.z}$ | | |
| 18 | 0.41 | 1.38 | 1.19 | 1.40 | | -0.04124 | | |
| 19 | 0.42 | 1.33 | 1.15 | 1.33 | | | | |
| 20 | 0.43 | 1.42 | 1.23 | 1.41 | | | | |
| 21 | 0.46 | 1.42 | 1.23 | 1.46 | | | | |
| 22 | 0.46 | 1.33 | 1.15 | 1.33 | | | | |
| 23 | 0.48 | 1.38 | 1.19 | 1.38 | | | | |
| 24 | 0.48 | 1.42 | 1.23 | 1.44 | | | | |
| 25 | 0.53 | 1.28 | 1.10 | 1.22 | | | | |
| 26 | 0.55 | 1.28 | 1.10 | 1.15 | | | | |
| 27 | 0.55 | 1.33 | 1.15 | 1.37 | | | | |
| 28 | 0.63 | 1.19 | 1.02 | 1.25 | | | | |

=KORREL(D1:D28;B1:B28)

=KORREL(D1:D28;A1:A28)

=KORREL(B1:B28;A1:A28)

=(G4-(H4*F4))/((1-H4^2)*(1-F4^2))^0,5

**Fig. 4.36** Partial correlation with Excel (high-octane petrol)

(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 16:**
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:

(a) $\bar{x}^2 = 3.20$

(b) $\sum_{i=1}^{n} (x_i - \bar{x})^2 = 3,042.36$

(c) $\sum_{i=1}^{n} y_i = -309$

(d) $\sum_{i=1}^{n} y_i^2 = 10,545$

(e) $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = 2,987.81$

(f) What is the (unweighted) mean value for reading comprehension?
(g) What is the empirical standard deviation for reading comprehension?
(h) What is the variation coefficient for reading comprehension?
(i) Determine the empirical variance for scientific literacy.
(j) Determine the covariance between the variables x and y.

(k) Determine the strength of the linear metric association between reading comprehension and scientific literacy.
(l) Determine the rank correlation coefficient under the assumption that the sum of the squared rank differences for statistical series is 54.

**Exercise 17:**
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:
(c) $\sum_{i=1}^{n} (x_i - \bar{x})^2 = 22,500$ and $\sum_{i=1}^{n} (y_i - \bar{y})^2 = 17,000$
(d) Describe the association between the traits based on the calculated correlation coefficient. Explain.

**Exercise 18:**
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 | 9,329.4 | 4,703.4 | −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 19:**
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 20:**
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 | .3688 |
| | | Std. Deviation | .0943 |
| | Yes | Mean | .4090 |
| | | Std. Deviation | .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.

# Regression Analysis 5

## 5.1 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 toward 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).

Let us approach the topic of regression analysis with an example. 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

Chapter 5 Translated from the German original, Cleff, T. (2011). 5 Regressionsanalyse. In *Deskriptive Statistik und moderne Datenanalyse* (pp. 147–185) © Gabler Verlag, Springer Fachmedien Wiesbaden GmbH, 2011.

T. Cleff, *Exploratory Data Analysis in Business and Economics*,
DOI 10.1007/978-3-319-01517-0_5, © Springer International Publishing Switzerland 2014

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 square centimetres.

Let us first consider the case where future demand is estimated from the sales of a similar dress from the previous year. Figure 5.1 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 $r = 0.42$.

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. 5.2.

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. 5.1. 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.

## 5.2 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 $\alpha$. It determines the distance of the line along the y-axis to the origin.

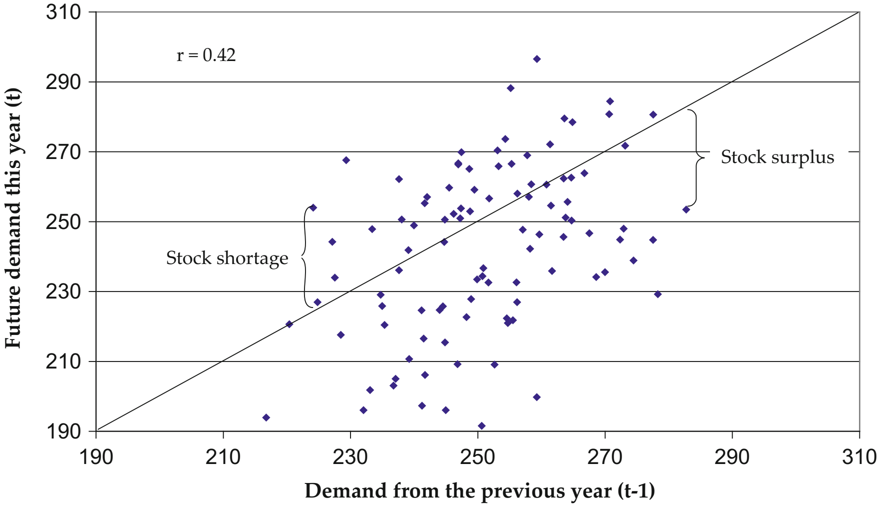


**Fig. 5.1** Demand forecast using equivalence

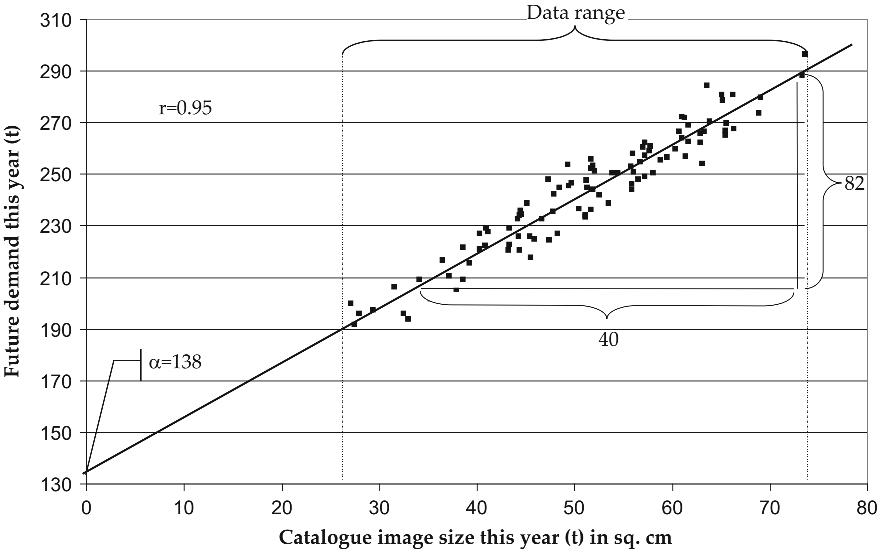


**Fig. 5.2** Demand forecast using image size

- 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 2, the value on the y-axis changes by 2 units, while the value on the x-axis changes by 1 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. 5.2 can be represented with the algebraic linear equation

$$\widehat{y} = \alpha + \beta x \tag{5.1}$$

This equation intersects the y-axis at the value 138, so that $\alpha = 138$ (see Fig. 5.2). Its slope is calculated from the slope triangle (quotient) $\beta = 82/40 \approx 2.1$. When the image size increases by 10 square centimetres, the demand increases by 21 dresses. The total linear equation is

$$\widehat{y} = 138 + 2.1 \cdot x. \tag{5.2}$$

For a dress with an image size of 50 square centimetres, we can expect sales to be

$$\widehat{y} = 138 + 2.1 \cdot 50 = 243. \tag{5.3}$$

With an image size of 70 square centimetres, the equation is

$$\widehat{y} = 138 + 2.1 \cdot 70 = 285 \text{ dresses}. \tag{5.4}$$

This linear estimation approximates the average influence of x variables on y variables using a mathematical function. The estimated values are indicated by $\widehat{y}$ and the realized y values are indicated by y. 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. 5.3. The marked data point corresponds to dress model 23, which was advertised with an image size of 47.4 square centimetres 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:

$$u_i = (y_i - \widehat{y}_i) \tag{5.5}$$

For dress model 23 the residual is:

$$u_{23} = (y_{23} - \widehat{y}_{23}) = 248 - 237.5 = 10.5 \tag{5.6}$$

In this way, every data point can be expressed as a combination of the result of the linear regression $\widehat{y}$ and its residual:

$$y_i = \widehat{y}_i + u_i \tag{5.7}$$

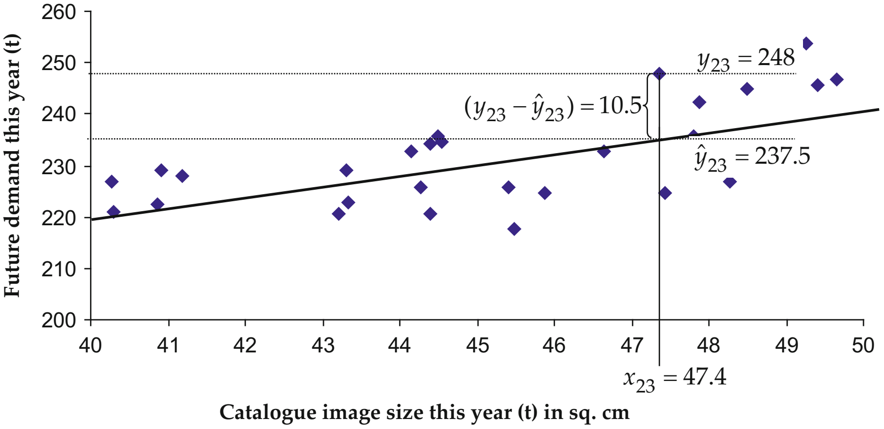


**Fig. 5.3** Calculating residuals

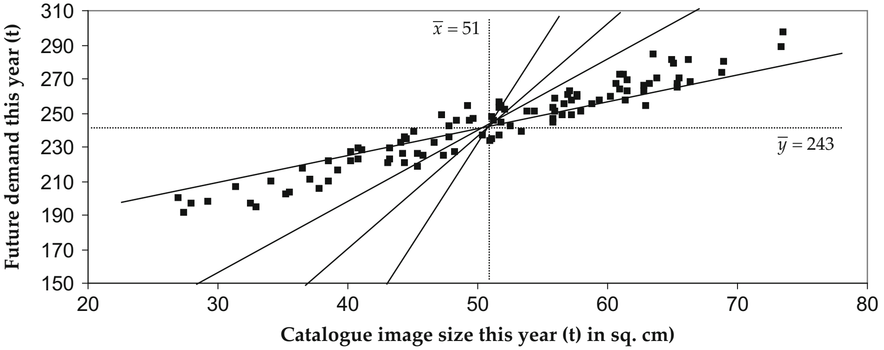


**Fig. 5.4** Lines of best fit with a minimum sum of deviations

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. 5.2 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 $\widehat{y}_{ii}$ is as close to 0 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. 5.4.

The reason for this is simple: the deviations above and below cancel each other out, resulting in a sum of 0. All lines that run through the bivariate centroid – the value pair of the averages of x and y – fulfil the condition

$$\sum_{i=1}^{n} (y_i - \widehat{y}_i) = 0 \tag{5.8}$$

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 deviation error. Frequently, statisticians create a line that minimizes *the sum of the squared deviations* of the actual data points $y_i$ from the points on the line $\widehat{y}_{i}i$. The minimization of the entire deviation error is

$$\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \to \min . \tag{5.9}$$

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 $\widehat{y}_{i}i$, we get:

$$f(\alpha;\beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \to \min . \tag{5.10}$$

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.

(i)

$$\frac{\partial f(\alpha,\beta)}{\partial \alpha} = \sum_{i=1}^{n} 2 \cdot (y_i - \alpha - \beta * x_i) \cdot (-1) = 0 \Longleftrightarrow \sum_{i=1}^{n} y_i = n \cdot \alpha + \beta \sum_{i=1}^{n} x_i \Longleftrightarrow \alpha = \overline{y} - \beta \cdot \overline{x} \tag{5.11}$$

(ii)

$$\frac{\partial f(\alpha,\beta)}{\partial \beta} = \sum_{i=1}^{n} 2 * (y_i - \alpha - \beta \cdot x_i) \cdot (-x_i) = 0 \Longleftrightarrow \sum_{i=1}^{n} (x_i y_i) = \alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 \tag{5.12}$$

The reformulation in (i) yields the formula for the constant α. We then equate (i) and (ii) to get:

$$n \cdot a + \beta\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i = \alpha\sum_{i=1}^{n} x_i + \beta\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i, \tag{5.13}$$

so that

$$\beta = \frac{\alpha\sum_{i=1}^{n} x_i - n \cdot a - \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i}{\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^2\right)} \tag{5.14}$$

By inserting this equation in (i) we get

$$\alpha = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{5.15}$$

If we place the latter in (ii) we get:

$$\sum_{i=1}^{n} (x_i y_i) = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \sum_{i=1}^{n} x_i + \beta\sum_{i=1}^{n} x_i^2 \tag{5.16}$$

After several reformulations, we arrive at the formula for the slope coefficient:

$$\frac{n\sum_{i=1}^{n} (x_i y_i) - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\mathrm{cov}(x, y)}{S_x^2} = \frac{r * S_y}{S_x} = \beta \tag{5.17}$$

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 5.5 discusses the use of computer applications for calculating regression.

## 5.3 Multivariate 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 multivariate regression is analogous to that using bivariate regression. We again assume that $\alpha = \beta_0$ and that $\beta_1$ and $\beta_2$ 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:

$$y = X \cdot \beta + u = \begin{bmatrix} y_0 \\ \cdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1+ & x_{11} & +\cdots+ & x_{k1} \\ \cdots+ & \cdots & +\cdots+ & \cdots \\ 1+ & x_{1n} & +\cdots+ & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \cdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ \cdots \\ u_n \end{bmatrix}$$

$$= \begin{bmatrix} \beta_0+ & \beta_1 x_{11} & +\cdots+ & \beta_k x_{k1} & +u_1 \\ \cdots+ & \cdots & +\cdots+ & \cdots & +\cdots \\ \beta_0+ & \beta_1 x_{1n} & +\cdots+ & \beta_k x_{kn} & +u_n \end{bmatrix} \tag{5.18}$$

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

$$\beta = \left(X'X\right)^{-1} X'y \tag{5.19}$$

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 $\beta_1$, which describes the relationship between catalogue image size and demand; and
- the second slope coefficient $\beta_2$, which describes the relationship between previous sales and demand.

The equation to determine the multivariate regression is thus:

$$\begin{aligned} \hat{y} &= \alpha + \beta_1 \cdot \textit{catalogue image size} + \beta_2 \cdot \textit{previous sales} \\ &= \alpha + \beta_1 x_1 + \beta_2 x_2 \end{aligned} \tag{5.20}$$

## 5.4 The Goodness of Fit of Regression Lines

A regression seeks to describe the average association of two or more variables. In Figs. 5.1 and 5.2 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. 5.2) is better suited for predicting sales than the previous sales ("equivalence") method (Fig. 5.1), 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 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.

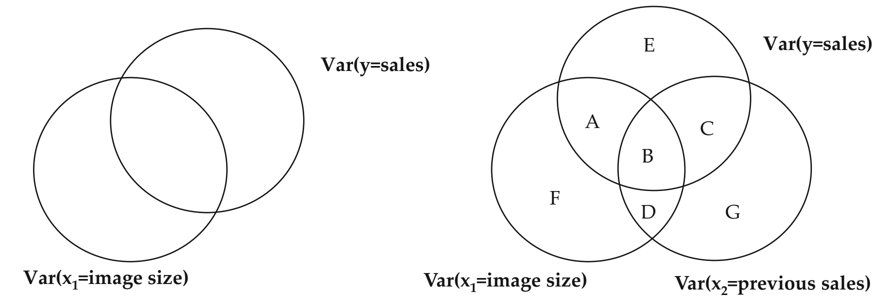
The coefficient of determination also indicates the share of y variance explained by x variance. In our example (Fig. 5.2) the coefficient of determination is $R^2 = 0.96^2 = 0.9216 = 92.16\%$. This means that 92.16% of the variance in sales (y variable) is explained through variance in catalogue image size (x variable).

Figure 5.5 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: the previous year's sales ($x_2$). Here the intersection between y variance (sales) and $x_1$ variance (image size) and the previous year's sales ($x_2$) increases. With the regression lines $\widehat{y}$, the variances of the independent x variables explain

$$R^2 = \left(\frac{(A+B+C)}{(A+B+C+E)}\right) \cdot 100 \text{ per cent} \tag{5.21}$$

of y variance. The general formula for $R^2$ in a multivariate regression can thus be expressed as follows:



**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. 5.5** The concept of multivariate analysis

$$R^2 = \frac{S_{\widehat{y}}^2}{S_y^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{y}_i - \overline{y}\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2} \tag{5.22}$$

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* $RSS = \sum_{i=1}^{n}\left(\widehat{y}_i - \overline{y}\right)^2$ divided by the *total sum of squares* $TSS = \sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$:

$$R^2 = \frac{RSS}{TSS} = \frac{\sum_{i=1}^{n}\left(\widehat{y}_i - \overline{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2} \tag{5.23}$$

R$^2$can also be calculated using the unexplained variance of y variables:

$$S_e^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2 \tag{5.24}$$

In part 2 above, the unexplained variance is represented by region E. The correlation of determination can then be defined as follows:

$$R^2 = 1 - \frac{S_\epsilon}{S_y} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{5.25}$$

Expressed using the *residual or error sum of squares* $ESS = \sum_{i=1}^{n}(y_i - \overline{y})^2$, $R^2$ is

$$R^2 = 1 - \frac{ESS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{5.26}$$

Another way to evaluate goodness of fit with multivariate regression is the adjusted coefficient of determination. We will learn about this approach in Sect. 5.6.

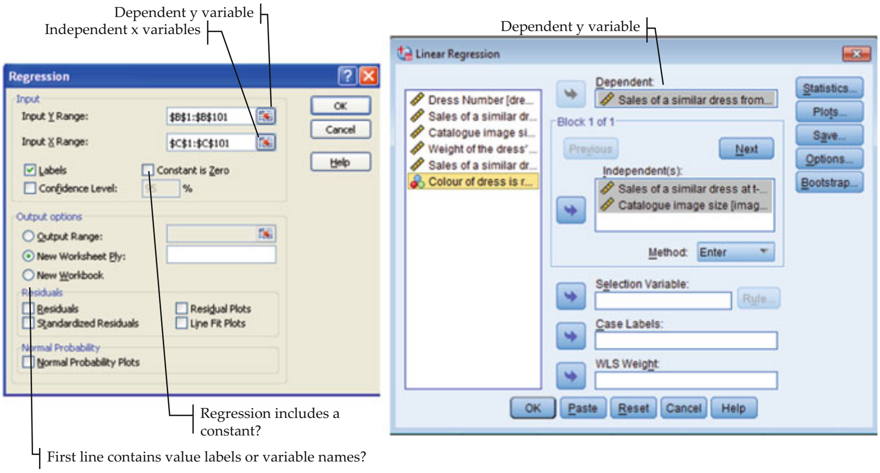
## 5.5 Regression Calculations with the Computer

### 5.5.1 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 part 1 of Fig. 5.6 appears. Next assign the fields for dependent and independent variables. Keep in mind that the independent variables must be arranged next to each other in the Excel tables

[1] To use the *Linest function* for the dataset *mail_order_business.xls*, 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 5 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 $\beta_1$ to $\beta_k$ are displayed. The last row of the first line contains the value of the constant α. 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-In Manager can be accessed via *File→Options→Add-ins→* Manage: Excel Add-ins →Go. . .


Dependent y variable
Independent x variables
Regression includes a constant?
First line contains value labels or variable names?
Dependent y variable

Regression

Input

Input Y Range: $B$1:$B$101

Input X Range: $C$1:$C$101

☑ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

OK Cancel Help

Linear Regression

Dress Number [dre...
Sales of a similar dr...
Catalogue image si...
Weight of the dress'...
Sales of a similar dr...
Colour of dress is r...

Dependent: Sales of a similar dress from...

Block 1 of 1 Previous Next

Independent(s): Sales of a similar dress at t-... Catalogue image size [imag...

Method: Enter

Selection Variable: Rule...

Case Labels:

WLS Weight:

Statistics... Plots... Save... Options... Bootstrap...

OK Paste Reset Cancel Help

Part 1: Regression with Excel Part 2: Regression with SPSS

**Fig. 5.6** Regression with Excel and SPSS

and may contain no missing values. Our example uses the file *mail_order_business.xls*. The interpretation of the output is in Sect. 5.5.2. The output from all statistical software applications I discuss is the same.

## 5.5.2 Regression Calculations with SPSS and Stata

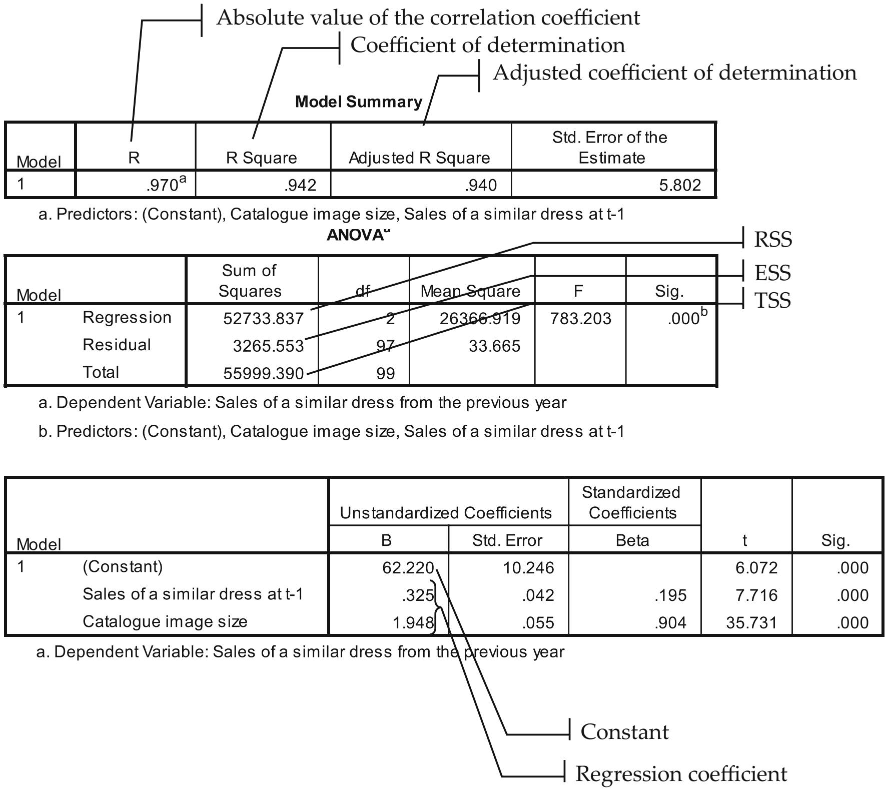
The calculation is similar using SPSS and Stata. In SPSS, open the *Linear Regression* window shown in part 2 of Fig. 5.6 by selecting *Analyze→Regression→Linear*. Then assign the dependent and independent variables and confirm the selection by clicking *OK*.

With Stata access the regression menu by choosing *Statistics→Linear models and related→Linear regression*. Then enter the dependent variables in the *dependent variable* field and the independent variables in the *independent variable* field and click *OK* or *Submit*.

Each programme displays the calculation results in 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 5.7 shows the result tables of the regression function with SPSS.

From these results we can determine the sales of period (t) using the following equation:

$$\hat{y} = 62.22 + 1.95 \cdot catalogue\ image\ size + 0.33 \cdot previous\ sales\ (t-1) \quad (5.27)$$


Absolute value of the correlation coefficient
Coefficient of determination
Adjusted coefficient of determination
RSS
ESS
TSS
Constant
Regression coefficient

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .970[a] | .942 | .940 | 5.802 |

a. Predictors: (Constant), Catalogue image size, Sales of a similar dress at t-1

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 52733.837 | 2 | 26366.919 | 783.203 | .000[b] |
| | Residual | 3265.553 | 97 | 33.665 | | |
| | Total | 55999.390 | 99 | | | |

a. Dependent Variable: Sales of a similar dress from the previous year

b. Predictors: (Constant), Catalogue image size, Sales of a similar dress at t-1

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 62.220 | 10.246 | | 6.072 | .000 |
| | Sales of a similar dress at t-1 | .325 | .042 | .195 | 7.716 | .000 |
| | Catalogue image size | 1.948 | .055 | .904 | 35.731 | .000 |

a. Dependent Variable: Sales of a similar dress from the previous year

**Fig. 5.7** Output from the regression function for SPSS

If a dress is advertised with an image of 50 square centimetres and a similar dress sold 150 times last year, then we can expect average sales of

$$\widehat{y} = 62.22 + 1.95 \cdot 50 + 0.33 \cdot 150 \approx 209 \text{ dresses} \tag{5.28}$$

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 = 3,265.553. From this we can also calculate the coefficient of determination, were it not already indicated above:

$$R^2 = \frac{52,733.873}{55,999.390} = 94.2\% \tag{5.29}$$

The variance of the independent x variables (the demand of a similar dress in the previous season; the catalogue image size) explains the variance of the dependent variable (the sales of a dress in the current season) for $R^2 = 94.2\%$.

## 5.6 Goodness of Fit of Multivariate 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*.

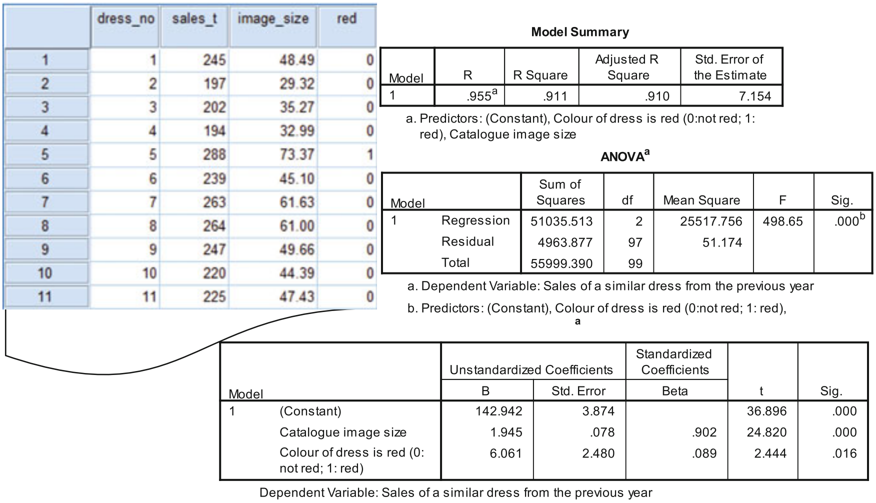
In practice, statisticians frequently calculate what is called the adjusted $R^2$, which penalizes overparametrization. With every additional variable, the penalization increases. The adjusted coefficient of determination can be calculated by the following equation, where n is the number of observations and k the number of variables in the model (including constants):

$$R^2_{adj} = R^2 - \frac{\left(1 - R^2\right)(k-1)}{(n-k)} = 1 - \left(1 - R^2\right)\frac{n-1}{n-k} \tag{5.30}$$

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 must 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.[3]

[3] For $R^2 = 0$ and $k > 1$ the following equation applies: $R^2_{adj} = 0 - \frac{(1-0)(k-1)}{(n-k)} = \left(-\frac{(k-1)}{(n-k)}\right) < 0$.



| | 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),

a

| 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. 5.8** Regression output with dummy variables

## 5.7 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 Sect. 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 sq. cm) 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 5.8 shows the results of the regression.

The regression can be expressed with the following algebraic equation:

$$\hat{y} = 142.9 + 1.95 \cdot catalogue\ image\ size + 6.1 \cdot red \tag{5.31}$$

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

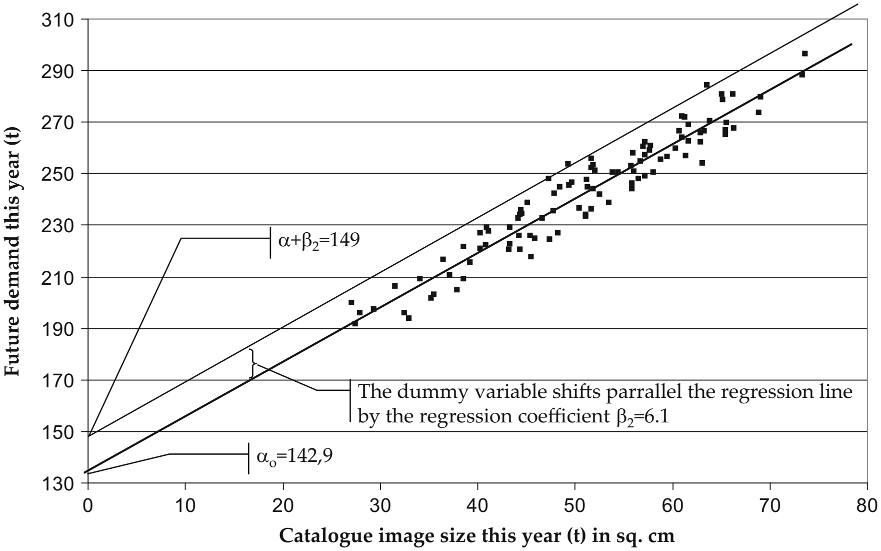


**Fig. 5.9** The effects of dummy variables shown graphically

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. 5.9).

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 3 colours in relation to the remaining dress colours (neither red nor green nor blue). Say we obtain the following regression:

$$\widehat{y} = 140 + 1.9 \cdot \textit{catalogue image size} + 6 \cdot \textit{red} + 5 \cdot \textit{green} + 4 \cdot \textit{blue} \qquad (5.32)$$

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.

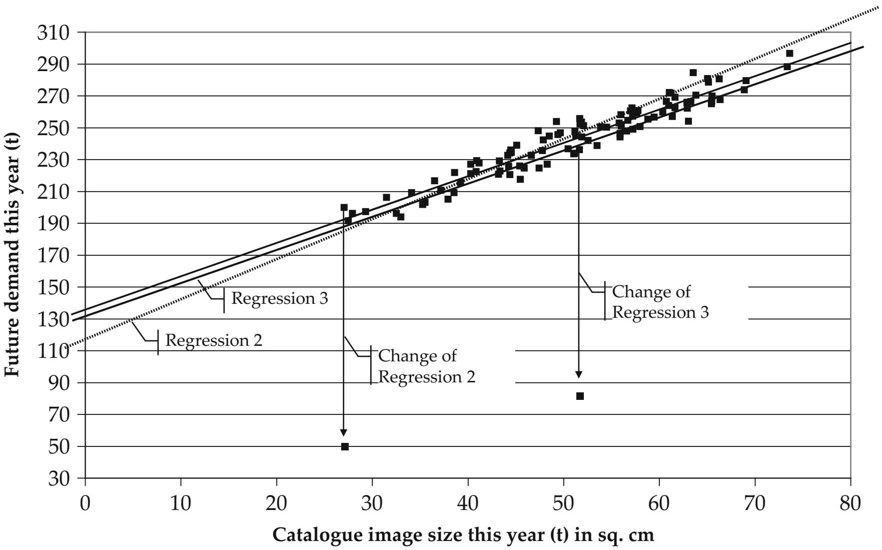


**Fig. 5.10** Leverage effect

## 5.8 Leverage Effects of Data Points

Let's look at the data points for the mail order business shown in Fig. 5.10. Consider the graph's first data point, which represents a dress advertised with 27.1 square centimetres 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. 5.10 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 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. 5.8).

## 5.9 Nonlinear 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 ($\alpha$, $\beta_1$, $\beta_2$,. . ., $\beta_k$) 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. 5.7 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 ($\alpha$, $\beta_1$, $\beta_2$,. . ., $\beta_k$) remain constant. In part 4 of Fig. 5.11, 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 Figure 5.11, a least squares regression can be carried out, although the regression is non-linear.

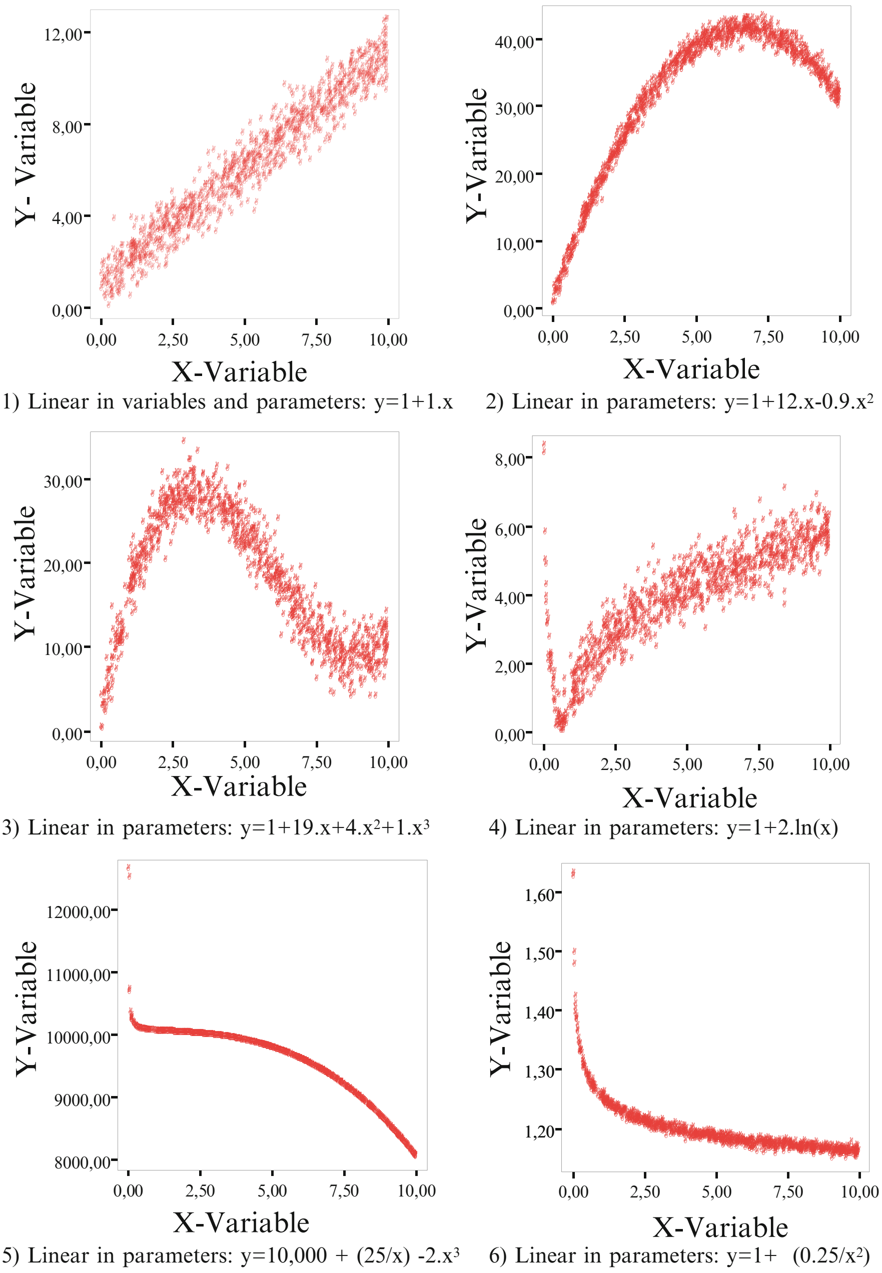
Using the least squares regression we can also represent non-linear associations: a regression need not be limited to the form of a straight line. Let's look at an example to understand how to approach regressions with variables that have a non-linear association. Figure 5.12 displays monthly sales figures [in €10,000s] and the number of consultants in 27 store locations. A linear regression calculated from these data produces the following regression line:

$$\widehat{y} = 0.0324 \cdot x + 55.945; R^2 = 0.66 \tag{5.33}$$

If the number of consultants in a district increases by 1, sales increases on average by

$$\Delta\widehat{y} = 0.0324 \cdot 1 \cdot [€10,000] = €3,240 \tag{5.34}$$

Yet a closer examination reveals that this regression line contains systematic errors. In a district containing between 20 and 100 consultants, the regression line



1) Linear in variables and parameters: y=1+1.x

2) Linear in parameters: $y=1+12.x-0.9.x^2$

3) Linear in parameters: $y=1+19.x+4.x^2+1.x^3$

4) Linear in parameters: y=1+2.ln(x)

5) Linear in parameters: $y=10{,}000 + (25/x) -2.x^3$

6) Linear in parameters: $y=1+ (0.25/x^2)$

**Fig. 5.11** 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 |

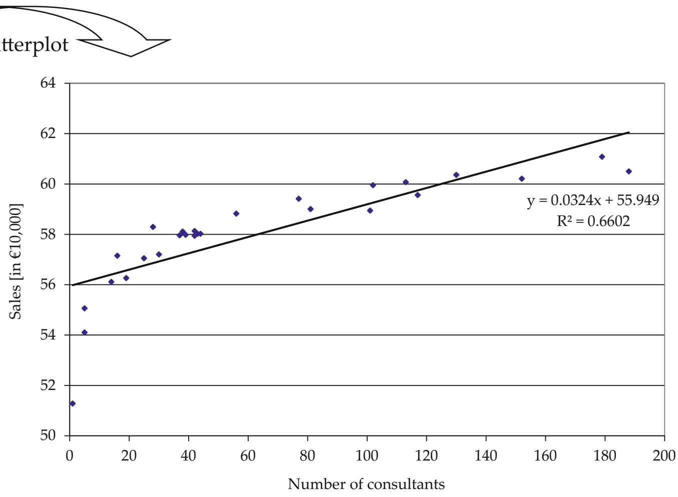


**Fig. 5.12** Regression with non-linear variables (1)

underestimates sales throughout, while in a district with 140 consultants or more, the regression lines overestimates sales throughout. The reason: a non-linear association exists between the x and y values, leading to a non-linear regression line.

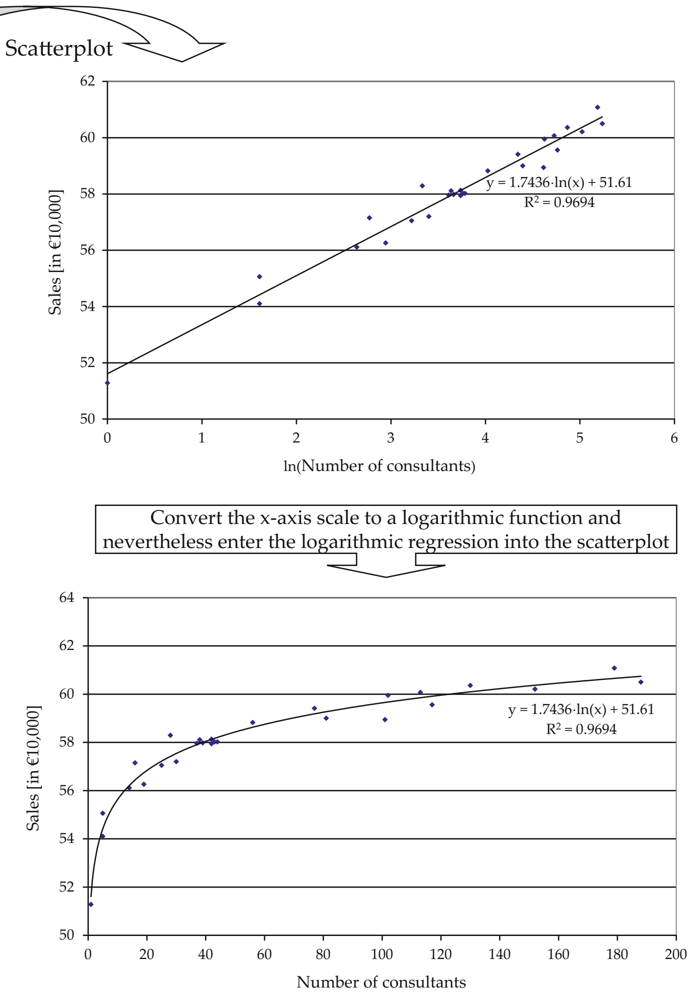
If we convert the x variable to a logarithmic function – the form of the scatterplot suggests a logarithmic regression – we get the upper scatterplot in Fig. 5.13. Here the x-axis does not track the number of consultants but the logarithmic function of the number of consultants. Now the regression line

$$\widehat{y} = 1.7436 \cdot \ln(x) + 51.61 \tag{5.35}$$

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 $\mathrm{R}^2 = 0.97$.

Of course, we can also choose *not* to convert the x-axis scale to a logarithmic function (see the lower scatterplot in Fig. 5.13) 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 ($\widehat{y} = 1.7436 \cdot \ln(x) + 51.61$).

| 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 |



Convert the x-axis scale to a logarithmic function and nevertheless enter the logarithmic regression into the scatterplot

**Fig. 5.13** Regression with non-linear variables (2)

## 5.10 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 $\widehat{y} = 62.22 + 1.95 \cdot \textit{catalogue image size} + 0.33 \cdot \textit{previous sales}\ (t-1)$. 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

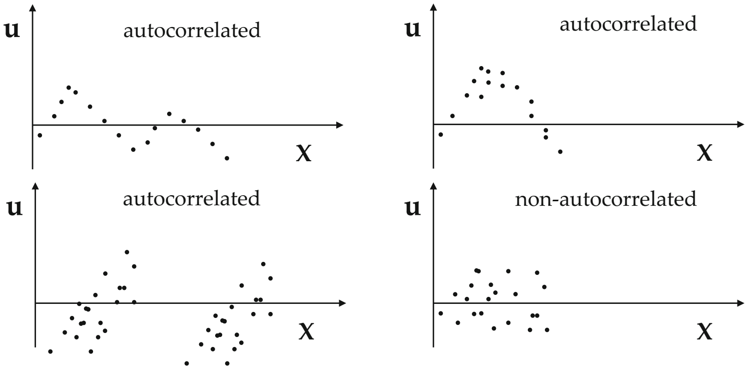


**Fig. 5.14** Autocorrelated and non-autocorrelated distributions of error terms

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 value can be expressed as a combination of the value estimated by the regression ($\widehat{y}_i$) and the accompanying error term ($u_i$). Since $\widehat{y}_i$ represents the outcome of the regression equation from $x_i$, we get:

$$y_i = \widehat{y}_i + u_i = \alpha + \beta \cdot x_i + u_i \tag{5.36}$$

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 (u). The case described in Fig. 5.8 – where x-axis deviations only appear in a certain direction (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:

$$\mathrm{Cov}\left(u_i; u_j\right) = 0 \text{ i} \neq \text{j} \tag{5.37}$$

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 square centimetres or smaller and with image sizes 60 square centimetres or larger, and mostly negative deviations obtain with image sizes between 40 and 60 square centimetres. Figure 5.14 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 non-linear

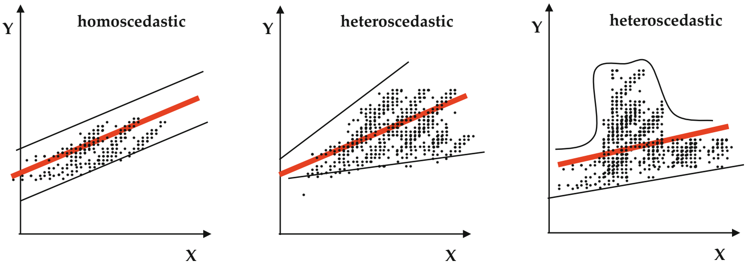


**Fig. 5.15** Homoscedasticity and Heteroscedasticity

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: $\text{Var}(u_i) = \sigma^2$. 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. 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 5.15 provides examples of the undesirable effect. Here too, the model must be checked for error specification (missing variables or an erroneous selection of the functional distribution).

   We can examine the quality conditions for the error term $u$ with a graphical analysis (see for instance Figs. 5.14 and 5.15). 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 one or more x variables is too large, so-called *multicollinearity* occurs, which falsifies the regression outcome.

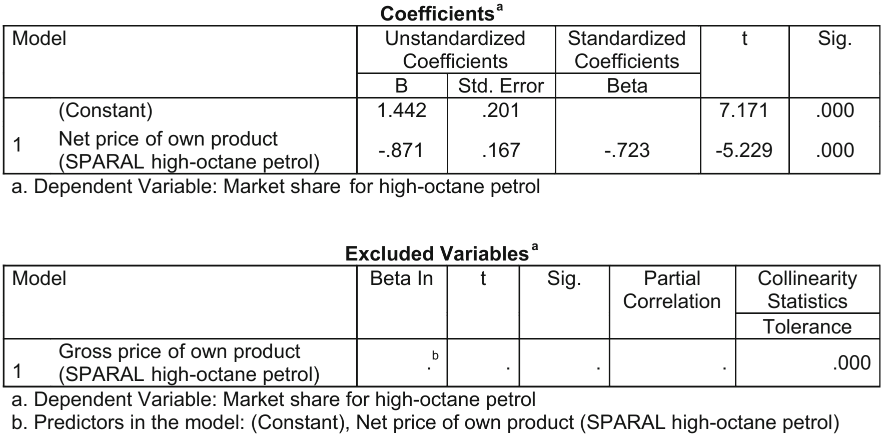
Ultimately, this condition entails nothing more than choosing two variables for the predictor x variables whose meaning is different or at least dissimilar. If we estimate the market share for petrol using gross and net prices from the SPSS file *multicollinearity_petrol_example.sav*, we get the output displayed in Fig. 5.16.

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:

$$\textit{net price} = \textit{gross price}/1.19 \tag{5.38}$$

The regression

$$\widehat{y} = \beta_o + \beta_1 \cdot \textit{net price} + \beta_2 \cdot \textit{gross price} \tag{5.39}$$



**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. 5.16** Solution for perfect multicollinearity

can be converted into:

$$\widehat{y} = \beta_o + \left(\frac{\beta_1}{1.19} + \beta_2\right) \cdot \mathit{gross\ price} \Longleftrightarrow \widehat{y} = \alpha + \beta \cdot \mathit{gross\ price} \qquad (5.40)$$

It would have been necessary to calculate the two regression coefficients $\beta_1$ and $\beta_2$, although there is only one linearly independent variable (gross or net). If perfect multicollinearity exists, it is impossible to determine certain regression coefficients.[4] 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:

- 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

[4] In Sect. 5.3 we calculated the regression coefficients $\beta$=($\alpha$=$\beta_0$; $\beta_1$; ... ; $\beta_k$) 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.

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.

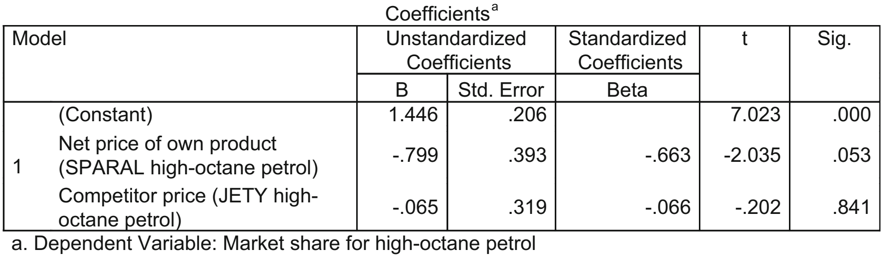
- The algebraic sign of the regressor is unusual. The competitor's price appears to have the same direction of effect on market share as the company's own price, i.e. the higher the competitor's price, the lower the market share.
- Removing or adding an observation from the dataset leads to large changes 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 *multicollinearity_petrol_example.sav* (see Sect. 4.7.1) 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:

$$x_1 = \alpha_o + \alpha_1 \cdot x_2 + \alpha_2 \cdot x_3 + \alpha_3 \cdot x_4 + \alpha_4 \cdot x_5 \tag{5.41}$$

The larger the coefficient of determination $R^2_{Aux(1)}$ 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

$$\text{VIF}_i = \frac{1}{1\text{-}R^2_{Aux(i)}} \tag{5.42}$$

If the $R^2_{Aux(i)}$ value of the auxiliary regression of an independent variable is (close to) 0, no multicollinearity exists and VIF = 1. If, by contrast, the $R^2_{Aux(i)}$ 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



Coefficients[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 1.446 | .206 | | 7.023 | .000 |
| | Net price of own product (SPARAL high-octane petrol) | -.799 | .393 | -.663 | -2.035 | .053 |
| | Competitor price (JETY high-octane petrol) | -.065 | .319 | -.066 | -.202 | .841 |

a. Dependent Variable: Market share for high-octane petrol

**Fig. 5.17** Solution for imperfect multicollinearity

multicollinearity and, when the VIF is conspicuously high, check the robustness of the results. Keep in mind, though, that a VIP as low as 5.3 already has a very high multiple correlation, namely, r = 0.9. For this reason, whenever the VIP is 1.7 or higher – VIP = 1.7 translates into a multiple correlation of r = 0.64 – 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 Tolerance(i) = (1-$R^2_{Aux(i)}$). 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. 5.17 the VIFs and the Tolerances of the dataset *multicollinearity_petrol_example.sav* 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 dataset. 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.

## 5.11 Chapter Exercises

**Exercise 21:**

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 $\widehat{y} = 1.26 - 0.298 \cdot \text{price}$

(a) What average market share can be expected when the coffee price is 3 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 22:**

You have a hunch that the product sales mentioned in Exercise 8 (p. 70) are not determined by price alone. So you perform a multivariate regression using Excel (or statistics software like SPSS). The results of the regression are listed in the tables below.

(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.

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]

| | | Unstandardized coefficients | | | |
|---|---|---|---|---|---|
| Model | | B | Std. error | t | Sig. |
| 1 | (Constant) | 24.346 | 3.107 | 7.84 | .000 |
| | Number of dealerships | .253 | .101 | 2.50 | .047 |
| | Unit price [in 1,000s of MUs] | − .647 | .080 | −8.05 | .000 |
| | Advertising budget [in 100,000s MUs] | − .005 | .023 | −0.24 | .817 |

**Exercise 23:**

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 below:

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 | | | |

| | | Unstandardized coefficients | | | |
|---|---|---|---|---|---|
| Model | | B | Std. error | t | Sig. |
| 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?

**Exercise 24:**

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 3 years. You attempt to estimate the factors influencing weekly market share using the regression below. 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),
- Senso White advertised with leaflets that contain images 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 | | | |

| | | Unstandardized coefficients | | | |
|---|---|---|---|---|---|
| Model | | B | Std. error | Standardized coefficients BETA | Sig. |
| 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 |
| | Senso White advertised in daily newspapers | −404.053 | 87.042 | – .18 | .000 |
| | Other toothpaste brands 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 the above table.

(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 adjustedR$^2$. 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 25:**

The fast food chain Burger Slim wants to introduce a new children's meal. The company decides to test different meals at its 2,261 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 following results:

**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 | 2,256 | ??? | | |
| | Total | 136636463021.389 | 2,260 | | | |

| | | Unstandardized coefficients | | | | |
|---|---|---|---|---|---|---|
| Model | | B | Std. error | Standardized coefficients BETA | t | Sig. |
| 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 | 2,255 | ??? | | |
| | Total | 136636463021.389 | 2,260 | | | |

| | | Unstandardized coefficients | | | | | |
|---|---|---|---|---|---|---|---|
| Model | | B | Std. error | Standardized coefficients BETA | Sig. | Tolerance | VIF |
| 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 $R^2$ from regression 1.
(b) What is the adjusted $R^2$ needed for?
(c) Using regression 1 determine the average revenue generated by a meal that costs 5 euros and contains a slim burger, a 0.5 l soft drink, a salad, and a toy.
(d) Using regression 2 determine which variable has the largest influence. Explain your answer.
(e) Compare the results of regressions 1 and 2. Which of the solutions would you consider in a presentation for the client?
(f) Consider the following scatterplot. What's the problem? Describe the effects of the results from regressions 1 and 2 on interpretability. How can the problem be eliminated?

# 6 Time Series and Indices

In the preceding chapter 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 6.1 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 (2012) and Wooldridge (2009).

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

Chapter 6 Translated from the German original, Cleff, T. (2011). 6 Zeitreihen- und Indexrechnung. In *Deskriptive Statistik und moderne Datenanalyse* (pp. 187–204) © Gabler Verlag, Springer Fachmedien Wiesbaden GmbH, 2011

T. Cleff, *Exploratory Data Analysis in Business and Economics*,
DOI 10.1007/978-3-319-01517-0_6, © Springer International Publishing Switzerland 2014

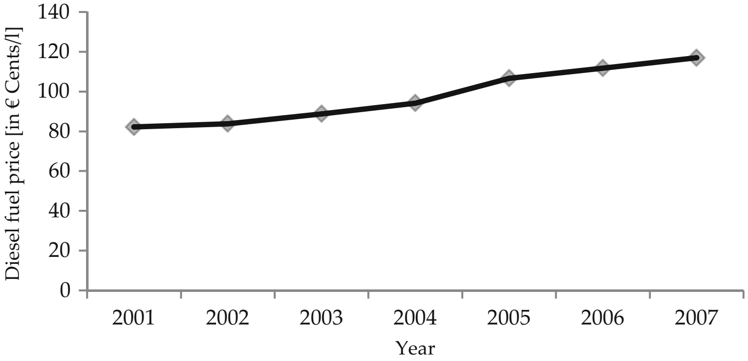


**Fig. 6.1** Diesel fuel prices by year, 2001–2007

in rapid succession. Times series data are similar, as they allow us to recognize movements and trends and to project them into the future. Below we investigate the most frequently used technique for measuring dynamic phenomena: index figures.

## 6.1 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 6.1 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*:

$$P^{*}_{\text{base year}=0,\text{reporting year}=t} = \frac{\text{Price in reporting year } \left(p_t\right)}{\text{Price in base year } \left(p_0\right)} \tag{6.1}$$

$$P^{*}_{2001,2007} = \frac{p_{2007}}{p_{2001}} = \frac{117.0}{82.2} = 1.42 \tag{6.2}$$

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 6.1. 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.

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

**Table 6.1** 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 1,000 t and (share of consumption in %) | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
| High octane | 18,979 | 18,785 | 18,140 | 17,642 | 16,870 | 16,068 | 15,718 |
|  | (33.6 %) | (33.7 %) | (33.7 %) | (32.7 %) | (32.5 %) | (31.5 %) | (31.2 %) |
| Regular | 8,970 | 8,409 | 7,710 | 7,395 | 6,561 | 6,181 | 5,574 |
|  | (15.9 %) | (15.1 %) | (14.3 %) | (13.7 %) | (12.6 %) | (12.1 %) | (11.1 %) |
| Diesel | 28,545 | 28,631 | 27,944 | 28,920 | 28,531 | 28,765 | 29,059 |
|  | (50.5 %) | (51.3 %) | (52.0 %) | (53.6 %) | (54.9 %) | (56.4 %) | (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, http://www.mwv.de. Based on the author's calculations.

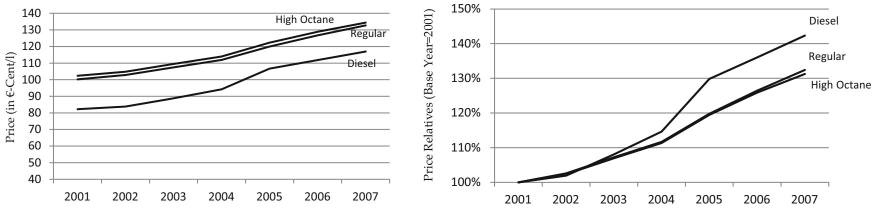


Figure 1

Figure 2

**Fig. 6.2** Fuel prices over time

time (see Fig. 6.2, part 1) with their index series from the base year 2001 (see Fig. 6.2, 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 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:

$$P^*_{\text{new base year,t}} = \frac{P^*_{\text{old base year}}}{P^*_{\text{new base year}}} \tag{6.3}$$

Now we can see that the percentage change between 2004 and 2007 is 23 %:

$$P^*_{2004,2007} = \frac{P^*_{2001,2007}}{P^*_{2001,2004}} = \frac{1.42}{1.15} = 1.23 \tag{6.4}$$

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

[1] Strictly speaking, a base shift need only be undertaken when the market basket linked to the time series is changed (see Sect. 6.5).

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. 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):

$$\mathrm{P}^{\mathrm{L}}_{0,t} = \frac{\sum\limits_{i=1}^{n} \frac{p_{i,t}}{p_{i,0}} \cdot p_{i,0} \cdot q_{i,0}}{\sum\limits_{i=1}^{n} p_{i,0} \cdot q_{i,0}} = \frac{\sum\limits_{i=1}^{n} p_{i,t} \cdot q_{i,0}}{\sum\limits_{i=1}^{n} p_{i,0} \cdot q_{i,0}} \tag{6.5}$$

Usually, Laspeyres index figures are multiplied by 100 or, as with the DAX stock market index, by 1,000. For example, the Federal Statistical Office of Germany expresses the inflation rate as $\mathrm{P}^{\mathrm{L}}_{0,t}$ multiplied by 100:[4]

$$100 \cdot \mathrm{P}^{\mathrm{L}}_{0,t} = 100 \cdot \frac{\sum\limits_{i=1}^{n} p_{i,t} \cdot q_{i,0}}{\sum\limits_{i=1}^{n} p_{i,0} \cdot q_{i,0}} \tag{6.6}$$

[2] For more, see the information on consumer price statistics at http://www.destatis.de. The site calculates the price increase rate for individuals with a personal inflation calculator.

[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.

In the example with diesel and petrol, total demand is 28,545,000 t of diesel ($q_{\text{diesel},2001}$), 8,970,000 t of regular ($q_{\text{regular},2001}$) and 18,979,000 t 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 ($q_{i,2001}$) and add them together. This total goes in the numerator. Next we weight the amounts consumed in 2001 with the prices of 2001 ($p_{i,2001}$) and add them together. This total goes in the denominator. Now we have the following weighted percentage change in price:

$$P^{L}_{\text{base year},\text{Berichtsjahr}} = \frac{\sum_{i=1}^{n} p_{i,\text{report year}} \cdot q_{i,\text{base year}}}{\sum_{i=1}^{n} p_{i,\text{base year}} \cdot q_{i,\text{base year}}} = P^{L}_{0,t} = \frac{\sum_{i=1}^{n} p_{i,t} \cdot q_{i,0}}{\sum_{i=1}^{n} p_{i,0} \cdot q_{i,0}} \tag{6.7}$$

$$P^{L}_{2001,2007} = \frac{134.4 \cdot 18,979 + 132.7 \cdot 8,970 + 117.0 \cdot 28,545}{102.4 \cdot 18,979 + 100.2 \cdot 8,970 + 82.2 \cdot 28,545} = 1.3647 \tag{6.8}$$

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:

$$P^{L}_{0,t} = \frac{\sum_{i=1}^{n} p_{i,t} \cdot q_{i,0}}{\sum_{i=1}^{n} p_{i,0} \cdot q_{i,0}} = \frac{\sum_{i=1}^{n} p_{i,t} \cdot \frac{q_{i,0}}{\sum_{j=1}^{n} q_{j,0}}}{\sum_{i=1}^{n} p_{i,0} \cdot \frac{q_{i,0}}{\sum_{j=1}^{n} q_{j,0}}} = \frac{\sum_{i=1}^{n} p_{i,t} \cdot f_{q_{i,0}}}{\sum_{i=1}^{n} p_{i,0} \cdot f_{q_{i,0}}} \tag{6.9}$$

$$P^{L}_{2001,2007} = \frac{134.4 \cdot 33.6\% + 132.7 \cdot 15.9\% + 117.0 \cdot 50.5\%}{102.4 \cdot 33.6\% + 100.2 \cdot 15.9\% + 82.2 \cdot 50.5\%} = 1.3647 \tag{6.10}$$

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 the 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 250g 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 (Krämer 2008, p. 87).

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 computer 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 2008, for instance, the Federal Statistical Office of Germany changed the base year from 2000 to 2005.

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.

[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.

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:

$$P^P_{\text{base year,report year}} = \frac{\sum_{i=1}^{n} p_{i,\text{report year}} \cdot q_{i,\text{report year}}}{\sum_{i=1}^{n} p_{i,\text{base year}} \cdot q_{i,\text{report year}}} = P^P_{0,t} = \frac{\sum_{i=1}^{n} p_{i,t} \cdot q_{i,t}}{\sum_{i=1}^{n} p_{i,0} \cdot q_{i,t}} \quad (6.11)$$

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 t of diesel ($q_{\text{Diesel},2007}$), 5,574,000 t of regular ($q_{\text{regular},2007}$) and 15,718,000 t 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:

$$P^P_{2001,2007} = \frac{\sum_{i=1}^{n} p_{i,2007} \cdot q_{i,2007}}{\sum_{i=1}^{n} p_{i,2001} \cdot q_{i,2007}} \quad (6.12)$$

$$P^P_{2001,2007} = \frac{134.4 \cdot 15,718 + 132.7 \cdot 5,574 + 117.0 \cdot 29,059}{102.4 \cdot 15,718 + 100.2 \cdot 5,574 + 82.2 \cdot 29,059} = 1.3721 \quad (6.13)$$

This is then weighted by the shares of the total expenditure:

$$P^P_{2001,2007} = \frac{134.4 \cdot 31.2\% + 132.7 \cdot 11.1\% + 117.0 \cdot 57.7\%}{102.4 \cdot 31.2\% + 100.2 \cdot 11.1\% + 82.2 \cdot 57.7\%} = 1.3721 \quad (6.14)$$

Based on this calculation, price levels rose by 37.2 % from 2001 to 2007assuming 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*:

$$P^{F}_{0,t} = \sqrt{P^{L}_{0,t} \cdot P^{P}_{0,t}} \tag{6.15}$$

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.

## 6.2 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 6.1 shows the quantity relative for the change in diesel sales:

$$Q^{*}_{0,t} = \frac{\text{Quantity in the report year } \left(q_t\right)}{\text{Quantity in the base year } \left(q_0\right)} \tag{6.16}$$

[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).

$$Q^{*}_{2001,2003} = \frac{q_{t=2003}}{q_{t=2001}} = \frac{27,944}{28,545} = 0.98 \tag{6.17}$$

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:

Laspeyres quantity index:

$$Q^{L}_{0,t} = \frac{\sum_{i=1}^{n} q_{i,t} \cdot p_{i,0}}{\sum_{i=1}^{n} q_{i,0} \cdot p_{i,0}} \tag{6.18}$$

Paasche quantity index:

$$Q^{P}_{0,t} = \frac{\sum_{i=1}^{n} q_{i,t} \cdot p_{i,t}}{\sum_{i=1}^{n} q_{i,0} \cdot p_{i,t}} \tag{6.19}$$

Fisher quantity index:

$$Q^{F}_{0,t} = \sqrt{Q^{L}_{0,t} \cdot Q^{P}_{0,t}} \tag{6.20}$$

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 6.1)?

Laspeyres quantity index (constant prices from 2001):

$$Q^{L}_{2001,2007} = \frac{\sum_{i=1}^{n} q_{i,2007} \cdot p_{i,2001}}{\sum_{i=1}^{n} q_{i,2001} \cdot p_{i,2001}} \tag{6.21}$$

$$Q^{L}_{2001,2007} = \frac{15,718 \cdot 102.4 + 5,574 \cdot 100.2 + 29,059 \cdot 82.2}{18,979 \cdot 102.4 + 8,970 \cdot 100.2 + 28,545 \cdot 82.2} = 0.8782 \tag{6.22}$$

Paasche quantity index (constant prices from 2007):

$$Q^{P}_{2001,2007} = \frac{\sum_{i=1}^{n} q_{i,2007} \cdot p_{i,2007}}{\sum_{i=1}^{n} q_{i,2001} \cdot p_{i,2007}} \tag{6.23}$$

$$Q^{P}_{2001,2007} = \frac{15,718 \cdot 134.4 + 5,574 \cdot 132.7 + 29,059 \cdot 117}{18,979 \cdot 134.4 + 8,970 \cdot 132.7 + 28,545 \cdot 117} = 0.8830 \tag{6.24}$$

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.

## 6.3 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:

[7] $W_{0,t} = \frac{\sum_{i=1}^{n} p_{i,t} \cdot q_{i,t}}{\sum_{i=1}^{n} p_{i,0} \cdot q_{i,0}} \neq P^{L}_{0,t} \cdot Q^{L}_{0,t} = \frac{\sum_{i=1}^{n} p_{i,t} \cdot q_{i,0}}{\sum_{i=1}^{n} p_{i,0} \cdot q_{i,0}} \cdot \frac{\sum_{i=1}^{n} q_{i,t} \cdot p_{i,0}}{\sum_{i=1}^{n} q_{i,0} \cdot p_{i,0}}$

$$W_{0,t}=\frac{\sum_{i=1}^{n}p_{i,t}\cdot q_{i,t}}{\sum_{i=1}^{n}p_{i,0}\cdot q_{i,0}}=Q_{0,t}^{F}\cdot P_{0,t}^{F}=Q_{0,t}^{L}\cdot P_{0,t}^{P}=Q_{0,t}^{P}\cdot P_{0,t}^{L} \tag{6.25}$$

According to this equation, fuel sales in 2007 rose by 20.5 % relative to those of 2001. The calculations are as follows:

$$W_{2001,2007}=Q_{2001,2007}^{L}\cdot P_{2001,2007}^{P}=0.8782\cdot 1.3721=1.2050, \tag{6.26}$$

or

$$W_{2001,2007}=Q_{2001,2007}^{P}\cdot P_{2001,2007}^{L}=0.8830\cdot 1.3647=1.2050 \tag{6.27}$$

## 6.4 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 6.2 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 1 increased by 0.5 % (=106.5−106.0) from 2003 to 2004. 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 2004 $L_t^{real}$ [in €], we thus find:

$$L_t^{real}=\frac{L_t^{no\min al}}{P_{0,t}^{L}}\rightarrow L_{2004}^{real}=\frac{L_{2004}^{no\min al}}{P_{0,2004}^{L}}=\frac{1,917.00}{1.055}=€1,817.06 \tag{6.28}$$

In 2003 the real average salary was €1,834.62 per month (see Table 6.2). This means that, in 2004, employees lost some purchasing power compared to 2003. While nominal average monthly salaries between 2000 and 2008 increased by 12.5 %, from €1,800 to €2,025, the real salary in 2008 was only

**Table 6.2** 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] |
| 2000 | 1,800.00 | 100.0 | 100.0 | 1,800.00 | 100.0 | 1,850.00 | 98.3 | 99.0 | 1,868.69 | 99.3 | 100.0 |
| 2001 | 1,854.00 | 103.0 | 102.0 | 1,817.65 | 101.0 | 1,868.50 | 99.3 | 99.7 | 1,874.12 | 99.6 | 100.3 |
| 2002 | 1,845.00 | 102.5 | 103.0 | 1,791.26 | 99.5 | 1,881.45 | 100.0 | 100.0 | 1,881.45 | 100.0 | 100.7 |
| 2003 | 1,908.00 | 106.0 | 104.0 | 1,834.62 | 101.9 | 1,868.50 | 99.3 | 101.0 | 1,850.00 | 98.3 | 99.0 |
| 2004 | 1,917.00 | 106.5 | 105.5 | 1,817.06 | 100.9 | 1,877.75 | 99.8 | 102.5 | 1,831.95 | 97.4 | 98.0 |
| 2005 | 1,926.00 | 107.0 | 106.5 | 1,808.45 | 100.5 | 1,951.75 | 103.7 | 103.0 | 1,894.90 | 100.7 | 101.4 |
| 2006 | 1,962.00 | 109.0 | 108.0 | 1,816.67 | 100.9 | 1,979.50 | 105.2 | 103.0 | 1,921.84 | 102.1 | 102.8 |
| 2007 | 1,998.00 | 111.0 | 109.0 | 1,833.03 | 101.8 | 1,998.00 | 106.2 | 103.5 | 1,930.43 | 102.6 | 103.3 |
| 2008 | 2,025.00 | 112.5 | 109.5 | 1,849.32 | 102.7 | 2,025.75 | 107.7 | 104.0 | 1,947.84 | 103.5 | 104.2 |

Source: Author's research.

$$Y_{2008}^{real} = \frac{2,025.00}{1.095} = €1,849.32 \tag{6.29}$$

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 6.2.

## 6.5 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. 6.1, 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:

$$I_{\tau,t}^{new} = \frac{I_{0,t}^{old}}{I_{0,\tau}^{old}} \tag{6.30}$$

Let us consider the example from Table 6.2. The index for the change of real income values in company 2 is based on 2002 (see the second-to-last column). If we now want to base this index series on the year 2000 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 2000. This produces the final column in Table 6.2.

**Table 6.3** Chain indices for forward and backward extrapolations

| | | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|---|
| 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/ (1.06·1.05)= | 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= | 1.06·1.04= | 1.06·1.05= | 1.06·1.05·1.01= |
| | | 1.05 | 1.06 | 1.10 | 1.11 | 1.12 |

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 ($\tau$). For the *forward extrapolation*, the index with older observations ($I^1$between periods 0 and $\tau$) 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 $\tau$):

Forward extrapolation:

$$\tilde{I}_{0,t} = \begin{cases} I^1_{0,t} & \text{for } t \le \tau \\ I^1_{0,\tau} \cdot I^2_{\tau,t} & \text{for } t > \tau \end{cases} \tag{6.31}$$

With the *backward extrapolation*, the index with the younger observations ($I^2$starting out time$\tau$) 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$\tau$):

Backward extrapolation:

$$\tilde{I}_{0,t} = \begin{cases} \dfrac{I^1_{0,\tau}}{I^2_{\tau,t}} & \text{for } t < \tau \\ I^2_{\tau,t} & \text{for } t \ge \tau \end{cases} \tag{6.32}$$

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 6.3 gives a sample of chain indices for backward and forward extrapolations.

## 6.6 Chapter Exercises

**Exercise 26:**

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 27:**

You are given the following information:

| | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|
| Nominal values | \$100,000 | \$102,000 | \$105,060 | \$110,313 | \$114,726 |
| Nominal values index [2005 = 100] | | | | | |
| Real values | | | | | |
| Real value index [2005 = 100] | | | | | |
| Price index 1 [2004 = 100] | 101.00 | 102.00 | 102.50 | | |
| Price index 2 [2007 = 100] | | | 100.00 | 103.00 | 103.50 |
| Price index 3 [2004 = 100] | | | | | |
| Price index 3 [2004 = 100] | | | | | |

(a) Calculate the nominal value index [2005 = 100].
(b) Chain the price trends for base year 2004.
(c) With the resulting index series, shift the base to 2005.
(d) Calculate the real value changes and the real value index for the base year 2005.

# 7 Cluster Analysis

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*: "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.

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.

T. Cleff, *Exploratory Data Analysis in Business and Economics*,
DOI 10.1007/978-3-319-01517-0_7, © Springer International Publishing Switzerland 2014

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:

1. Hierarchical cluster analysis
2. K-means cluster analysis

The following sections offer a brief introduction to both types of cluster analysis.

## 7.1 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, 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.

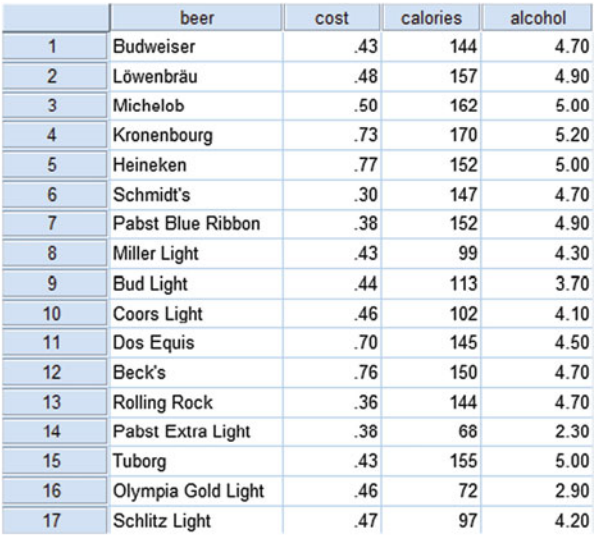
Using a dataset employed by Bühl (2012, p. 627), let's take a look at the methods of hierarchical cluster analysis and the problems associated with them.

Our sample dataset on 17 beers (see Fig. 7.1) 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

[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. 7.1** Beer dataset (Source: Bühl 2012, p. 627)

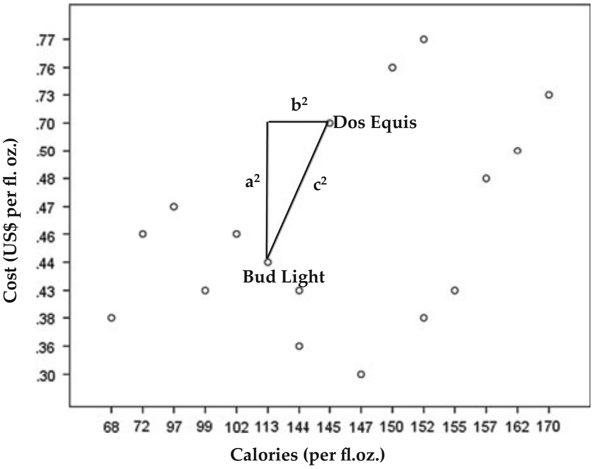
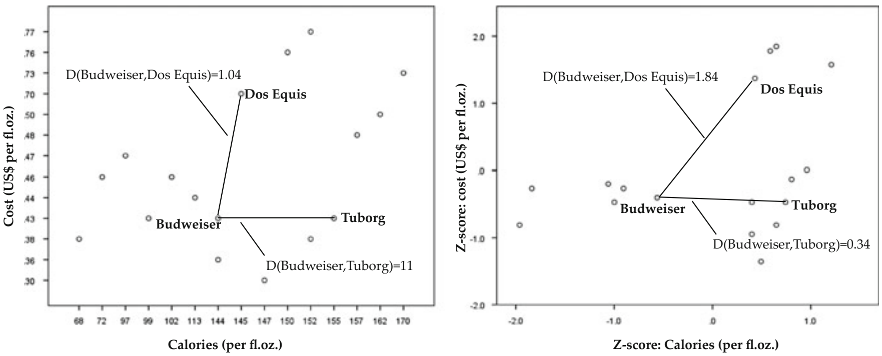


**Fig. 7.2** Distance calculation 1

(see Fig. 7.2). 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:

$$\text{Distance}(\text{Dos Equis, Bud Light}) = \sqrt{a^2 + b^2} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (7.1)$$

$$\text{Distance}(\text{Dos Equis, Bud Light}) = \sqrt{(70 - 44)^2 + (145 - 113)^2} = 41.23 \quad (7.2)$$



Part 1: Non-standardized values

Part 2: Standardized values

**Fig. 7.3** Distance calculation 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 *Euclidian distance* between two observations[2]:

$$\text{Distance}(\text{A}, \text{B}) = \sqrt{\left(x_1^b - x_1^a\right)^2 + \left(x_2^b - x_2^a\right)^2 + \cdots + \left(x_n^b - x_n^a\right)^2} \tag{7.3}$$

Using this information, we can now determine the distances between, say, Tuborg, Dos Equis, and Budweiser.

In part 1 of Fig. 7.3 the distance between Budweiser and Tuborg is 11 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.

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 11 units separating Tuborg and Budweiser – have a stronger impact on the distance than the differences in cost – e.g. the 0.27 units separating Tuborg 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 the same units of measure. If the properties are in different units of measures, the variables must be made "unit free" before being measured. Usually,

[2] In the case of two dimensions, the Euclidean distance and the Pythagorean theorem provide the same results.

| | | |
|---|---|---|
| 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. 7.4** Distance and similarity measures

this is done by applying a *z-transform* to all variables in order to standardize them.[3] 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. 7.3. 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 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 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 7.4 presents the most important distance and similarity measures grouped by scale.

It is important to note that only one proximity measure may be used in a hierarchical cluster analysis. For instance, the chi-square may not be used for some variables and the squared Euclidean distance for others. If two different variable scales exist at the same time, we must find a proximity measure permitted

[3] 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: $z_i = \frac{x_i - \bar{x}}{S}$.

for both. If, say, we have binary and metric variables, we must use the squared Euclidean distance. Backhaus et al. (2008, p. 401) 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.[4]

Let us return again to our beer example. Using the squared Euclidean distance, we obtain the following *distance matrix* (see Fig. 7.5):

After determining the distance between each observation, we aggregate the closest pair into a cluster. These are Heineken (no. 5) and Becks (no. 12), which are separated by 0.009.

The new cluster configuration consists of 15 observations and a cluster containing Heineken and Beck's. 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 (no. 17) and Coors Light (no. 10). The configuration now consists of 13 different data objects and two clusters of two beers each.

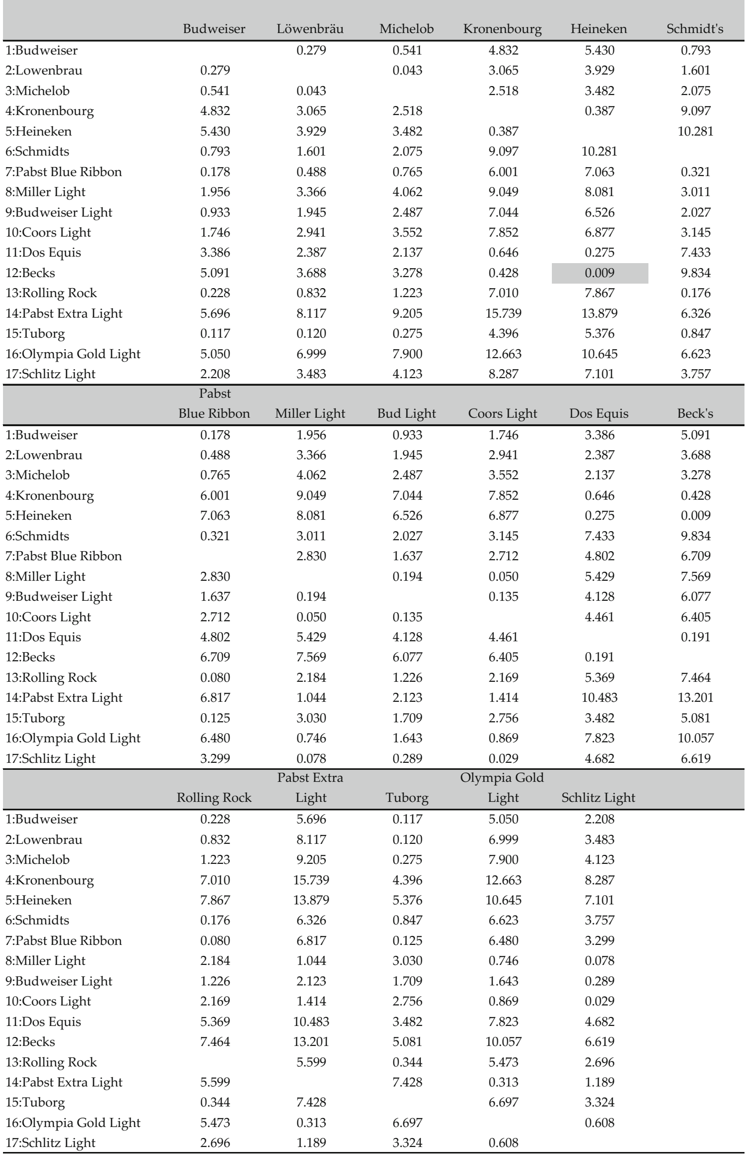
We continue to repeat the distance measurement and linkage steps. We can link beers with other beers, beers with clusters, or clusters with other clusters. Figure 7.6 shows the sequence of steps in the linkage process.

With every step, the heterogeneity of linked objects tends to rise. In the first step, the distance between Heineken and Beck's is only 0.009; by the 10th step, the linkage of Pabst Extra Light (no. 14) and Olympia Gold Light (no. 16) exhibits a distance of 0.313. The sequence of linkage steps and their associated distance values can be taken from the *agglomeration schedule*. For each step, the combined observations are given under *cluster combined* and the linkage distances under *coefficients*. 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 (Fig. 7.7).

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*. This technique tends to form long and snakelike chains of clusters (see Fig. 7.8).
2. The *complete linkage method*, by contrast, uses the furthest two observations of two clusters as the basis for distance measurement. This method generates wide

[4] Say we wanted to dichotomize *calories per fl. oz.* using three calorie variables. Calorie variable 1 assumes the value of 1 when the calories in a beer lie between 60 and 99.99 calories, otherwise it is equal to zero. Calorie variable 2 assumes the value 1 when the calories in a beer lie between 100 and 139.99 calories, otherwise it is equal to zero. Calorie variable 3 assumes the value 1 when the calories in a beer lie between 140 and 200 calories, 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 |

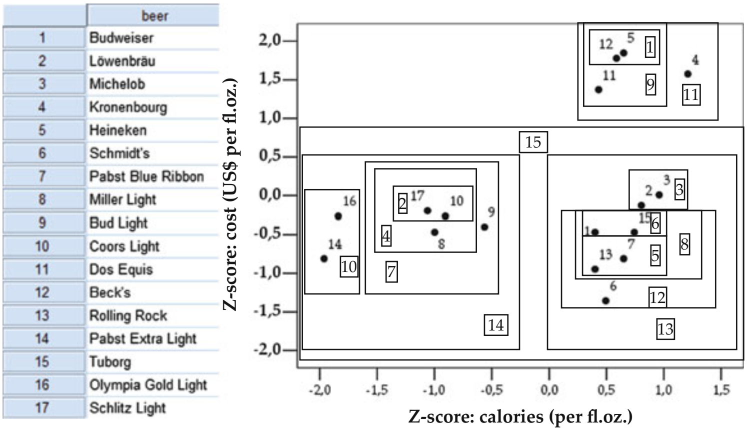
| | Pabst<br>Blue Ribbon | Miller Light | Bud Light | Coors Light | Dos Equis | Beck's |
|---|---|---|---|---|---|---|
| 1:Budweiser | 0.178 | 1.956 | 0.933 | 1.746 | 3.386 | 5.091 |
| 2:Lowenbrau | 0.488 | 3.366 | 1.945 | 2.941 | 2.387 | 3.688 |
| 3:Michelob | 0.765 | 4.062 | 2.487 | 3.552 | 2.137 | 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<br>Light | Tuborg | Olympia Gold<br>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. 7.5** Distance matrix

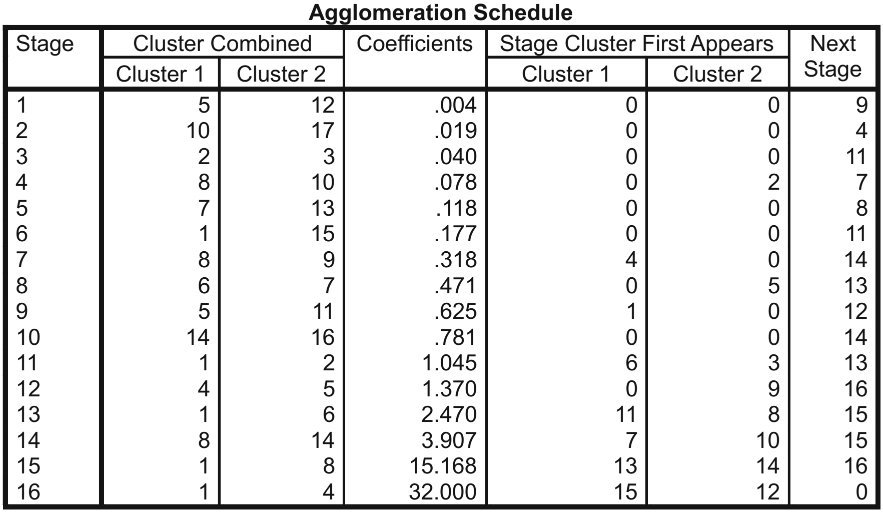
yet compact cluster solutions. This technique may not be used when *elongated* cluster solutions exist in the dataset.

3. The *centroid linkage method* 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.



| | 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. 7.6** Sequence of steps in the linkage process

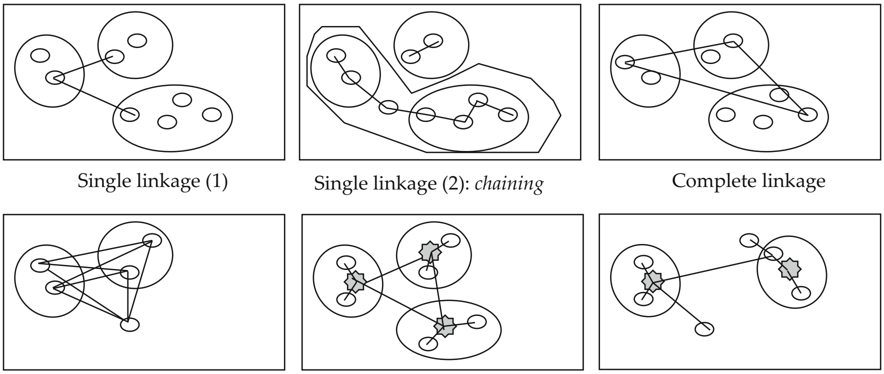


**Agglomeration Schedule**

| 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. 7.7** Agglomeration schedule

4. Centroid linkage should not be confused with the *average linkage* method, which determines the average distance between the observations of two clusters. Generally, this technique forms neither chains nor 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. Wardin 1963) 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



Single linkage (1)

Single linkage (2): *chaining*

Complete linkage

Average linkage

Centroid linkage

Ward linkage

**Fig. 7.8** Linkage methods

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: 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:

(1) The observation values 2 and 4 with a mean of 3,

(2) The observation values 2 and 5 with a mean of 3.5, or

(3) The observation values 4 and 5 with a mean of 4.5.

These clusters yield the following error sums of squares:

(1) QS $= [(2–3)^2 + (4–3)^2] + (5–5)^2 = 1$

(2) QS $= [(2–3.5)^2 + (5–3.5)^2] + (4–4)^2 = 2.25$

(3) QS $= (2–2)^2 + [(4–4.5)^2 + (5–5.5)^2] = 0.25$

The value for the third linkage is the lowest. Its aggregated cluster raises the error sum of squares for all clusters by 0.25, 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 7.8 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" (Mooi and Sarstedt 2011, p. 252). 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 onto 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 clusters 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. 7.7; 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. The length of each branch equals the heterogeneity level of the cluster, with values normalized to a scale between 0 and 25. Reading the following dendrogram of the beer example from left to right, we see that the beer cluster 4, 5, 11, and 12 has a short branch, 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).

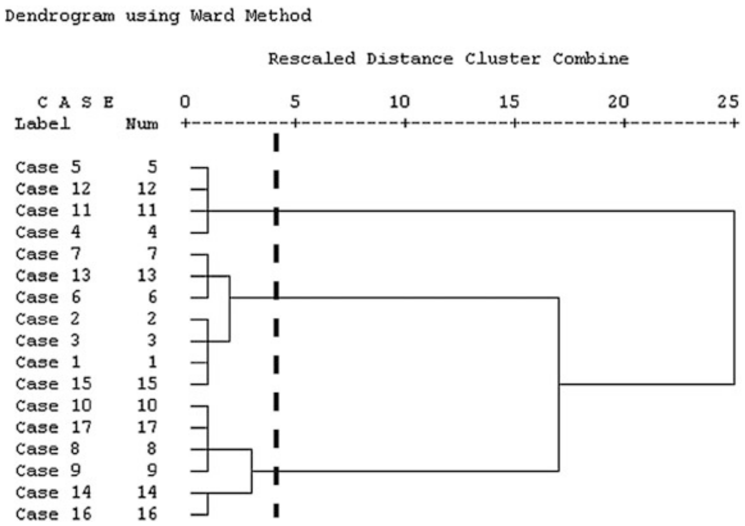
When determining the optimal number of clusters, we proceed as a gardener who begins pruning his tree at the first big branch on the left. This "pruning" is indicated by the dotted line in Fig. 7.9. The number of branches to be pruned corresponds to the number of clusters – in this case, three.

In a *scree plot* the number of clusters are 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 a line that converges asymptotically on the abscissa (Fig. 7.10). Our beer example yields the following scree plot (confirmed by a three-cluster solution):

Though scree plots and dendrograms are frequently used in social and economic research, they do not always yield objective and unambiguous results.

2. The second criterion is obtained by calculating the quotient of the variance 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 1, 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



```
Dendrogram using Ward Method

                         Rescaled Distance Cluster Combine

    C A S E      0         5        10        15        20        25
  Label     Num  +---------+---------+---------+---------+---------+

  Case 5      5
  Case 12    12
  Case 11    11
  Case 4      4
  Case 7      7
  Case 13    13
  Case 6      6
  Case 2      2
  Case 3      3
  Case 1      1
  Case 15    15
  Case 10    10
  Case 17    17
  Case 8      8
  Case 9      9
  Case 14    14
  Case 16    16
```

**Fig. 7.9** Dendrogram

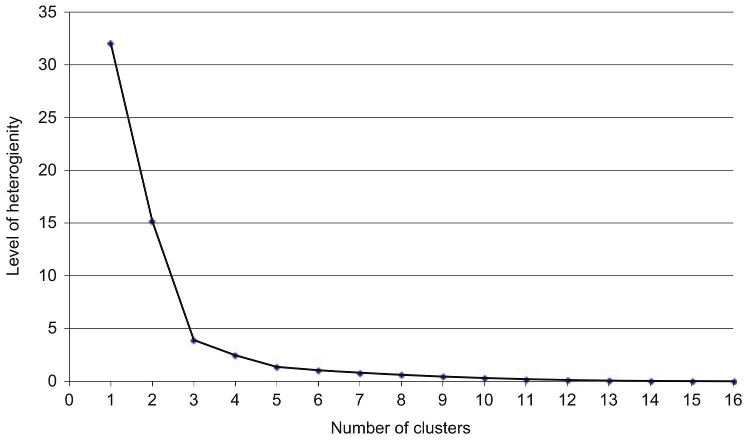
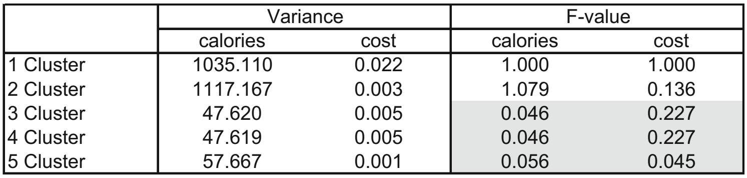


**Fig. 7.10** Scree plot identifying heterogeneity jumps

variances individually. Figure 7.11 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:

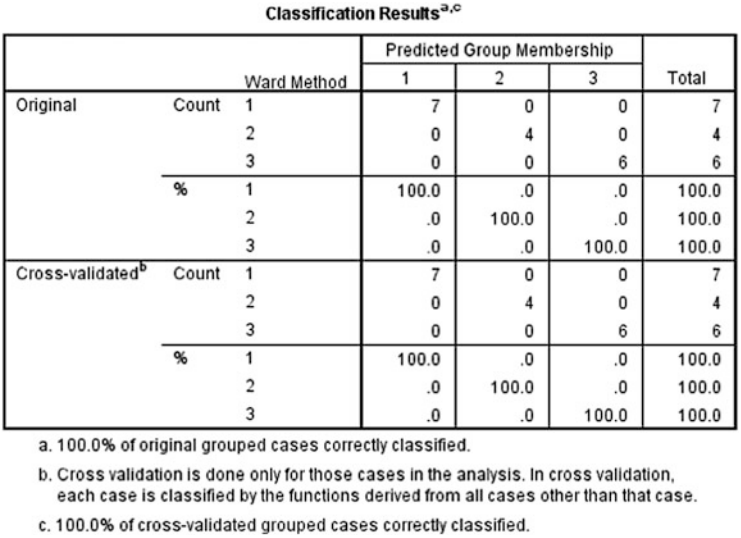
$$F = \frac{1117.167}{1035.110} = 1.079 \tag{7.4}$$

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.



| | 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. 7.11** F-value assessments for cluster solutions 2–5



**Classification Results**[a,c]

| | | Ward Method | Predicted Group Membership | | | Total |
|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | |
| 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.

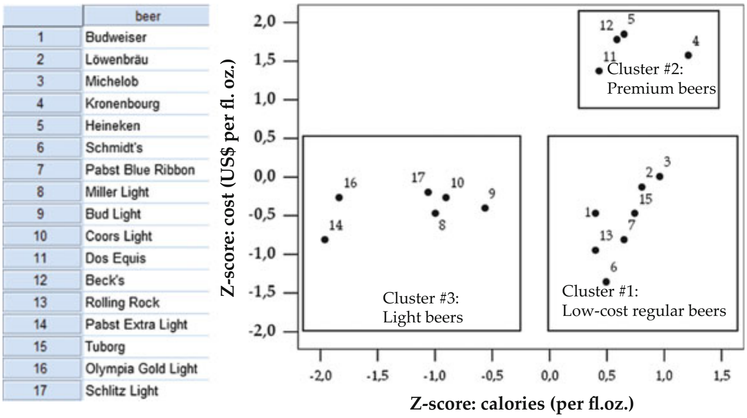
c. 100.0% of cross-validated grouped cases correctly classified.

**Fig. 7.12** Cluster solution and discriminant analysis

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 2 and 5 clusters can be correctly classified using discriminant analysis. A sample result for the three-cluster solution is provided in Fig. 7.12.

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*:

Cluster 3 contains all light beers with a lower-than-average calorie count and a lower-than-average cost. Cluster 1 contains all low-cost regular beers with a higher-than-average calorie count. The premium beers in cluster 2 exhibit both



| | 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. 7.13** Cluster interpretations

higher-than-average costs and higher-than-average calorie counts (Fig. 7.13). 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 Janssens et al. (2008, p. 71) you can learn about how variance analysis (ANOVA) can be used to check group differences for significance.[5]

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 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.

[5] When we apply this technique to the variables in our sample – cluster membership is an independent variable and cost and calories are dependent variables – we ascertain significant differences among the three groups. According to the post-hoc method, premium beers are significantly more expensive than other beers and light beers have significantly fewer calories than other beers. Scheffé's and Tamhane's tests yield similar significance differences.

## 7.2 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 determine 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 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. 7.14. This clustering is different from that provided by hierarchical analysis. Here Bud Light (no. 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 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 (Fig. 7.14), we notice that Bud Light has a distance of 0.790 from its own centroid[7] (cluster #1) and a distance of 0.65 from the centroid of cluster #3.[8] 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. Backhaus et al. (2008, p. 444) prefer 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.

[6] The centroid is determined by calculating the mean for every variable for all observations of each cluster separately.

[7] $(-0.401\text{-}(-0.401))^2 + (-1.353\text{-}(-0.563))^2 = 0.79^2$: Distance: 0.79

[8] $(-0.571\text{-}(-0.401))^2 + (0.486\text{-}(-0.563))^2 = 0.65^2$: Distance: 0.65

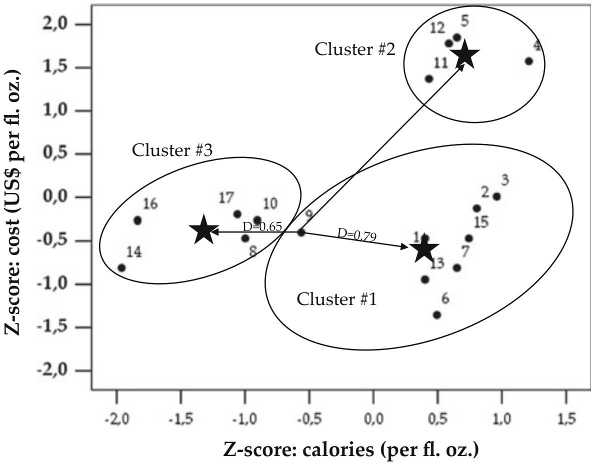


**Fig. 7.14** Initial partition for k-means clustering

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.

## 7.3 Cluster Analysis with SPSS and Stata

This section uses the SPSS and Stata sample datasets *beer.sav* and *beer.dta*. Follow the steps outlined in the following Figs. 7.15, 7.16, and 7.17.

## 7.4 Chapter Exercises

**Exercise 28:**

The share of the population living in urban areas and infant deaths per 1,000 births were collected for 28 European countries. Afterward the data underwent a hierarchical cluster analysis (Fig. 7.18). The results are presented below:

(a) Sketch the basic steps of the cluster analysis method from the agglomeration table.

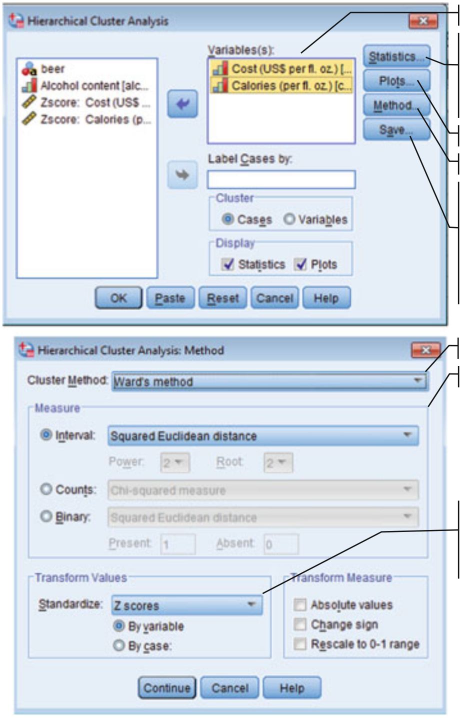
(b) How many clusters make sense from a methodological standpoint? Explain your answer.

**Exercise 29:**

A market research institute studied the relationship between income and personal satisfaction (Fig. 7.19). They used hierarchical clustering to analyse their data:

(a) Assume you decide for a four-cluster solution. Circle the four clusters in the following graphic.

➔ Select *Analyze* → *Classify* → *Hierarchical Cluster…* to open the dialogue box



Hierarchical Cluster Analysis

Variables(s): Cost (US$ per fl. oz.) [... ; Calories (per fl. oz.) [c...

Cluster: ● Cases ○ Variables

Display: ☑ Statistics ☑ Plots

Hierarchical Cluster Analysis: Method

Cluster Method: Ward's method

Measure: ● Interval: Squared Euclidean distance; Power: 2; Root: 2

○ Counts: Chi-squared measure

○ Binary: Squared Euclidean distance; Present: 1; Absent: 0

Transform Values: Standardize: Z scores; ● By variable; ○ By case

Transform Measure: □ Absolute values; □ Change sign; □ Rescale to 0-1 range

Select the variables/items are to be used for cluster analysis.

In the *Statistics…* window: use *Agglomeration schedule* and *Proximity matrix* (indicate first proximity matrix). Cluster Membership displays information about which clusters the observations belong to. It can be set for different numbers of clusters.

In the *Plots…* window: use *Dendrogram*.

The *Methods…* window: see below.

The *Save…* window allows you to save cluster memberships.

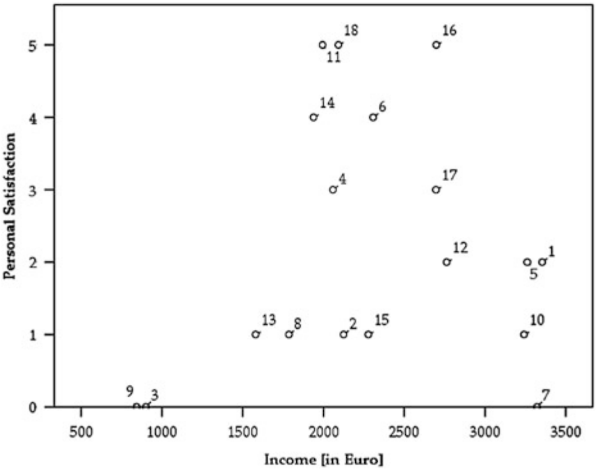
- *None*: no cluster membership will be saved as a variable,
- *Single solution*: generates a new variable indicating cluster membership for a single cluster solution,
- *Range of solutions*: just like *single solution* but for multiple cluster solutions.

Define the cluster method

Determine the proximity measure for the scale.

Determine whether the original values of the distance measure should be used or whether the values should be transformed and how they should be transformed (e.g. z - transform).

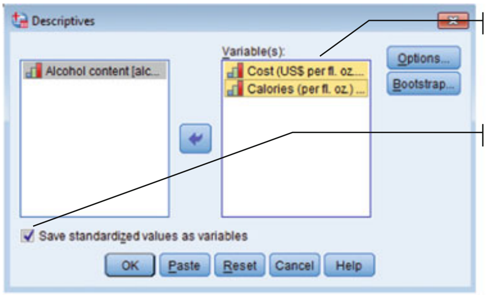
**Fig. 7.15** Hierarchical cluster analysis with SPSS



(b) Characterize the contents of the four clusters.

(c) Assume you decide for a five-cluster solution. Circle the five clusters in the above graphic.

➔ Select *Analyze* → *Descriptive statistics* → *Descriptives*… to standardize variables



Alcohol content [alc...

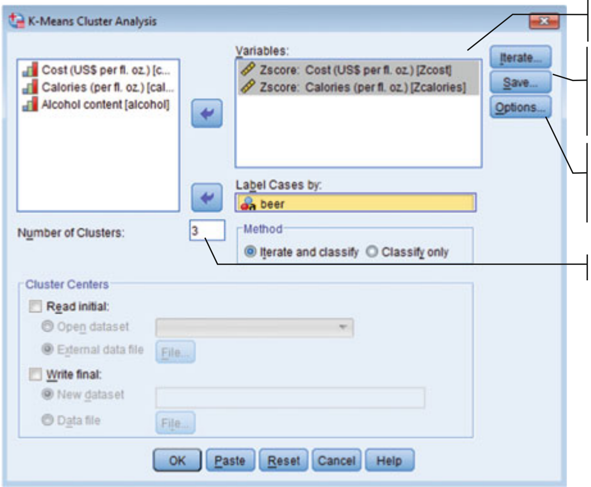
Variable(s):
Cost (US$ per fl. oz....
Calories (per fl. oz.) ...

☑ Save standardized values as variables

Select the variables to be used standardized

Chose the option *Save standardized values as variables*

➔ Then select *Analyze* → *Classify* → *K-Means Cluster*… to obtain the K-means cluster analysis



Cost (US$ per fl. oz.) [c...
Calories (per fl. oz.) [cal...
Alcohol content [alcohol]

Variables:
Zscore: Cost (US$ per fl. oz.) [Zcost]
Zscore: Calories (per fl. oz.) [Zcalories]

Label Cases by:
beer

Number of Clusters: 3

Method
● Iterate and classify ○ Classify only

Cluster Centers
□ Read initial:
○ Open dataset
● External data file
□ Write final:
● New dataset
○ Data file

Select the variables (z-scores) to be used.

The *Save*… window allows you to save cluster membership and a new variable indicating the Euclidean distance between each case and its classification center.

In the *Options*… window you can select the following statistics: initial cluster centers, ANOVA table, and cluster information for each case.

Specify the number of clusters.

**Fig. 7.16** K-means cluster analysis with SPSS

(d) Which approach makes the most sense from a methodological standpoint?

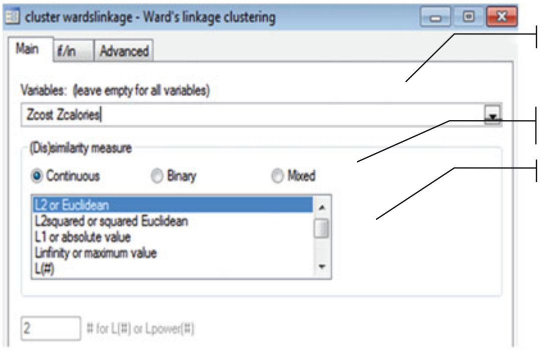
(e) Now, the market research institute used K-means clustering to analyse their data. Please interpret the table *Final Cluster Centres*. What is the difference between the three-cluster-solution of the hierarchical and the K-means approach?

**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* → *Wardś linkage* (or another method)



cluster wardslinkage - Ward's linkage clustering

Variables: (leave empty for all variables)
Zcost Zcalories

(Dis)similarity measure
● Continuous ○ Binary ○ Mixed

L2 or Euclidean
L2squared or squared Euclidean
L1 or absolute value
Linfinity or maximum value
L(#)

2 # for L(#) or Lpower(#)

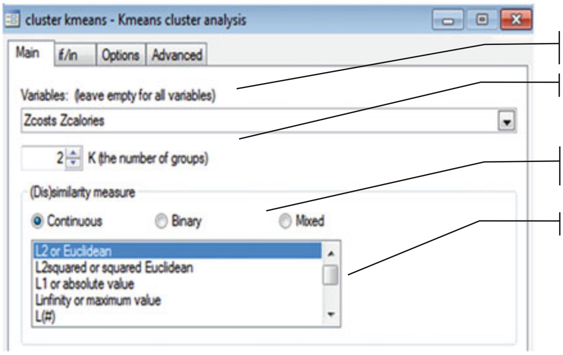
Select the variables (z-scores) to be used.

Select the measurement scale of the selected variables.

Select the (dis)similarity measure.

**K-means cluster analysis commands:**

➔ Select *Statistics* → *Multivariate analysis* → *Cluster analysis* → *Cluster data* → *Kmeans*



cluster kmeans - Kmeans cluster analysis

Variables: (leave empty for all variables)
Zcosts Zcalories

2 K (the number of groups)

(Dis)similarity measure
● Continuous ○ Binary ○ Mixed

L2 or Euclidean
L2squared or squared Euclidean
L1 or absolute value
Linfinity or maximum value
L(#)

Select the variables (z-scores) to be used.

Select the number of clusters.

Select the measurement scale of the selected variables.

Select the (dis)similarity measure.

**Postclustering commands:**

➔ Select *Statistics* → *Multivariate analysis* → *Cluster analysis* → *Postclustering*

- *Dendrogram*: displays dendrogram.
- *Summary variables from cluster analysis*: save a specific cluster solution.

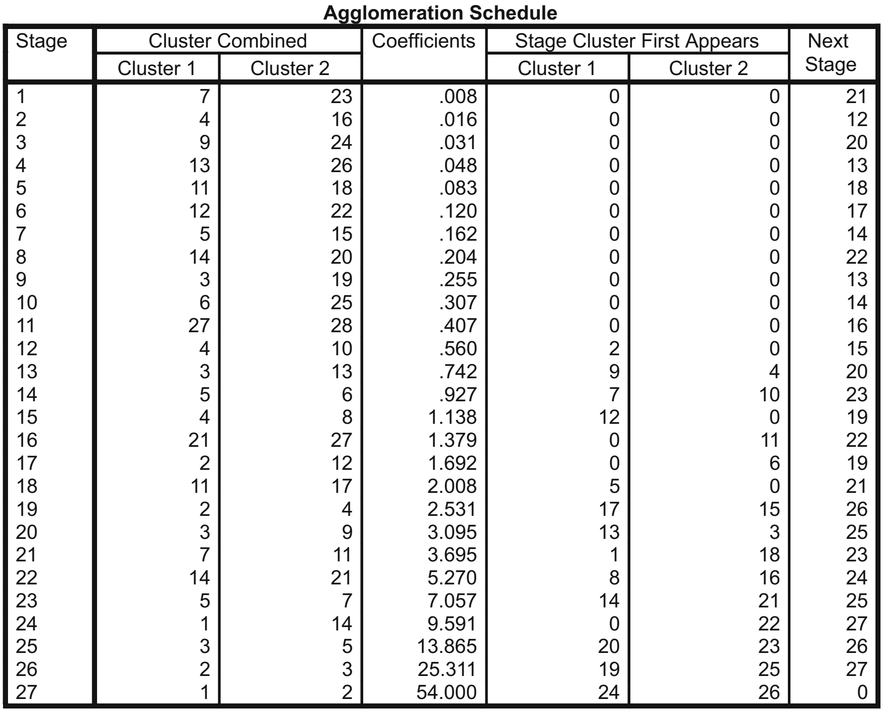
**Important syntax commands for variance analysis:**

cluster averagelinkage; cluster centroidlinkage; cluster completelinkage; cluster wardslinkage; cluster singlelinkage; cluster medianlinkage; cluster waveragelinkage; cluster dendrogramm; cluster generate; cluster kmeans

**Fig. 7.17** Cluster analysis with Stata

| Final cluster centers | | | |
|---|---|---|---|
| | Cluster | | |
| | 1 | 2 | 3 |
| Zscore: Income [in euros] | .81388 | −.04781 | −1.34062 |
| Zscore: Personal satisfaction | −.52984 | 1.08662 | −.97436 |

| 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 |



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. 7.18** Hierarchical cluster analysis (Source: Bühl 2012, p. 627)

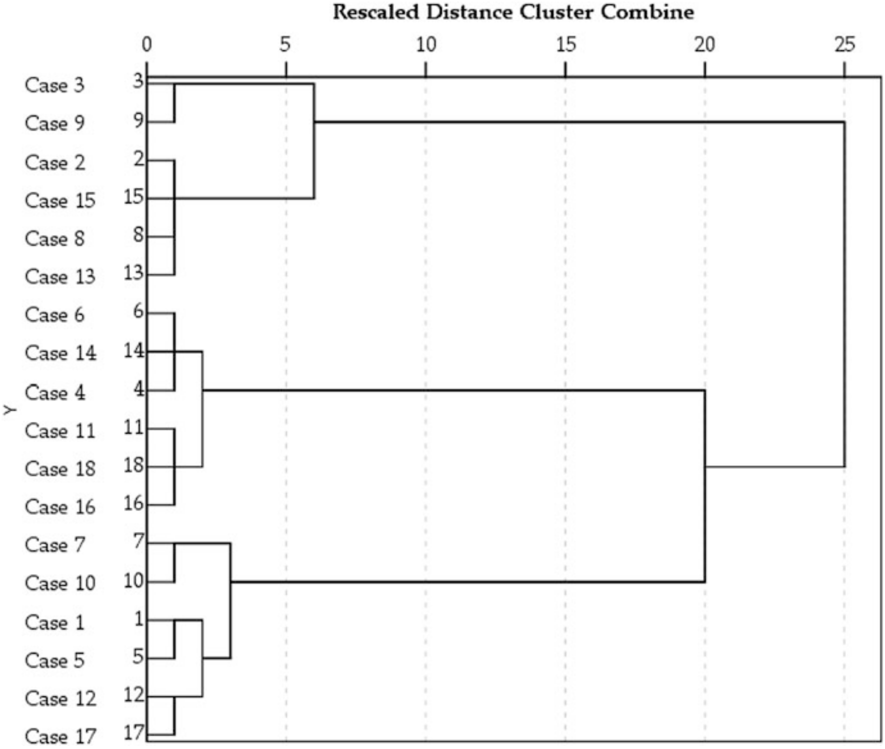


**Fig. 7.19** Dendrogram

# Factor Analysis 8

## 8.1 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 (2010, p. 639). Thirty people were asked the following questions (Fig. 8.1):

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 1 to 7 – are simply added together for each person. One problem with this method occurs when questions are formulated negatively, as with question 5 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.
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 8.2 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.

T. Cleff, *Exploratory Data Analysis in Business and Economics*,
DOI 10.1007/978-3-319-01517-0_8, © Springer International Publishing Switzerland 2014

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1. It is important to buy a toothpaste that prevents cavities. | Disagree completely | □<br>1 | □<br>2 | □<br>3 | □<br>4 | □<br>5 | □<br>6 | □<br>7 | Agree completely |
| 2. I like a toothpaste that gives me shiny teeth. | Disagree completely | □<br>1 | □<br>2 | □<br>3 | □<br>4 | □<br>5 | □<br>6 | □<br>7 | Agree completely |
| 3. A toothpaste should strengthen your gums. | Disagree completely | □<br>1 | □<br>2 | □<br>3 | □<br>4 | □<br>5 | □<br>6 | □<br>7 | Agree completely |
| 4. I prefer a toothpaste that freshens breath. | Disagree completely | □<br>1 | □<br>2 | □<br>3 | □<br>4 | □<br>5 | □<br>6 | □<br>7 | Agree completely |
| 5. Prevention of tooth decay is not an important benefit offered by toothpaste. | Disagree completely | □<br>1 | □<br>2 | □<br>3 | □<br>4 | □<br>5 | □<br>6 | □<br>7 | Agree completely |
| 6. The most important consideration in buying toothpaste is attractive teeth. | Disagree completely | □<br>1 | □<br>2 | □<br>3 | □<br>4 | □<br>5 | □<br>6 | □<br>7 | Agree completely |

**Fig. 8.1** Toothpaste attributes

```
             |   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. 8.2** Correlation matrix of the toothpaste attributes

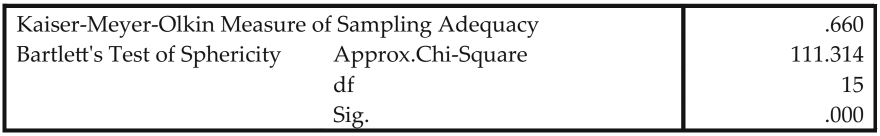
If those surveyed do not show similar patterns in their responses, then the high level of data heterogeneity and low level of data correlation render the results unusable for factor analysis. Backhaus et al. (2008, p. 333) gives five criteria for determining whether the correlation matrix is suitable for running a factor analysis:

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 $\chi 2$ 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.

**Table 8.1** 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. 8.3** Correlation matrix check

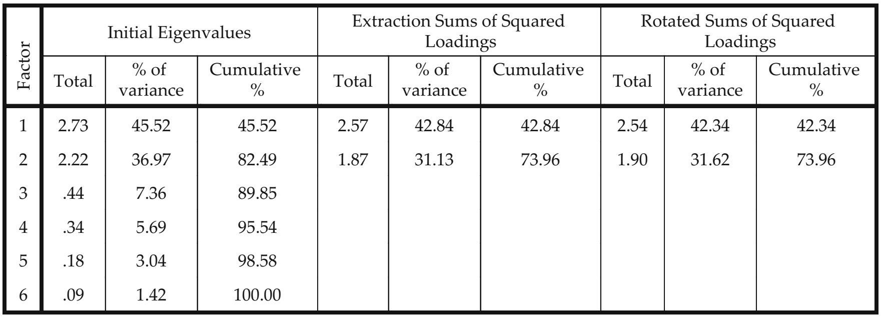
4. A factor analysis should not be performed when, in an anti-image covariance matrix (AIC),[1] more than 25 % of elements below the diagonal have values larger than 0.09.
5. The Kaiser-Meyer-Olkin measure (or KMO measure) is generally considered by researchers to be the best method for testing the suitability 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 above Table 8.1 suggests how KMO might be interpreted:

If the correlation matrix turns out to be suitable for factor analysis, we can assume that regular patterns exist between responses and questions (Fig. 8.3). 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 ($p < 0.05$).

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.

[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.



| 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. 8.4** Eigenvalues and stated total variance for toothpaste attributes

- *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. 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 2.541 and the other with an eigenvalue of 1.897. This results in an importance of 57.26 % for factor 1 and 42.74 % for factor 2. Later I will explain this importance in more detail (Fig. 8.4).

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 matrix. If there are no large deviations ($\leq 0.05$) between the reproduced and the original correlation matrix, then the reproduction– the representability of the original data – is considered very good. Figure 8.5 shows the reproduced correlation

| 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. 8.5** Reproduced correlations and residuals

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.

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. 8.4) 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* toward a less steep decline, all further factors after the one starting the elbow are omitted. The plot in Fig. 8.6 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 8.7 shows the factor loadings for attributes from our toothpaste

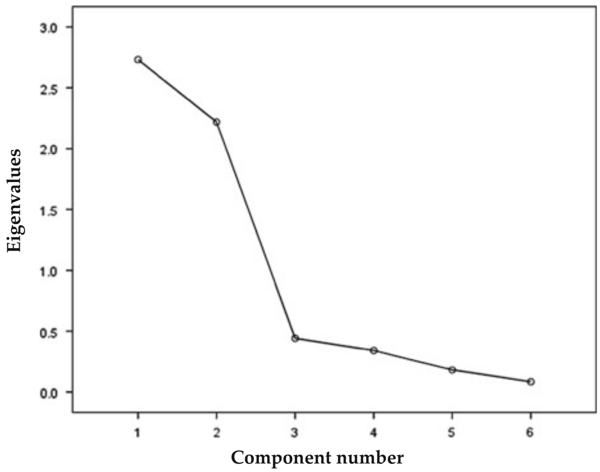


**Fig. 8.6** Screeplot of the desirable toothpaste attributes

| | Unrotated factors | | Rotated factors | |
|---|---|---|---|---|
| Toothpaste should | 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. 8.7** 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.

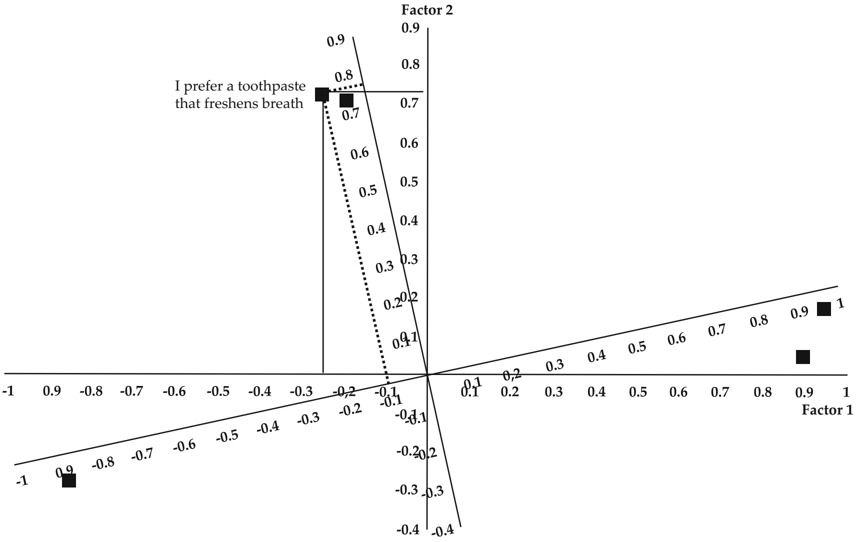
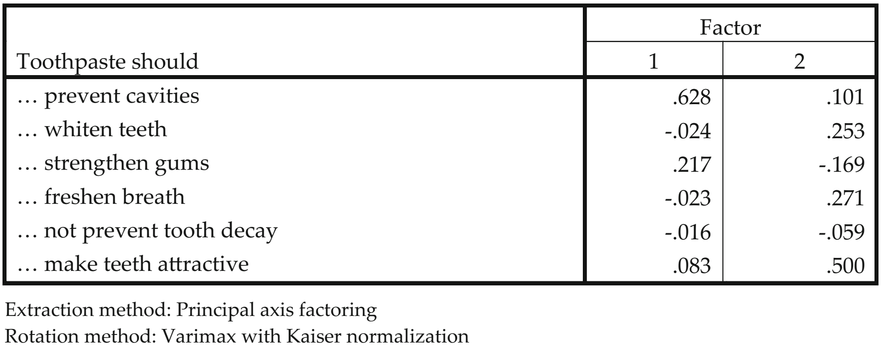


**Fig. 8.8** Varimax rotation for toothpaste attributes

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 8.8 below 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 1 (health attributes) and of 0.734 for factor 2 (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 1 now has the value of −0.090 and factor 2 the value of 0.769 for the item *freshen breath*. The varimax rotation reduces the loading of factor 1 and increases the loading of factor 2, making factor assignments of items more obvious. This is the basic idea of the varimax method: the coordinate system is rotated until the sum of the variances of the squared loadings is maximized. In most cases, this simplifies the interpretation.[2]

[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 | 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. 8.9** Factor score coefficient matrix

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 ($\alpha_{ij}$)from the factor score coefficient matrix (see Fig. 8.9):

$$\mathrm{F_i} = \alpha_{i1} \cdot z_1 + \alpha_{i2} \cdot z_2 + \alpha_{i3} \cdot z_3 + \alpha_{i4} \cdot z_4 + \cdots + \alpha_{in} \cdot z_n \tag{8.1}$$

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 #3[3] has a value of

$$\begin{aligned}\mathrm{F_1} = {} & 0.628 \cdot 1.04 - 0.024 \cdot (-1.38) + 0.217 \cdot 1.41 - 0.023 \cdot (-0.07) \\ & - 0.166 \cdot (-1.31) + 0.083 \cdot (-0.84) = 1.14\end{aligned} \tag{8.2}$$

for factor 1 and a value of

$$\begin{aligned}\mathrm{F_2} = {} & 0.101 \cdot 1.04 + 0.253 \cdot (-1.38) - 0.169 \cdot 1.41 - 0.271 \cdot (-0.07) \\ & - 0.059 \cdot (-1.31) + 0.5 \cdot (-0.84) = (-0.84)\end{aligned} \tag{8.3}$$

for factor 2. This indicates a higher-than-average preference for health benefits and a lower-than-average preference for social benefits.

[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).

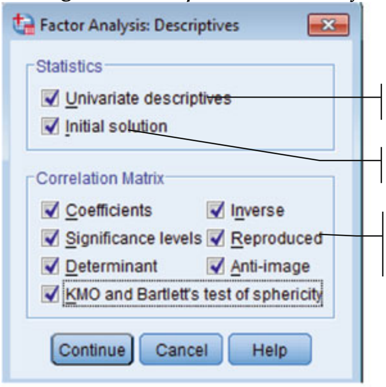
Factor analysis may only be used for metrically scaled variables. Some researchers have described certain conditions under which ordinal scales permit metric variables (Pell 2005; Carifio and Perla 2008). In any case, different item measurements (5-point scale versus 7-point scale, say) require prior standardization. Factor values may only be calculated for persons or observations for which no missing values exist for any of the items being analyzed. However, it is possible to impute missing data, enabling a broad analysis with the complete data set. Different imputation techniques like mean imputation, regression imputation, stochastic imputation, multiple imputation, etc. are recommended in literature (see Enders 2010).

## 8.2 Factor Analysis with SPSS and Stata

This section uses the SPSS and Stata sample datasets *toothpaste_attributes.sav* and *toothpaste_attributes.dta*. For SPSS, 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 factor analysis. Follow then the steps outlined in Fig. 8.10.

For Stata, select *Statistics* → *Multivariate analysis* → *Factor and principal component analysis* → *factor analysis* to open the *Factor Analysis* dialog box. In the menu that opens (*Model*), first select the variables (items) are to be used for factor analysis. Follow then the steps outlined in Fig. 8.11.

**Dialog box *Descriptives…*:** Usually, all options should be selected.



Factor Analysis: Descriptives

Statistics
- ☑ Univariate descriptives
- ☑ Initial solution

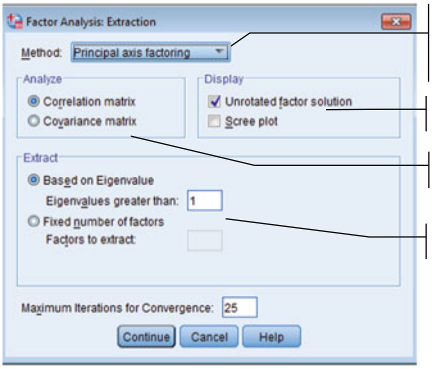
Correlation Matrix
- ☑ Coefficients ☑ Inverse
- ☑ Significance levels ☑ Reproduced
- ☑ Determinant ☑ Anti-image
- ☑ KMO and Bartlett's test of sphericity

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.

**Dialog box *Extraction*:**



Factor Analysis: Extraction

Method: Principal axis factoring

Analyze
- ● Correlation matrix
- ○ Covariance matrix

Display
- ☑ Unrotated factor solution
- □ Scree plot

Extract
- ● Based on Eigenvalue
  - Eigenvalues greater than: 1
- ○ Fixed number of factors
  - Factors to extract:

Maximum Iterations for Convergence: 25

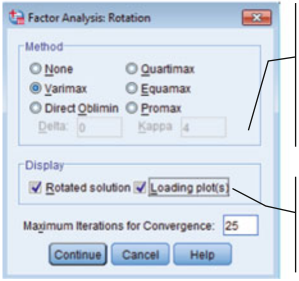
Specify the method of factor extraction. Available methods are principal components, unweighted least squares, generalized least squares, maximum likelihood, principal axis factoring, alpha factoring, and image factoring.

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.

**Dialog boxes *Rotation* and *Scores*:**



Factor Analysis: Rotation

Method
- ○ None ○ Quartimax
- ● Varimax ○ Equamax
- ○ Direct Oblimin ○ Promax
- Delta: 0 Kappa 4

Display
- ☑ Rotated solution ☑ Loading plot(s)

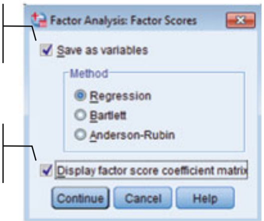
Maximum Iterations for Convergence: 25

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.



Factor Analysis: Factor Scores

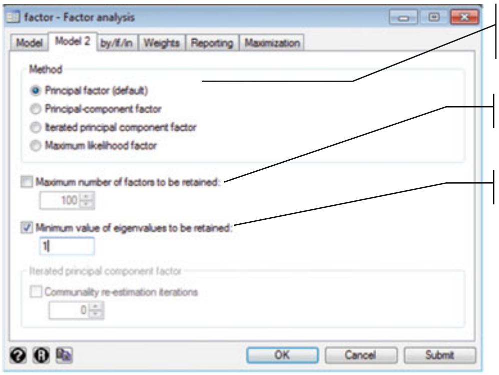
- ☑ Save as variables

Method
- ● Regression
- ○ Bartlett
- ○ Anderson-Rubin

- ☑ Display factor score coefficient matrix

**Fig. 8.10** Factor analysis with SPSS

**Under Model 2:**


factor - Factor analysis

Model | Model 2 | by/if/in | Weights | Reporting | Maximization

Method

● Principal factor (default)

○ Principal-component factor

○ Iterated principal component factor

○ Maximum likelihood factor

□ Maximum number of factors to be retained: 100

☑ Minimum value of eigenvalues to be retained: 1

Iterated principal component factor

□ Communality re-estimation iterations 0

Specify the method: Stata can produce principal factor, iterated principal factor, principal-components factor, and maximum likelihood factor analyses.

Set upper limit for number of factors.

Set the minimum value of eigenvalues to be obtained (should be one).

**Rotation Commands**

- Type in the command line *rotate, varimax* and hit <enter>.

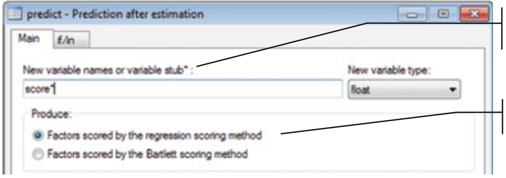
**Reports Commands**

*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.*


predict - Prediction after estimation

Main | if/in

New variable names or variable stub*: score*

New variable type: float

Produce:

● Factors scored by the regression scoring method

○ Factors scored by the Bartlett scoring method

Indicate variable names under which to save the factors (e.g. score*).

Determine how factor scores are produced. For our example, select regression scoring method.

**Important syntax commands for factor analysis:**

factor; factor postestimation; pca; pca postestimation; rotate; rotatemat; scoreplot; screeplot; alpha; canon; estimates, estat; predict

**Fig. 8.11** Factor analysis with Stata

## 8.3 Chapter Exercises

### Exercise 30:

Interpret the results of the following factor analysis about university students.

| KMO and Bartlett's test | | |
|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | | .515 |
| Bartlett's test of sphericity | Approx. Chi-Square | 37.813 |
| | df | 15 |
| | Sig. | .001 |

## Anti-Image Matrices

| | | Intelligence quotient | Independent preparation | Motivation | Self-confidence | Assessment preparation | Contact hours |
|---|---|---|---|---|---|---|---|
| Anti-Image Covariance | Intelligence quotient | .397 | −.100 | −.121 | −.114 | .076 | .052 |
| | Independent preparation (in hours) | −.100 | .191 | .115 | −.095 | −.065 | −.112 |
| | Motivation [1:*very low* to 50:*very high*] | −.121 | .115 | .202 | −.139 | .059 | −.124 |
| | Self-confidence [1:*very low* to 50:*very high*] | −.114 | −.095 | −.139 | .416 | −.017 | .104 |
| | Assessment preparation (in hours) | .076 | −.065 | .059 | −.017 | .391 | −.061 |
| | Contact hours (in hours) | .052 | −.112 | −.124 | .104 | −.061 | .114 |
| Anti-Image Correlation | Intelligence quotient | .643[a] | −.362 | −.427 | −.281[a] | .192 | .246 |
| | Independent preparation (in hours) | −.362 | .487[a] | .584 | −.338 | −.237[a] | −.755 |
| | Motivation [1:*very low* to 50:*very high*] | −.427 | .584 | .385[a] | −.479 | .210 | −.815[a] |
| | Self-confidence [1:*very low* to 50:*very high*] | −.281 | −.338 | −.479 | .536 | −.042 | .475 |
| | Assessment preparation (in hours) | .192 | −.237 | .210 | −.042 | .816 | −.288 |
| | Contact hours (in hours) | .246 | −.755 | −.815 | .475 | −.288 | .450 |

[a] Measures of sampling adequacy(MSA)

Communalities

| | Initial | Extraction |
|---|---|---|
| Intelligence quotient | .603 | .725 |
| Independent preparation (in hours) | .809 | .713 |
| Motivation [1 = very low to 50 = very high] | .798 | .622 |
| Self-confidence [1 = very low to 50 = very high] | .584 | .556 |
| Assessment preparation (in hours) | .609 | .651 |
| Contact hours (in hours) | .886 | .935 |

Extraction method: Principal axis factoring

## Total Variance Explained

| | Initial eigenvalues | | | Extraction sums of squared loadings | | | Rotation sums of squared loadings | | |
|---|---|---|---|---|---|---|---|---|---|
| Factor | 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 | .57 | 9.51 | 89.23 | | | | | | |
| 4 | .34 | 5.74 | 94.97 | | | | | | |
| 5 | .24 | 4.04 | 99.01 | | | | | | |
| 6 | .06 | .99 | 100.00 | | | | | | |

Extraction method: Principal axis factoring

Rotated factor matrix[a]

| | Factor | |
|---|---|---|
| | 1 | 2 |
| Intelligence quotient | −.004 | .851 |
| Independent preparation (in hours) | .839 | .091 |
| Motivation [1 = very low to 50 = very high] | .264 | .743 |
| Self-confidence [1 = very low to 50 = very high] | −.166 | .727 |
| Assessment preparation (in hours) | .759 | −.273 |
| Contact hours (in hours) | .946 | .201 |

Extraction method: Principal axis factoring
Rotation method: Varimax with Kaiser normalization[a]
[a]Rotation converged in 3 iterations

# 9 Solutions to Chapter Exercises

**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 at the book's website.

**Solution 4:**

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

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

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

**Solution 5:**

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 6:**

(a) First sort the dataset, then: $\tilde{x} = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) = \frac{1}{2}\left(x_{(5)} + x_{(6)}\right) = \frac{1}{2}(4+5) = 4.5$

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

T. Cleff, *Exploratory Data Analysis in Business and Economics*,
DOI 10.1007/978-3-319-01517-0_9, © Springer International Publishing Switzerland 2014

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

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

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

(f) Next calculate the lower and upper quartiles.

$\mathrm{x}_{0.25}$: (n + 1)·p = (10 + 1)·0.25 = 2.75 → $\mathrm{x}_{0.25} = (1-\mathrm{f})\cdot\mathrm{x_i} + \mathrm{f}\cdot\mathrm{x_{i+1}} = 0.25\cdot x_2 + 0.75\cdot\mathrm{x}_3 = 0.25\cdot 2 + 0.75\cdot 3 = 2.75$

$\mathrm{x}_{0.75}$: (n + 1)·p = (10 + 1)·0.75 = 8.25 → $\mathrm{x}_{0.75} = 0.75\cdot\mathrm{x}_8 + 0.25\cdot\mathrm{x}_9 = 0.75\cdot 6 + 0.25\cdot 8 = 6.5$.

The interquartile range is $\mathrm{x}_{0.75} - \mathrm{x}_{0.25} = 3.75$.

**Solution 7:**

In the old sample (n = 50) the sum of all observations is $\sum_{i=1}^{50}x_i = n\cdot\bar{x} = 50\cdot 10 = 500$. The new sample has two more observations, for a total sum of $\sum_{i=1}^{52}x_i = 500 + 18 + 28 = 546$. The value for the arithmetic mean is thus $\bar{x}_{new} = \frac{\sum_{i=1}^{52}x_i}{50+2} = \frac{546}{52} = 10.5$. To calculate empirical variance, the following generally applies:

$S^2_{emp} = \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2\right) - \bar{x}^2$. For the original sample n = 50, $S^2_{emp_{old}} = 4 = \frac{1}{50}\left(\sum_{i=1}^{50}x_i^2\right) - 10^2$ applies, producing the following sum of squares $\sum_{i=1}^{50}x_i^2 = 50\cdot\left(4 + 10^2\right) = 5,200$. From this we can determine the empirical variance of the new sample:

$$S^2_{emp_{new}} = \frac{1}{n+2}\left(\sum_{i=1}^{n}x_i^2 + x_{51}^2 + x_{52}^2\right) - \bar{x}^2_{new} = \frac{1}{52}\left(5{,}200 + 18^2 + 28^2\right) - 10.5^2 = 11.06.$$

To determine the empirical standard deviation, we must extract the root from the result, for $S_{emp_{new}} = 3.33$.

**Solution 8:**

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

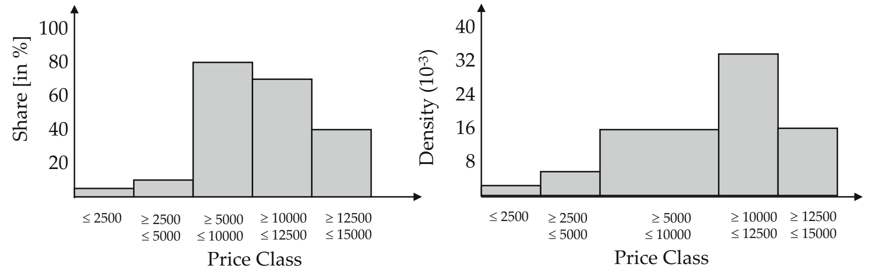
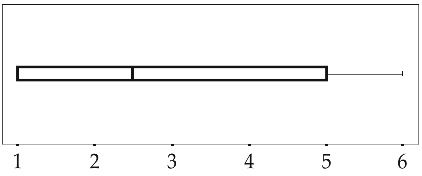


**Fig. 9.1** Bar graph and histogram



(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) $\bar{x}\ geom = \sqrt[3]{(1+0.02)(1+0.04)(1+0.01)} - 1 =$ 2.3 %

**Solution 9:**

(a) The middle price class is twice as large as the other price classes. A bar graph (see the figure on the left) would be misleading here because the €5,000 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 are as follows: 2/2,500 = 0.0008; 8/2,500 = 0.0032; 80/5,000 = 0.016; 70/2,500 = 0.028; 40/2,500 = 0.016 (Fig. 9.1).

(b) The mean can be calculated from the class midpoint: $\bar{x}$ = €9,850; 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 + 2,500·5/35 = €10,357.14; modal class: €10,000 – 12,500.

(c) $x_{0.55}$ = 10,000 + 2,500·(5 + 5)/35 = 10,714.28;

(d) $x_{0.2}$ = 5,000 + 5,000·(15)/40 = €6,875. The cars on the other used market are more expensive on average.

**Solution 10:**

The question is about growth rates. Here the geometric mean should be applied. $\bar{x}_{geom} = \sqrt[4]{(1+0.04)(1+0.03)(1+0.02)(1+0.01)} - 1 = 0.024939 = 2.49\%$

**Solution 11:**

$\mathrm{CR}_2 = 76.67\ \%$

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

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

$$\text{GINI}_{\text{norm.}} = \tfrac{\text{n}}{\text{n}-1} \cdot \text{GINI} = 0.27$$

**Solution 12:**

(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)

$$\chi^2 = \frac{(130-89.25)^2}{89.25} + \frac{(30-26.25)^2}{26.25} + \ldots. + \frac{(130-85.5)^2}{85.5} = 84.41$$

(d)

$$V = \sqrt{\frac{\chi^2}{n \cdot (Min(number\ \ of\ \ columns,\ number\ \ of\ \ rows) - 1)}} = \sqrt{\frac{84.41}{400 \cdot 1}} = 0.46$$

**Solution 13:**

(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 |
| 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 3 rows are added together due to their sparseness, we get: $\chi^2 = 0 + 4 + 0.44 + 0 + 1.45 + 0.16 = 6.06$.

(d) $V = \sqrt{\frac{\chi^2}{n\cdot(\ Min(\ number\ \ of\ \ columns,\ number\ \ of\ \ rows)-1)}} = \sqrt{\frac{9.77}{300\cdot2}} = 0.1276$. If the last 3 rows are added together due to their sparseness, we get: V = $\sqrt{\frac{6.06}{300\cdot1}} = 0.142$

(e) Phi is only permitted with two rows or two columns.

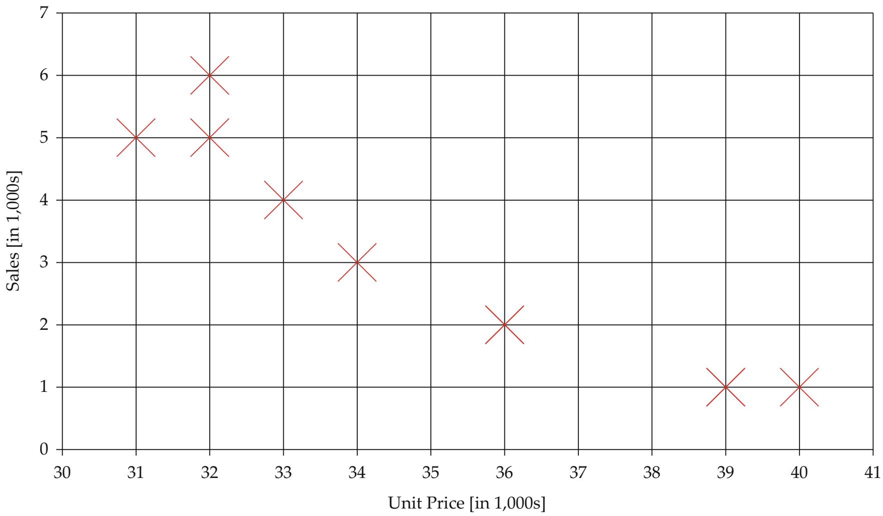
**Solution 14:**

(a) f(Region = Region3|assessment = good) = 2/15·100% = 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 15:**

(a) Y: Sales; X: Price [in 1,000 s]



(b)

| Country | Sales [in 1,000s] | Unit price [in 1,000s] | Sales$^2$ [in 1,000s] | Unit price$^2$ [in 1,000s] | Sales Price | R (Sales) | R (Price) | $d_i$ | $d_i^2$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 6 | 32 | 36 | 1,024.00 | 192.00 | 10 | 2.5 | 7.5 | 56.25 |
| 2 | 4 | 33 | 16 | 1,089.00 | 132.00 | 7 | 4 | 3 | 9 |
| 3 | 3 | 34 | 9 | 1,156.00 | 102.00 | 6 | 5 | 1 | 1 |
| 4 | 5 | 32 | 25 | 1,024.00 | 160.00 | 8.5 | 2.2 | 6 | 36 |
| 5 | 2 | 36 | 4 | 1,296.00 | 72.00 | 4.5 | 6.5 | −2 | 4 |
| 6 | 2 | 36 | 4 | 1,296.00 | 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 | 1,521.00 | 39.00 | 2 | 8.5 | −6.5 | 42.25 |
| 9 | 1 | 40 | 1 | 1,600.00 | 40.00 | 2 | 10 | −8 | 64 |
| 10 | 1 | 39 | 1 | 1,521.00 | 39.00 | 2 | 8.5 | −6.5 | 42.25 |
| Sum | 30 | 352 | 122 | 12,488.00 | 1,003.00 | 55 | 55 | 0 | 315 |
| Mean | 3.0 | 35.2 | 12.2 | 1,248.80 | 100.30 | 5.5 | 5.5 | 0.0 | 31.5 |

*Unit price [in 1,000 s of MUs]:*

$$\overline{x} = \frac{1}{10}(32 + 33 + 34 + \cdots + 39) = 35.2$$

$$S_{emp} = \sqrt{\frac{(x_i - \overline{x})^2}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2} = \sqrt{\frac{1}{10}12,488 - 35.2^2} = \sqrt{9.76} = 3.12$$

*Sales:*

$$\overline{y} = \frac{1}{10}(6 + 4 + 3 + \cdots + 1) = 3.0$$

$$S_{emp} = \sqrt{\frac{(y_i - \overline{y})^2}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2} = \sqrt{\frac{1}{10}122 - 3^2} = \sqrt{3.2} = 1.79$$

*Covariance:*

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i - \overline{x} \cdot \overline{y} = \frac{1}{10}(6 \cdot 32 + \cdots + 1 \cdot 39) - 35.2 \cdot 3 = 100.3 - 105.6$$
$$= -5.3$$

(c) $r = \frac{S_{xy}}{S_x S_y} = \frac{-5.3}{1.79 \cdot 3.12} = -0.95$

(d) $\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot \left(7.5^2 + 3^2 + \cdots + \left(-6.5^2\right)\right)}{10 \cdot \left(10^2 - 1\right)} = 1 - \frac{6 \cdot 315}{10 \cdot \left(10^2 - 1\right)} = -0.909$. 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 16:**

(a)

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{-309}{14} = -22.07$$

(b)

$$S_{emp} = \sqrt{\frac{\sum_{i=1}^{n} y_i^2}{n} - \overline{y}^2} = \sqrt{\frac{10,545}{14} - 22.07^2} = \sqrt{266.129} = 16.31$$

(c) Coefficient of Variation $\frac{S_{emp}}{|\bar{y}|} = \frac{16.31}{|-22.07|} = 0.74$

(d)

$$S^2_{emp} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} = \frac{3,042.36}{14} = 217.31$$

(e)

$$S_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n} = 213.42$$

(f)

$$r = \frac{S_{xy}}{S_x \cdot S_y} = 0.89$$

(g)

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot 54}{14 \cdot \left(14^2 - 1\right)} = 0.88$$

**Solution 17:**

(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 18:**

(a) Using the table we can calculate the following: $\frac{1}{5}\sum_{i=1}^{5} (x_i - \bar{x})(y_i - \bar{y}) = 2,971.6$.

Pearson's correlation is then: $r = \frac{2,971.6}{432.96 \cdot 7.49} = 0.916$. 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 19:**

(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: $\mathrm{r}_{\mathrm{xy.z}}=\frac{\mathrm{r}_{\mathrm{xy}}-\mathrm{r}_{\mathrm{xz}}\mathrm{r}_{\mathrm{yz}}}{\sqrt{\left(1-\mathrm{r}_{\mathrm{xz}}^2\right)\cdot\left(1-\mathrm{r}_{\mathrm{yz}}^2\right)}}$.

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:

$$\mathrm{r}_{\mathrm{xz.y}}=\frac{\mathrm{r}_{\mathrm{xz}}-\mathrm{r}_{\mathrm{xy}}\mathrm{r}_{\mathrm{zy}}}{\sqrt{\left(1-\mathrm{r}_{\mathrm{xy}}^2\right)\cdot\left(1-\mathrm{r}_{\mathrm{zy}}^2\right)}}=\frac{0.3347-((-0.7383)\cdot(-0.4624))}{\sqrt{\left(1-(-0.7383)^2\right)\cdot\left(1-(-0.4624)^2\right)}}=-0.011$$

(c) The association in (a) is a spurious correlation. In reality there is no association between toilet paper price and potato chip sales.

**Solution 20:**

$$r_{pb}=\frac{\overline{y}_1-\overline{y}_0}{S_y}\sqrt{\frac{n_0\cdot n_1}{n^2}}=\frac{0.41-0.37}{0.095}\sqrt{\frac{2,427\cdot 21,753}{24,180^2}}=0.127$$

**Solution 21:**

(a) Market share = 1.26 − 0.298 · Price = 1.26 − 0.298 · 3 = 36.6 %.

(b) $0.40=1.26-0.298\cdot\text{price}\Leftrightarrow\text{price}=\frac{0.40-1.26}{-0.298}=€2.89$

(c) 42% of the variance in market share is explained by variance in the independent variable price.

(d)

$$R^2=1-\frac{ESS}{TSS}\Longleftrightarrow\text{TSS}=\frac{\text{ESS}}{1-\text{R}^2}=\frac{0.08}{0.58}=0.14$$

**Solution 22:**

(a) $\widehat{y}=24.346+0.253\cdot x_1-0.647\cdot x_2-0.005\cdot x_3$, where:

$\mathrm{x}_1$: number of locations;

$\mathrm{x}_2$: item price [in 1,000 s of MUs];

$\mathrm{x}_3$: advertising budget [in 100,000 s of MUs]

The low (insignificant) influence of advertising budget would, in practice, eliminate the variable $\mathrm{x}_3$ from the regression (see part d) of the exercise, yielding the following result: $\widehat{y}=24.346+0.253\cdot x_1-0.647\cdot x_2$

(b) We already know the coefficient of determination: $\mathrm{R}^2=0.951$.

(c) The regression coefficient for the item price is $\alpha_2=-0.647$. Since the item price is measured by 1,000 s of units, a price decrease of 1,000 MUs affects sales as follows: Δsales = (−1) · (−0.647) = 0.647. Sales is also measured by 1,000 s of units, which means that total sales increase by 1,000 ·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: Δsales = (+1)·

(−0.005) = (−0.005). Sales are measured in 1,000 s of units, which means they will sink by 1,000 · (−0,005) = (−5). The result arises because the variable *advertising budget* is an insignificant influence (close to 0); advertising appears to play no role in determining sales.

**Solution 23:**

(a) $\widehat{y} = 38.172 - 7.171 \cdot x_1 + 0.141 \cdot x_2$, where:
$x_1$: price of the company's product;
$x_2$: price of the competition's product put through the logarithmic function.
The low (insignificant) influence of the competition's price put through the logarithmic function would, in practice, eliminate the variable $x_2$ from the regression (see part e) of the exercise), yielding the following result: $\widehat{y} = 38.172 - 7.171 \cdot x_1$

(b)

$$R^2 = \frac{Explained\ \ Sum\ \ of\ \ Squares\ (RSS)}{Total\ \ Sum\ \ of\ \ Squares\ (TSS)} = \frac{124.265}{134.481} = 0.924;$$

$$R^2_{adj} = 1 - \left(1 - R^2\right)\frac{n-1}{n-k} = 1 - (1 - 0.924)\frac{27-1}{27-3} = 0.918$$

(c) RSS + ESS = TSS ⇔ ESS = TSS – 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)·(−7,171) = −7.171 percentage points.

**Solution 24:**

(a) $\widehat{y} = 9898 - 949.5 \cdot \text{price} + 338.6 \cdot HZ_{sw} - 501.4 \cdot HZ_{az} - 404.1 \cdot TZ_{az} + 245.8 \cdot TZ_{sw} + 286.2 \cdot HZ_{hz\_abb}$

(b) $\widehat{y} = 9898 - 949.5 \cdot 2.5 + 338.6 \cdot 0 - 501.4 \cdot 1 - 404.1 \cdot 0 + 245.8 \cdot 0 + 286.2 \cdot 0 \approx 7023$

(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 25:**

(a) $R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS} = 1 - \frac{34,515,190,843.303}{136,636,463,021.389} = 0.7474$

(b) In order to compare regressions with varying numbers of independent variables.

(c) Average proceeds = 25,949.5 + 5·4,032.79 – 7,611.182 + 6,079.44 = 44,581.752 MU

(d) Lettuce, because the standardized beta value has the second largest value.

(e) The price and size of the beverage in regression 2 show a high VIF value, i.e. a low tolerance. In addition, the $R^2$ of regression 1 to regression 2 has barely increased. The independent variables in regression 2 are multicollinear, distorting significances and coefficients. The decision impinges on regression 1.

(f) No linear association exists. As a result, systematic errors occur in certain areas of the x-axis in the linear regression. The residuals are auto correlated. The systematic distortion can be eliminated by using a logarithmic function or by inserting a quadratic term.

**Solution 26:**

| Good | Price1 | Quantity 1 | Price 3 | Quantity 3 | $p_3 \cdot q_1$ | $p_1 \cdot q_1$ | $p_3 \cdot q_3$ | $p_1 \cdot q_3$ |
|---|---|---|---|---|---|---|---|---|
| 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)

$$P^L_{1,3} = \frac{\sum_{i=1}^{4} p_{i,3} \cdot q_{i,1}}{\sum_{i=1}^{4} p_{i,1} \cdot q_{i,1}} = \frac{(8 \cdot 22) + (28 \cdot 4) + (13 \cdot 7) + (42 \cdot 3)}{(6 \cdot 22) + (27 \cdot 4) + (14 \cdot 7) + (35 \cdot 3)} = \frac{505}{443} = 1.14$$

$$Q^L_{1,3} = \frac{\sum_{i=1}^{4} q_{i,3} \cdot p_{i,1}}{\sum_{i=1}^{4} q_{i,1} \cdot p_{i,1}} = \frac{(23 \cdot 6) + (5 \cdot 27) + (10 \cdot 14) + (3 \cdot 35)}{(22 \cdot 6) + (4 \cdot 27) + (7 \cdot 14) + (3 \cdot 35)} = \frac{518}{443} = 1.17$$

The inflation rate between the two years is 14 %. During the same period, sales of the 4 goods assessed with the prices of the first year increased by 17 %.

(b)

$$P^P_{1,3} = \frac{\sum_{i=1}^{n} p_{i,3} \cdot q_{i,3}}{\sum_{i=1}^{4} p_{i,1} \cdot q_{i,3}} = \frac{(8 \cdot 23) + (28 \cdot 5) + (13 \cdot 10) + (42 \cdot 3)}{(6 \cdot 23) + (27 \cdot 5) + (14 \cdot 10) + (35 \cdot 3)} = \frac{580}{518} = 1.12$$

$$Q^P_{1,3} = \frac{\sum_{i=1}^{4} q_{i,3} \cdot p_{i,3}}{\sum_{i=1}^{4} q_{i,1} \cdot p_{i,3}} = \frac{(23 \cdot 8) + (5 \cdot 28) + (10 \cdot 13) + (3 \cdot 42)}{(22 \cdot 8) + (4 \cdot 28) + (7 \cdot 13) + (3 \cdot 42)} = \frac{580}{505} = 1.15$$

The inflation rate between the two years is 12 %. During the same period, sales of the 4 goods assessed with the prices of the third year increased by 15 %.

(c) 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).

(d)

$$P^F_{1,3} = \sqrt{P^L_{1,3} \cdot P^P_{1,3}} = \sqrt{1.14 \cdot 1.12} = 1.13$$

$$Q^F_{1,3} = \sqrt{Q^L_{1,3} \cdot Q^P_{1,3}} = \sqrt{1.17 \cdot 1.15} = 1.16$$

(e) $W_{1,3} = Q_{1,3}{}^F \cdot P_{1,3}{}^F = 1.16 \cdot 1.13 = Q_{1,3}{}^L \cdot P_{1,3}{}^P = 1.17 \cdot 1.12 = Q_{1,3}{}^P \cdot P_{1,3}{}^L = 1.15 \cdot 1.14 = 1.31$ The sales growth in the third year is 31 % more than the first year.

(f)

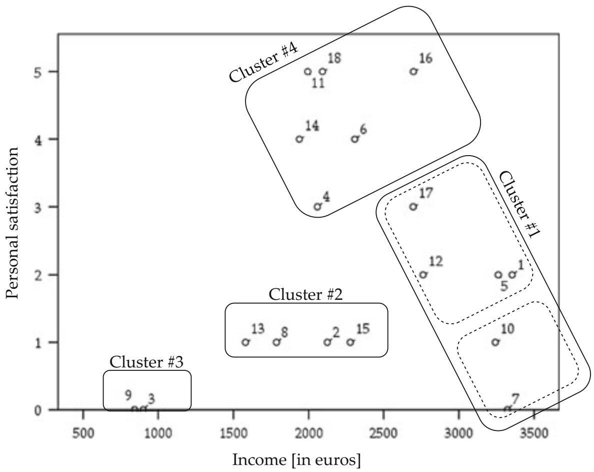
$$\bar{p}_{geom} = \sqrt[n]{\prod_{i=1}^{n} (1 + p_i)} - 1 = \sqrt[2]{(1 + 0.14)} - 1 = 0.0677 \rightarrow 6.77\ \%\ \text{price rate}$$

increase.

**Solution 27:**

| | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|
| Nominal Values | \$100,000 | \$102,00 | \$105,060 | \$110,313 | \$114,726 |
| Nominal Value Index [2005=100] | 100.00 | 102.00 | 105.06 | 110.31 | 114.73 |
| Real Values | \$100,00 | \$101,00 | \$103,523 | \$105,533 | \$109.224 |
| Real Value Index [2500=100] | 100.00 | 101.00 | 103.52 | 105.53 | 109.22 |
| Price Index 1 [2004=100] | 101.00 | 102.00 | 102.50 | | |
| Price Index 2 [2007=100] | | | 100.00 | 103.00 | 103.50 |
| Price Index 3 [2004=100] | 101.00 | 102.00 | 102.50 | 105.58 | 106.09 |
| Price Index 4 [2005=100] | 101.00 | 100.99 | 101.49 | 104.53 | 105.04 |

**Fig. 9.2** Cluster analysis (1)



Example calculations:

- Nominal value index [2005 = 100] for 2007: $W^{\text{nominal}}_{2005,2007} = \frac{\$105{,}060}{\$100{,}000} \cdot 100 = 105.06$
- Price index [2004 = 100] for 2008:

$$\tilde{P}_{2004,2008} = P_{2004,2007} \cdot P_{2007,2008} = 102.50 \cdot 103.00 = 105.58$$

- Shifting the base of the price index [2004 = 100] to[2005 = 100] for 2008:

$$\tilde{P}^{[2005=100]}_{2005,2008} = \frac{P^{[2004=100]}_{2004,2008}}{P^{[2004=100]}_{2004,2005}} = \frac{105.58}{101.00} \cdot 100 = 104.53$$

- Real value change for 2008: $W^{\text{real}}_{2008} = \frac{W^{\text{no min al}}_{2008}}{\tilde{P}^{[2005=100]}_{2005,2008}} = \frac{110{,}313}{1.0453} = \$105,533$
- Real value index [2005 = 100] for 2008: $W^{\text{no min al}}_{2005,2008} = \frac{\$105{,}533}{\$100{,}000} \cdot 100 = 105.53$

**Solution 28:**

(a) 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.

(b) A four-cluster-solution makes sense, since further linkage raises heterogeneity levels excessively. The last heterogeneity jump rises from 9.591 to 13.865.

**Solution 29:**

(a) Figure 9.2

(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.

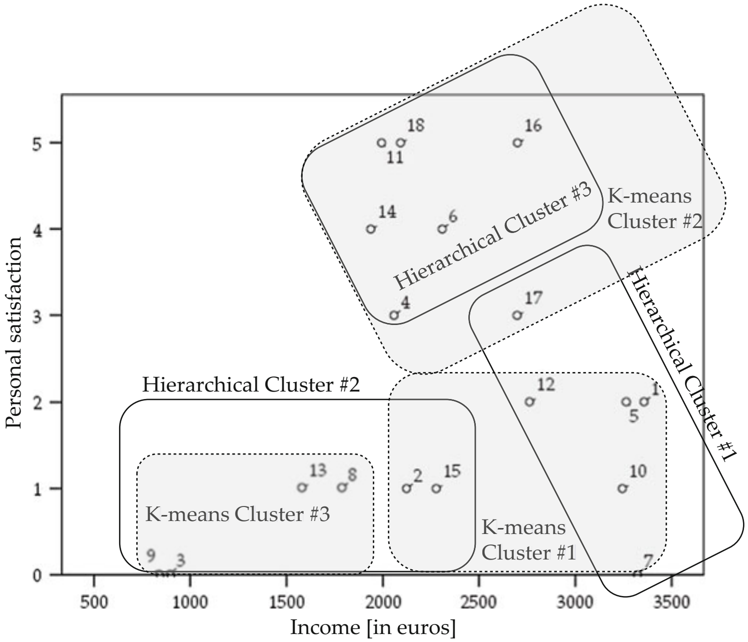


**Fig. 9.3** Cluster analysis (2)

(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) Figure 9.3

**Solution 30:**

- 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: Component 1 and 2 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

ACEA. *European Automobile Manufacturers' Association*. http://www.acea.be/index.php/collection/statistics

Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2008). *Multivariate analysemethoden. Eine Anwendungsorientierte Einführung* (12th ed.). Berlin, Heidelberg: Springer.

Berg, S. (1981). *Optimalität bei Cluster-Analysen*. Münster: Dissertation, Fachbereich Wirtschafts- und Sozialwissenschaften, Westfälische Wilhelm-Universität Münster.

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*. Berlin: Akademie-Verlag.

Bortz, J., Lienert, G. A., & Boehnke, K. (2000). *Verteilungsfreie Methoden der Biostatistik* (2nd ed.). Berlin: Springer.

British Board of Trade Inquiry Report. (1990). Report on the Loss of the 'Titanic' (S.S.). Gloucester (reprint).

Bowers, J. (1972). A note on comparing r-biserial and r-point biserial. *Educational and Psychological Measurement, 32*, 771–775.

Bühl, A. (2012). *SPSS 20. Einführung in die moderne Datenanalyse unter Windows* (13th ed.). Munich: Pearson Studium.

Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical Education, 42*, 1150–1152.

Cleff, T. (2011). *Deskriptive Statistik und moderne Datenanalyse. Eine computergestützte Einführung mit Excel, PASW (SPSS) und STATA* (2nd ed.). Wiesbaden: Gabler.

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

de Moivre, A. (1738). *Doctrine of chance* (2nd ed.). London: H. Woodfall.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Everitt, B. S., & Rabe-Hesketh, S. (2004). *A handbook of statistical analyses using Stata* (3rd ed.). Chapman & Hall: Boca Raton.

Faulkenberry, G. D., & Mason, R. (1978). Characteristics of nonopinion and no opinion response groups. *Public Opinion Quarterly, 42*, 533–543.

Greene, W. H. (2012). *Econometric analysis* (8th ed.). New Jersey: Pearson Education.

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

Glass, G. V. (1966). Note on rank-biserial correlation. *Educational and Psychological Measurement, 26*, 623–631.

Hair, J. et al. (2006). *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Prentice Hall International.

Harkleroad, D. (1996). Actionable competitive intelligence. In: Society of Competitive Intelligence Professionals (Ed.), *Annual international conference & exhibit conference proceedings* (pp. 43–52). Alexandria.

T. Cleff, *Exploratory Data Analysis in Business and Economics*,
DOI 10.1007/978-3-319-01517-0, © Springer International Publishing Switzerland 2014

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.

Janson, S., & Vegelius, J. (1982). Correlation coefficients for more than one scale type. *Multivariate Behaviorial Research, 17*, 271–284.

Janssens, W., Wijnen, K., de Pelsmacker, P., & van Kenvove, P. (2008). *Marketing research with SPSS*. Essex: Pearson Education.

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement, 34*, 111–117.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data*. New York: Wiley.

Krämer, W. (2005). *So lügt man mit Statistik* (7th ed.). Munich, Zurich: Piper

Krämer, W. (2008). *Statistik verstehen. Eine Gebrauchsanweisung* (8th ed.). Munich, Zurich: Piper.

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

Malhotra, N. K. (2010). *Marketing research. An applied approach* (6th Global Edition). London: Pearson.

Mooi, E., & Sarstedt, M. (2011). *A concise guide to market research. The process, data, and methods using IBM SPSS statistics*. Berlin and Heidelberg: Springer

Pell, G. (2005). Use and misuse of Likert scales. *Medical Education, 39*, 970.

Rinne, H. (2008). *Taschenbuch der Statistik* (4th ed.). Frankfurt/Main: Verlag Harri Deutsch.

Roderick, J. A., Little, R. C., & 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). London/New York: Plenum Press.

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

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

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

Schwarze, J. (2008). *Aufgabensammlung zur Statistik* (6th ed.). Herne/Berlin: NWB.

Swoboda, H. (1971). *Exakte Geheimnisse: Knauers Buch der modernen Statistik*. München, Zurich: Knauer.

Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*, 236–244.

Wooldridge, J. M. (2009). *Introductory econometrics: A modern approach, 4th International Student Edition*. Cincinnati: South-Western Cengage Learning: South-Western Cengage Learning.

# Index

T. Cleff, *Exploratory Data Analysis in Business and Economics*,
DOI 10.1007/978-3-319-01517-0, © Springer International Publishing Switzerland 2014