CHAPTER 17

Exploring, displaying and examining data

Chapter contents

Learning objectives

When you have read this chapter, you should understand:

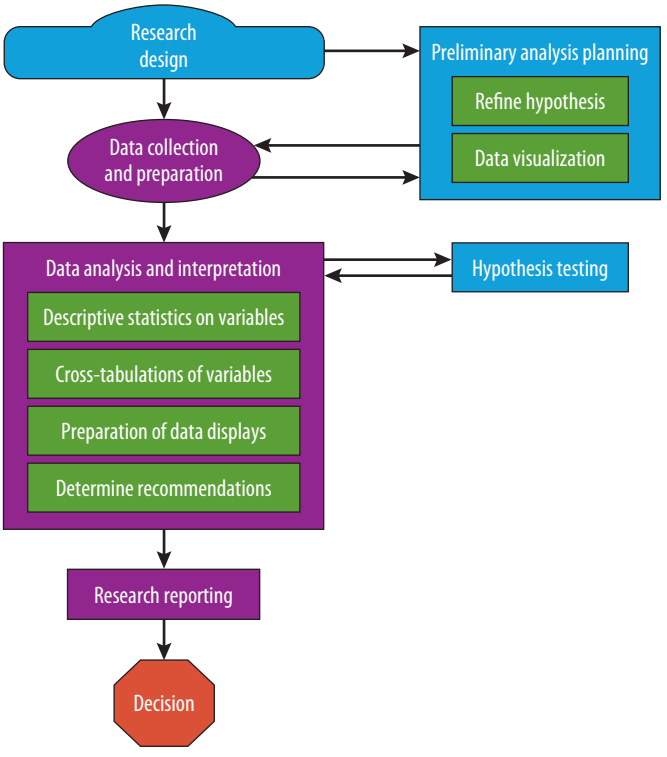
1. that exploratory data analysis techniques are a good diagnostic tool
2. how statistical process control charts can be used to evaluate point values, trends and special causes
3. how cross-tabulation is used to evaluate relationships involving categorical variables.

17.1 Introduction

The convenience of data entry via a spreadsheet, optimal mark recognition (OMR) or with the data editor of a statistical program makes it tempting to move directly to statistical analysis. Why waste time finding out if the data confirms the hypothesis that motivated the study? Why not obtain descriptive statistical summaries (based on our discussion in Chapter 16) and then test hypotheses?

In Chapter 1, we said research conducted scientifically is a puzzle-solving activity. We also noted that an attitude of curiosity, suspicion and imagination was essential to the discovery process. It is natural then that exploration of the data would be an integral part of our perspective. When the study's purpose is not the production of causal inferences, confirmatory data analysis is not required. When it is, we advocate discovering as much as possible about the data before selecting the appropriate means of confirmation. Exhibit 17.1 reminds you of the importance of data visualization as an integral element in the data analysis process and as a necessary step prior to hypothesis testing.

Exhibit 17.1 ***Data analysis in the research process.***



Depending on the type of management question, we can discover a great deal about our data through exploratory data analysis, statistical process control charting and cross-tabulation.

17.2 Exploratory data analysis

Exploratory data analysis (EDA) is both a data analysis perspective and a set of techniques.[1] In exploratory data analysis, the data guide the choice of analysis – or a revision of the planned analysis – rather than the analysis presuming to overlay its structure on the data without the benefit of the analyst's scrutiny. This is comparable to our position that research should be problem-oriented rather than tool-driven. The flexibility to respond to the

patterns revealed by successive iterations in the discovery process is an important attribute of this approach. By comparison, **confirmatory data analysis** occupies a position closer to classical statistical inference in its use of significance and confidence. But confirmatory analysis may also differ from traditional practices by using information from a closely related dataset or by validating findings through the gathering and analysing of new data.[2]

One authority has compared exploratory data analysis to the role of police detectives and other investigators, and confirmatory analysis to that of judges and the judicial system. The former are involved in the search for clues and evidence; the latter are preoccupied with evaluating the strength of what is found. Exploratory data analysis is the first step in the search for evidence, without which confirmatory analysis has nothing to evaluate.[3] Consistent with that analogy, EDA shares a commonality with exploratory designs, not formalized ones. Because it does not follow a rigid structure, it is free to take many paths in unravelling the mysteries in the data – to sift the unpredictable from the predictable.

A major contribution of the exploratory approach lies in the emphasis on visual representations and graphical techniques over summary statistics. Summary statistics, as you will see shortly, may obscure, conceal or even misrepresent the underlying structure of the data. When numerical summaries are used exclusively and accepted without visual inspection, the selection of confirmatory models may be precipitous and based on flawed assumptions. Consequently, it may produce erroneous conclusions.[4] For these reasons, data analysis should begin with visual inspection. After that, it is not only possible but also desirable to cycle between exploratory and confirmatory approaches.

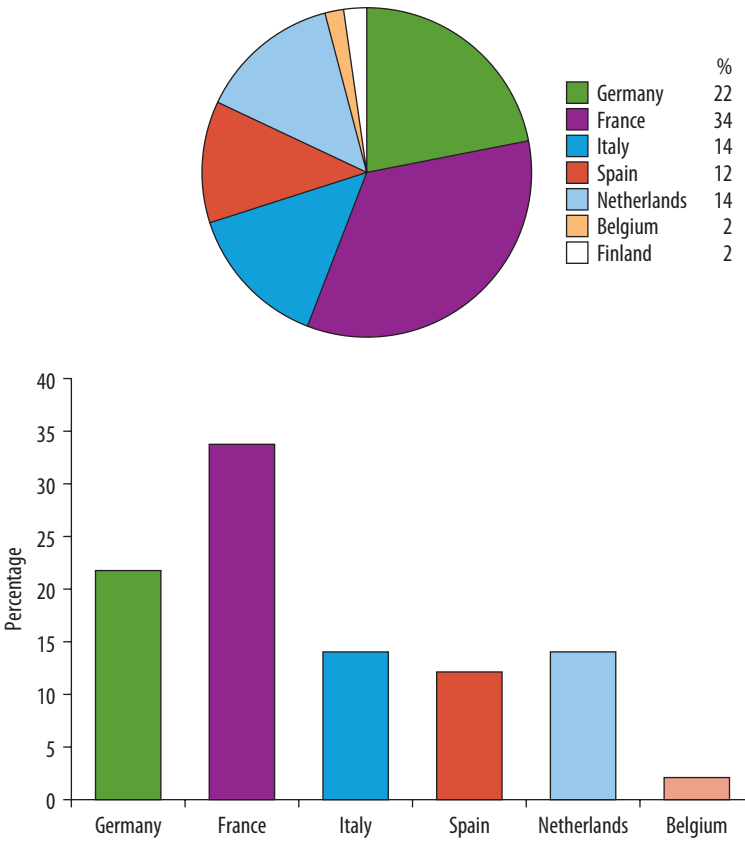
Frequency tables, bar charts and pie charts[5]

Several useful techniques for displaying data are not new to EDA. They are essential to any examination of the data. For example, a **frequency table** is a simple device for arraying data. An example is presented in Exhibit 17.2. It arrays data by assigned numerical value, with columns for per cent, per cent adjusted for missing values and cumulative per cent. Country, the nominal variable that describes the origins of the sampled corporations, provides the observations for this table. Although there are 50 observations, the small number of categories makes the variable easily tabulated. The same data are presented in Exhibit 17.3 using a bar chart and a pie chart. The values and percentages are more readily understood in this graphic format, and visualization of the country categories and their relative sizes is improved. The pie chart offers you a better visualization of the general distribution; you see easily what share the largest countries take relative to the whole sample. However, comparisons between two countries are more difficult in the pie chart: for example, it is hard to see that the slice of Spain is smaller than the slices of Italy and the Netherlands. In the bar chart such small differences are more easily seen, but you do not see immediately that more than 50 per cent of the companies are located in France and Germany.

***Exhibit 17.2** A frequency table of origin of EUROSTOXX50 companies.*

Value label	Value	Frequency	Per cent	Valid per cent	Cumulative per cent
Germany	1	11	22	22	22
France	2	17	34	34	56
Italy	3	7	14	14	70
Spain	4	6	12	12	82
Netherlands	5	7	14	14	96
Belgium	6	1	2	2	98
Finland	7	1	2	2	100
Total		50	100	100	
Valid Cases	50				
Missing Cases	0				

Exhibit 17.3 ***Nominal variable displays (EUROSTOXX50 list).***



When the variable of interest is measured on an interval-ratio scale and is one with many potential values, these techniques are not particularly informative. Exhibit 17.4 is a frequency table of the expected price-earnings ratios of the EUROSTOXX50 firms measured in percentages. Only a few values have a frequency greater than 1. Thus, the primary contribution of this table is an ordered list of values. If the table were converted to a bar chart, most bars would have an equal length and a few bars would be a little bit longer. In addition, bar charts do not reserve spaces for values where no observations occur within the range. Constructing a pie chart for this variable would also be pointless.

Histograms

The **histogram** is a conventional solution for the display of interval-ratio data. Histograms are used when it is possible to group the variable's values into intervals. Histograms are constructed with bars (or asterisks that represent data values) where each value occupies an equal amount of area within the enclosed area. Data analysts find histograms useful for (i) displaying all intervals in a distribution, even those without observed values, and (ii) examining the shape of the distribution for skewness, kurtosis and the modal pattern. When looking at a histogram, one might ask: Is there a single hump (a mode)? Are sub-groups identifiable when multiple modes are present? Are straggling data values detached from the central concentration?[6]

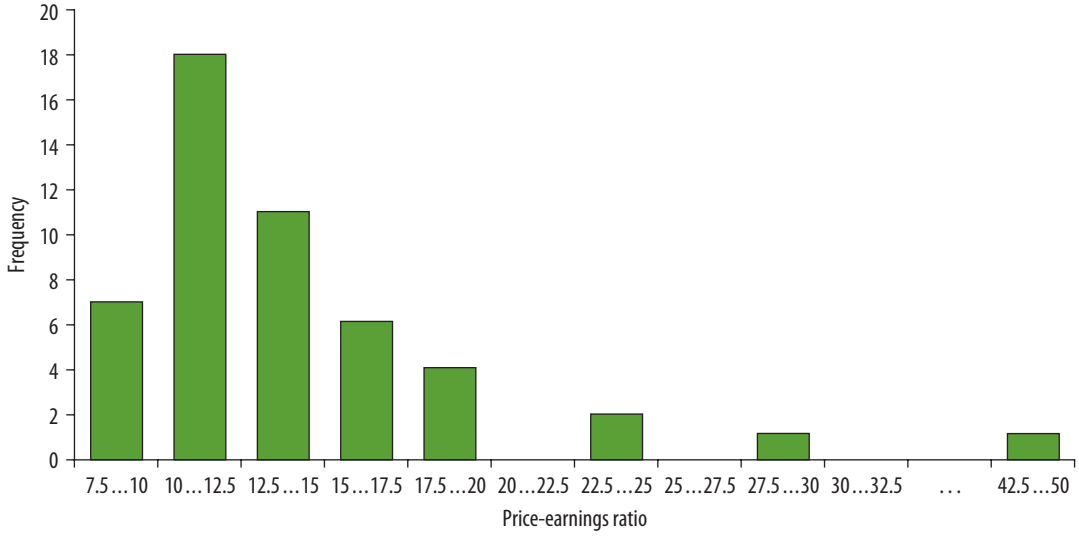
The values for the expected price-earnings variable presented in Exhibit 17.4 were measured on a ratio scale and are easily grouped. Other variables possessing an underlying order are similarly appropriate for histograms.

A histogram of the expected price-earning variable taken from the EUROSTOXX50 list ranked by performance listing is shown in Exhibit 17.5. The midpoint for each interval for the variable of interest, return to investors, is shown on the horizontal axis; the frequency or number of observations in each interval on the vertical axis. We erect a vertical bar above the midpoint of each interval on the horizontal scale. The height of the bar corresponds

Exhibit 17.4 *Frequency table price-earnings ratio of EUROSTOXX50 firms.*

	Value	Frequency	Percentage	Cumulative percentage		Value	Frequency	Percentage	Cumulative percentage
1	7.70	1	2	2	22	12.60	1	2	54
2	7.90	1	2	4	23	12.70	1	2	56
3	8.00	1	2	6	24	13.50	1	2	58
4	8.40	1	2	8	25	13.60	1	2	60
5	8.50	1	2	10	26	13.70	1	2	62
6	8.60	1	2	12	27	14.00	1	2	64
7	9.60	1	2	14	28	14.10	2	4	68
8	10.00	2	4	18	29	14.20	1	2	70
9	10.10	1	2	20	30	14.30	1	2	72
10	10.20	3	6	26	31	15.30	1	2	74
11	10.30	1	2	28	32	15.80	1	2	76
12	10.70	1	2	30	33	15.90	1	2	78
13	10.80	2	4	34	34	16.40	1	2	80
14	11.10	1	2	36	35	16.70	1	2	82
15	11.30	2	4	40	36	16.80	1	2	84
16	11.60	1	2	42	37	17.50	1	2	86
17	11.90	1	2	44	38	18.50	2	4	90
18	12.00	1	2	46	39	19.60	1	2	92
19	12.30	1	2	48	40	22.50	2	4	96
20	12.40	1	2	50	41	27.90	1	2	98
21	12.50	1	2	52	42	48.10	1	2	100

Exhibit 17.5 *Histogram of expected price-earnings ratios of EUROSTOXX50 companies.*



with the frequency of observations in the interval above which it is erected. This histogram was constructed with intervals 2.5 increments wide, and the final interval contains only one observation, 48.1. The intervals 20 . . . 22.5; 25 . . . 27.5 and all intervals between 30 and 47.5 contain no information. The large gap at the end is emphasized by an empty interval labelled '. . .'. These values are found in the expected price-earnings ratio frequency table (Exhibit 17.4). Intervals with 0 counts show gaps in the data and alert the analyst to look for problems with spread. When the upper tail of the distribution is compared with the frequency table, we find four extreme values (two times 22.5, 27.9 and 48.1). Along with the peaked midpoint and reduced number of observations in the upper tail, this histogram warns us of irregularities in the data.

Exhibit 17.6 A stem-and-leaf display of price-earnings ratios of EUROSTOXX50 firms.

7	79
8	0456
9	1
10	0012223788
11	13369
12	034567
13	567
14	01123
15	389
16	478
17	5
18	55
19	6
20	
21	
22	55
23	
24	
25	
26	
27	9
28	
29	
30	
31	
32	
33	
34	
35	
36	
37	
38	
39	
40	
41	
42	
43	
44	
45	
46	
47	
48	1

Stem-and-leaf displays[7]

The **stem-and-leaf display** is an EDA technique that is closely related to the histogram. It shares some of its features but offers several unique advantages. It is easy to construct by hand for small samples or may be produced by computer programs.

In contrast to histograms, which lose information by grouping data values into intervals, the stem-and-leaf presents actual data values that can be inspected directly without the use of enclosed bars or asterisks as the representation medium. This feature reveals the distribution of values within the interval and preserves their rank order for finding the median, quartiles and other summary statistics. It also eases the linking of a specific observation back to the data file and to the subject that produced it.

Visualization is the second advantage of stem-and-leaf displays. The range of values is apparent at a glance, and both shape and spread impressions are immediate. Patterns in the data – such as gaps where no values exist, areas where values are clustered or outlying values that differ from the main body of the data – are easily observed.

In order to develop a stem-and-leaf display for the data in Exhibit 17.4, the first two digits of each data item are arranged to the left of a vertical line. Next, we pass through the price-earnings ratios in the order that they were recorded and place the last digit for each item (the unit position, 0.1) to the right of the vertical line. The last digit for each item is placed on the horizontal row corresponding to its first digit(s). Now it is a simple matter to rank-order the digits in each row, creating the stem-and-leaf display shown in Exhibit 17.6.

Each line or row in this display is referred to as a stem and each piece of information on the stem is called a leaf. The first line or row is:

[7 | 7 9]

The meaning attached to this line or row is that there are two items in the dataset whose first digit is seven: 7.7; 7.9. The second line is:

[8 | 0 4 5 6]

and shows that there are four price-earnings ratios whose first digit is eight: 8.0; 8.4; 8.5 and 8.6. The stem is the digit(s) to the left of the vertical line (8 for this example) and the leaf is the digit(s) to the right of the vertical line (0, 4, 5, 6).

When the stem-and-leaf display shown in Exhibit 17.6 is turned upright (rotated 90° to the left), the shape is the same as that of the histogram shown in Exhibit 17.5.

Boxplots[8]

The **boxplot**, or box-and-whisker plot, is another technique used frequently in exploratory data analysis.[9] A boxplot provides a visual image of the distribution's location, spread, shape, tail length and outliers. Boxplots are extensions of the **five-number summary** of a distribution. This summary consists of the median, upper and lower quartiles, and the largest and smallest observations. The median and quartiles are used because they are particularly **resistant statistics**. Resistance is a characteristic that 'provides insensitivity to localized misbehaviour in data'.[10] Resistant statistics are unaffected by outliers and change only slightly in response to the replacement of small portions of the dataset.

Recall the discussion of the mean and standard deviation in Chapter 16. Now assume we take the dataset [5,6,6,7,7,7,8,8,9]. The mean of the set is 7, the standard deviation 1.23. If the 9 is replaced with 90, the mean becomes 16 and the standard deviation increases to 27.78. The mean is now two times larger than most of the numbers in the distribution and the standard deviation is more than 22 times its original size. Changing only one of nine values has disturbed the location and spread summaries to the point where they no longer represent the other eight values. Both the mean and the standard deviation are considered **non-resistant statistics**; they are susceptible to the effects of extreme values in the tails of the distribution and do not represent typical values well under conditions of asymmetry. The standard deviation is particularly problematic because it is computed from the squared deviations from the mean.[11] In contrast, the median and quartiles are highly resistant to change. When we changed the 9 to 90, the median remained at 7 and the lower and upper quartiles stayed at 6 and 8, respectively.

Because of the nature of quartiles, up to 25 per cent of the data can be made extreme without perturbing the median, the rectangular composition of the plot or the quartiles themselves. These characteristics of resistance are incorporated into the construction of boxplots.

Boxplots may be constructed easily by hand or by computer programs. The basic ingredients of the plot are (i) the rectangular plot that encompasses 50 per cent of the data values, (ii) a centre line (or other notation) marking the median and going through the width of the box, (iii) the edges of the box, called hinges, and (iv) the whiskers that extend from the right and left hinges to the largest and smallest values.[12] These values may be found within 1.5 times the interquartile range (IQR) from either edge of the box. These components and their relationships are shown in Exhibit 17.7.

We can create a boxplot of the market value of the EUROSTOXX50 firms.[13] Exhibit 17.8 shows the essential summary numbers for a boxplot of the market value.

Exhibit 17.7 Boxplot components.

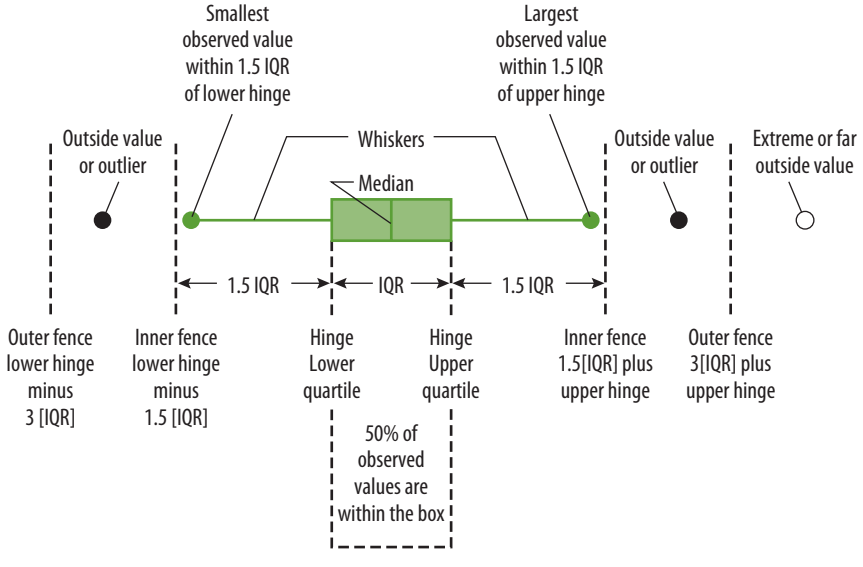
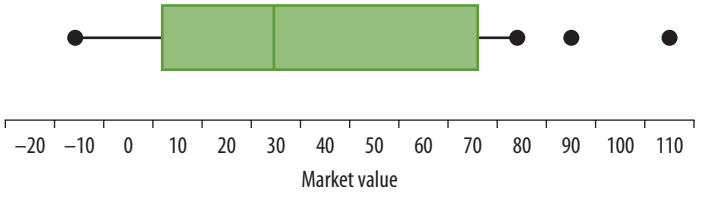


Exhibit 17.8 ***Summary of market value.***

Minimum	Lower hinge (p25)	Median	Upper hinge (p75)	Maximum
0.89	16.50	26.73	38.97	108.9
		Fence		
IQR	Distance to	(–)	(+)	
22.47	(±1.5) = 33.70	16.50 – 33.70 = –17.20	38.97 + 33.70 = 72.67	

Exhibit 17.9 ***Boxplot of market value of EUROSTOXX50 firms.***

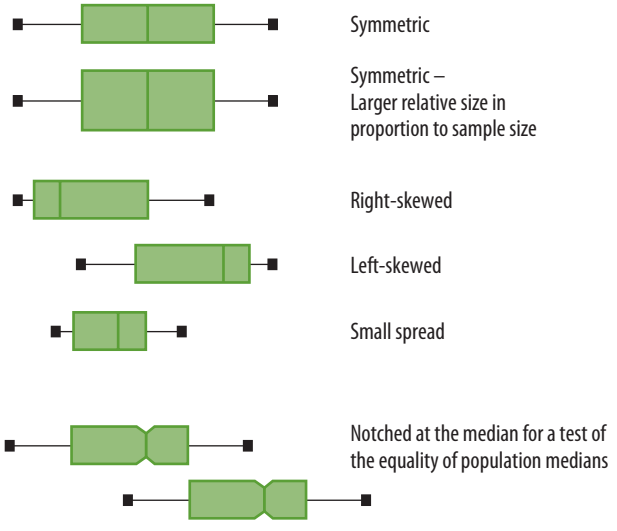


The plot shown in Exhibit 17.9 started with these data and the following calculations. You may construct your own boxplot with the data provided in the data file 'EUROSTOXX50' on the accompanying website and the SPSS Explore procedure. Beginning with the box, the ends are drawn using the lower and upper quartile (hinge) data. The median is drawn in at 26.73. Then the IQR is calculated (38.97 – 16.50 = 22.47). From this we can locate the lower and upper fences. The fences are –17.20 and 72.67. Next, the smallest and largest data values from the distribution within the fences are used to determine the whisker length. These values are 0.89 and 66.70. We are now able to see the outliers in relation to the 'main body' of the data. **Outliers** are data points that exceed +1.5 IQRs of a boxplot's hinges. Data values for the outliers are added, and identifiers may be provided for interesting values. The completed boxplot is shown in Exhibit 17.9.

When examining data, it is important to separate legitimate outliers from errors in measurement, editing, coding and data entry. Outliers that reflect unusual cases are an important source of information for the study. They are displayed or given special statistical treatment, or other portions of the dataset are sometimes shielded from their effects. Outliers that are mistakes should be corrected or removed.

Exhibit 17.10 summarizes several comparisons that are of help to the analyst. Boxplots are an excellent diagnostic tool, especially when graphed on the same scale. The upper two plots in the exhibit are both symmetric, but one is larger than the other. Larger box widths are sometimes used when the second variable, from the same measurement scale, comes from a larger sample size. The box widths should be proportional to the square root of the sample size, but not all plotting programs account for this.[14] Right- and left-skewed distributions, and those with reduced spread are also presented clearly in the plot comparison. Finally, groups may be compared by means of multiple plots. One variation, in which a notch at the median marks off a confidence interval to test the equality of group medians, takes us a step closer to hypothesis testing.[15] Here the sides of the box return to full width at the upper and lower confidence intervals. When the intervals do not overlap, we can be confident, at a specified confidence level, that the medians of the two populations are different.

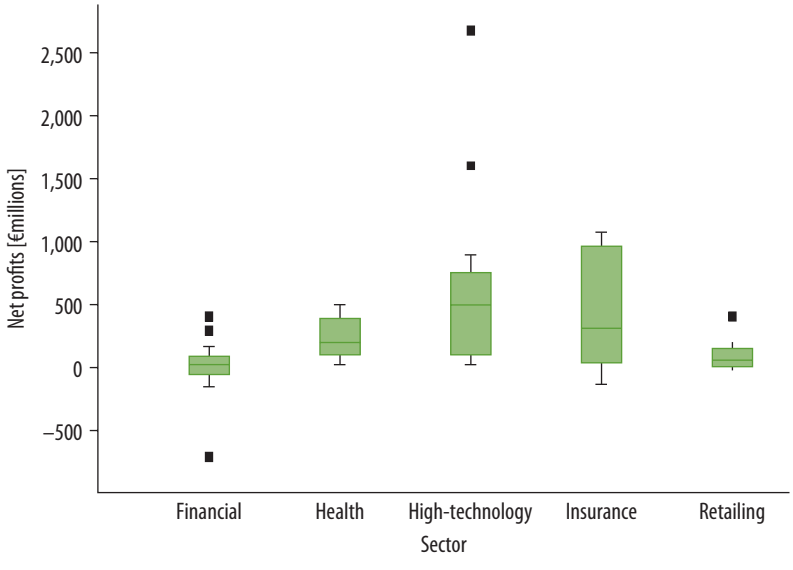
Exhibit 17.10 ***Diagnostics with boxplots.***



In Exhibit 17.11, multiple boxplots compare five sectors on the 'profit' variable. The overall impression is one of potential problems for the

analyst: unequal variances, skewness and extreme outliers. Note the similarities of the profiles of finance and retailing in contrast to the high-technology and insurance sectors. If hypothesis tests are planned, further examination of this plot for each sector would require a stem-and-leaf display and a five-number summary. From this, we could make decisions on test selection and whether the data should be transformed or re-expressed before further analysis.

Exhibit 17.11 ***Boxplot comparison of p/e ratios in different euro (€) countries.***



SPSS reference

There are many software packages that allow you to produce the graphics presented above. SPSS allows has a graphic modules and how to create such graphics in SPSS is shown in Pallant (2013) Chapter 7.

Transformation[16]

Some of the examples in this section have departed from normality. While this makes for good illustrations, such data pose special problems in data analysis. Transformation is one solution to this problem. **Transformation** is the re-expression of data on a new scale using a single mathematical function for each data point. Although nominal and ordinal data may be transformed, the procedures are beyond the scope of this book. We consider only interval-ratio scale transformations here.

The fact that data collected on one scale are found to depart from the assumptions of normality and constant variance does not preclude re-expressing them on another scale. What is discovered, of course, must be linked to the original data.

We transform data for several reasons: (i) to improve interpretation and compatibility with other datasets, (ii) to enhance symmetry and stabilize spread, and (iii) to improve linear relationships between and among variables. We improve interpretation when we find alternate ways to understand the data and discover patterns or relationships that may not have been revealed on the original scales. A **standard score (*Z* score)** may be calculated to improve compatibility among variables that come from different scales and require comparison. *Z* scores convey distance in standard deviation units with a mean of 0 and a standard deviation of 1. This is accomplished by converting the raw score, X_i, to

$$Z = \frac{X_i - \bar{X}}{s}$$

Z scores improve interpretation through their reference to the normal curve and our understanding of the areas under it.

Conversion of US dollars to euros, centimetres to inches, stones to pounds, litres to gallons, or Celsius to Fahrenheit are examples of linear conversions that change scale but do not change symmetry or spread. Many statisticians consider these data as manipulations rather than transformations.

Non-linear transformations are often needed to satisfy the other two reasons for re-expressing data. Normality and constancy of variance are important assumptions for many parametric statistical techniques. A transformation to reduce skewness and stabilize variance makes it possible to use various confirmatory techniques without violating their assumptions. Analysis of the relationship between variables also benefits from transformation. Improved predictions and better diagnostics of fit and residuals (as in regression analysis) are frequent payoffs.

Transformations are defined with power, p, as the re-expression of x with x^p.[17] Exhibit 17.12 shows the most frequently used power transformations.

Exhibit 17.12 Frequently used power transformations.

Power	Transformation
3	Cube
2	Square
1	No change, existing data
1/2	Square root
−1/2	Reciprocal square root
−1	Reciprocal
−2	Reciprocal square
−3	Reciprocal cube

We use spread-and-level plots to guide our choice of a power transformation. By plotting the log of the median against the log of the interquartile range, we can find the slope of the plot: where p, the power we are seeking, is equal to 1 – slope. Although 1/4 and 1/3 powers often result – and are sometimes preferred – many computer programs require rounding the transformation to the nearest half power.

The price-earnings variable is used to illustrate this concept. The data distribution shows a right skew. The five-number summary reveals an extreme score as the maximum data point (see Exhibit 17.13).

Exhibit 17.13 Summary price-earnings ratios.

Minimum	Lower hinge (p25)	Median	Upper hinge (p75)	Maximum
7.70	10.20	12.45	15.80	48.10

The largest observation, 48.10, is only approximately 3 IQRs beyond the main body of data, and there are only three other values beyond the fence.

A quick calculation of the ratio of the largest observation to the smallest (48.10/7.70 = 6.25) serves as the final confirmation that transformation might not be worthwhile. It is desirable for this informal index to be greater than 20; with ratios less than 2, transformation is not practical.[18] From this information, we might conclude that the price-earnings variable is not a good candidate for transformation.

When researchers communicate their findings to people not attached to scientific research (e.g. those involved in management or politics), the advantages of re-expression must be balanced against pragmatism: some transformed scales have no familiar analogies. Logarithmic euro can be explained, but how about reciprocal root euro? Attitude and preference scales might be better understood transformed, but the question of interpretation remains.

Throughout this section we have exploited the visual techniques of exploratory data analysis to look beyond numerical summaries and gain insight into the behaviour of the data. Few of the approaches have stressed the need for advanced mathematics, and all have an intuitive appeal for the analyst. When the more common ways of summarizing location, spread and shape have conveyed an inadequate picture of the data, we have used more resistant statistics to protect us from the effects of extreme scores and occasional errors. We have also emphasized the value of transforming the original scale of the data during preliminary analysis rather than at the point of hypothesis testing.

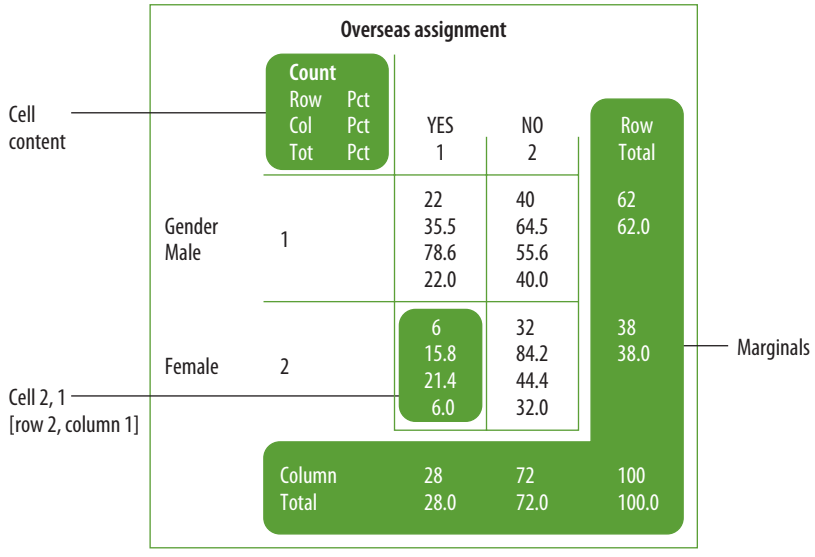
SPSS reference

SPSS has incorporated many mathematical functions that can be used when you make (COMPUTE in SPSS jargon) variables. You will find the details on how to do this in Pallant (2013) Chapter 8.

Cross-tabulation

Cross-tabulation is a technique for comparing two classification variables, such as gender and selection by one's company for an overseas assignment. The technique uses tables with rows and columns that correspond to the levels or values of each variable's categories. Exhibit 17.14 is an example of a computer-generated cross-tabulation. This table has two rows for gender and two columns for assignment selection. The combination of the variables with their values produces four cells. Each cell contains a count of the cases of the joint classification and also the row, column and total percentages. The number of row cells and column cells is often used to designate the size of the table, as in this 2×2 table. The cells are individually identified by their row and column numbers, as illustrated. Row and column totals, called **marginals**, appear at the bottom and right 'margins' of the table. They show the counts and percentages of the separate rows and columns.

Exhibit 17.14 ***SPSS cross-tabulation of gender by overseas assignment.***



When tables are constructed for statistical testing, we call them contingency tables, and the test determines if the classification variables are independent. Of course, tables may be larger than 2×2.

The use of percentages

Percentages serve two purposes in data presentation. First, they simplify the data by reducing all numbers to a range from 0 to 100. Second, they translate the data into standard form, with a base of 100, for relative comparisons. In a sampling situation, the number of cases that fall into a category is meaningless unless it is related to some base. A count of 28 overseas assignees has little meaning unless we know it is from a sample of 100. Using the latter as a base, we conclude that 28 per cent of this study's sample has an overseas assignment.

While the above is useful, it is even more useful when the research problem calls for a comparison of several distributions of data. Assume the previously reported data were collected five years ago and the present study had a

Exhibit 17.15 ***Comparison of percentages in cross-tabulation studies of gender by overseas assignment.***

Study 1

Overseas assignment

	Count / Row Pct / Col Pct / Tot Pct	YES 1	NO 2	Row Total
Gender Male	1	22 35.5 72.6 22.0	40 64.5 55.6 40.0	62 62.0
Female	2	6 15.2 21.4 6.0	32 64.2 44.4 32.0	38 38.0
	Column Total	28 28.0	72 72.0	100 100.0

Study 2

Overseas assignment

	Count / Row Pct / Col Pct / Tot Pct	YES 1	NO 2	Row Total
Gender Male	1	225 25.5 62.5 15.0	675 75.0 59.2 45.0	900 60.0
Female	2	135 22.5 37.5 9.0	465 77.2 40.8 31.0	600 40.0
	Column Total	360 24.0	1140 76.0	1500 100.0

sample of 1,500, of which 360 were selected for overseas assignments. By using percentages, we can see the relative relationships and shifts in the data (see Exhibit 17.15).

With two-dimension tables, the selection of a row or column will accentuate a particular distribution or comparison. This raises the question about which direction the percentages should be calculated. Most computer programs offer options for presenting percentages in both directions and interchanging the rows and columns of the table. However, in situations where one variable is hypothesized to be the presumed cause, is thought to affect or predict a response, or is simply antecedent to the other variable, we label it the independent variable. Percentages should then be computed in the direction of this variable. Thus, if the independent variable is placed on the row, select row percentages; if it is on the column, select column percentages. In which direction should the percentages run in the previous example? If the column percentages alone are reported, we imply that assignment status has some effect on gender. This is implausible. When percentages are reported by row, the implication is that gender influences selection for overseas assignments.

Care should be taken in interpreting percentages from tables. Consider again the data in Exhibit 17.15. From the first to the second study, it is apparent that the percentage of females selected for overseas assignment rose from 6 to 9 per cent of their respective samples. This is not to be confused with the percentage of women in the samples who happen to be assignees. Among all women eligible for selection in the first study, 15.8 per cent were assigned and 84.2 per cent were not. Among all overseas selectees in the first study, 21.4 per cent were women. Similar comparisons can be made for the other categories.

Percentages are used by virtually everyone dealing with numbers – and often incorrectly. The following guidelines, if used during analysis, will help to prevent errors in reporting:[19]

- *Averaging percentages*. Percentages cannot be averaged unless each is weighted by the size of the group from which it is derived. Thus, a simple average will not suffice; it is necessary to use a weighted average.
- *Use of too large percentages*. This often defeats the purpose of percentages – which is to simplify. A large percentage is difficult to understand and is confusing. If a 1,000 per cent increase is experienced, it is better to describe this as a tenfold increase.
- *Using too small a base*. Percentages hide the base from which they have been computed. A figure of 100 per cent when contrasted with 20 per cent would appear to suggest a sizeable difference. Yet if the base of the former is three, while the base of the latter is 250, the absolute differences would tell a different story.
- *Percentage decreases can never exceed 100 per cent*. This is obvious, but this type of mistake occurs frequently. The higher figure should always be used as the base. For example, if a price was reduced from €2 to 50 cents, the decrease would be 75 per cent (50/200).

Other table-based analysis

The recognition of a meaningful relationship between variables generally signals a need for further investigation. Even if one finds a statistically significant relationship, the questions of why and under what conditions remain. The introduction of a **control variable** to interpret the relationship is often necessary. Cross-tabulation tables serve as the framework.

Statistical packages like STATA, SAS and SPSS have among their modules many options for the construction of n-way tables with provision for multiple control variables. Suppose you are interested in creating a cross-tabulation of two variables with one control. Whatever the number of values in the primary variables, the control variable with five values determines the number of tables. For some applications, it is appropriate to have five separate tables; for others, it might be preferable to have adjoining tables or have the values of all the variables in one. Management reports are of the latter variety. Exhibit 17.16 presents an example in which all three variables are handled under the same banner. Programs such as this one can handle far more complex tables and statistical information.[20]

Exhibit 17.16 ***SPSS cross-tabulation with control and nested variables.***

	Control variable					
	Category 1			Category 2		
	Nested variable			Nested variable		
	cat 1	cat 2	cat 3	cat 1	cat 2	cat 3
Stub . . .	cells . . .					

	SEX OF EMPLOYEE			
	MALES		FEMALES	
	MINORITY CLASSIFICATION		MINORITY CLASSIFICATION	
	WHITE	NON-WHITE	WHITE	NON-WHITE
EMPLOYMENT CATEGORY				
CLERICAL	16%	7%	18%	7%
OFFICE TRAINEE	7%	3%	17%	2%
SECURITY OFFICER	3%	3%		
COLLEGE TRAINEE	7%	0%	1%	
EXEMPT EMPLOYEE	6%	0%	0%	
MBA TRAINEE	1%	0%	0%	
TECHNICAL	1%			

Summary

1 The objective of exploratory data analysis (EDA) is to learn as much as possible about the data. EDA simplifies this goal by providing a perspective and set of tools to search for clues and patterns. EDA augments rather than supplants traditional statistics. In addition to numerical summaries of location, spread and shape, EDA uses visual displays to provide a complete and accurate impression of distributions and variable relationships.

Frequency tables array data from highest to lowest values with counts and percentages. They are most useful for inspecting the range of responses and their repeated occurrence. Bar charts and pie charts are appropriate for relative comparisons of nominal data. Histograms are optimally used with continuous variables where intervals group the responses.

▶

Stem-and-leaf displays and boxplots are EDA techniques that provide visual representations of distributions. The former present actual data values using a histogram-type device that allows inspection of spread and shape. Boxplots use the five-number summary to convey a detailed picture of a distribution's main body, tails and outliers. Both stem-and leaf displays and boxplots rely on resistant statistics to overcome the limitations of descriptive measures that are subject to extreme scores. Transformation may be necessary to re-express metric data so as to reduce or remove problems of asymmetry, inequality of variance or other abnormalities.

2 The feature of **statistical process control (SPC)** that pertains to displaying data is its charting system. SPC charts provide reliable visuals for evaluating point values and trends, and depicting variation graphically. Control charts help managers focus on special causes of variation by revealing whether a system is under control (as an early warning of change) and substantiating results from improvements (confirming results).

A control chart displays sequential measurements of a process together with a centre line and control limits. The selection of a control chart depends on the level of data (data type) you are measuring. Managers should look for the following visual characteristics in reports containing control charts: (i) outliers – observations falling outside the control lines; (ii) runs – data points in a series over or below the central line; (iii) trends – the continual rise or fall of data points; and (iv) periodicity – data that show the same pattern of change over time, creating a cycle.

3 The evaluation of relationships involving categorical variables employs cross-tabulation. The tables used for this purpose consist of cells and marginals. The cells contain combinations of count, row, column and total percentages. The tabular structure is the framework for later statistical testing.

Computer software for cross-classification analysis makes table-based analysis with one or more control variables an efficient tool for decision-making.

Discussion questions

Terms in review

1 Define or explain:
- **a** marginals
- **b** standard scores (Z scores)
- **c** control chart
- **d** non-resistant statistics
- **e** lower control limit
- **f** the five-number summary
- **g** spread-and-level plots.

Making research decisions

2 How should the researcher handle missing values?

3 How do the following detect errors in the data?
- **a** Histogram
- **b** Stem-and-leaf display
- **c** Boxplot
- **d** Cross-tabulation

From concept to practice

4 Use the data in Exhibit 17.4 to construct a stem-and-leaf display.
- **a** Where do you find the main body of the distribution?
- **b** How many values reside outside the inner fence(s)?

5 Select the sales variable from Exhibit 17.17.
- **a** Create a five-number summary.
- **b** Construct a boxplot.
- **c** Interpret the distribution and results with summary measures and descriptive statistics.
- **d** Transform the variable into Z scores.
- **e** Identify and comment on outliers, if any.

6 Select the market value variable from Exhibit 17.17 and construct a histogram with available software.
- **a** What is the gain in information with 5,000-, 2,000- or 1,000-unit intervals?
- **b** Which would be the best interval to convey results to management?
- **c** Why would these data need re-expression?
- **d** What is the optimal power transformation for these data?

Exhibit 17.17 ***Data table for discussion questions 5 and 6.***

	Market value	Sales	Sector		Market value	Sales	Sector
1	24 983.00	8 966.00	2	26	9 009.00	17 533.00	4
2	31 307.00	126 932.00	3	27	7 842.00	11 113.00	2
3	57 193.00	54 574.00	7	28	5 431.00	19 671.00	8
4	57 676.00	86 656.00	4	29	5 811.00	11 389.00	5
5	60 345.00	62 710.00	7	30	16 257.00	15 242.00	2
6	22 190.00	96 146.00	3	31	16 247.00	10 211.00	7
7	36 566.00	39 011.00	2	32	18 548.00	9 593.00	7
8	44 646.00	36 112.00	7	33	13 620.00	9 691.00	7
9	25 022.00	50 220.00	4	34	10 750.00	12 844.00	3
10	26 043.00	25 099.00	1	35	12 450.00	18 398.00	2
11	13 152.00	53 794.00	2	36	16 729.00	20 276.00	7
12	11 234.00	25 047.00	5	37	16 532.00	8 730.00	7
13	26 666.00	23 966.00	4	38	5 111.00	17 635.00	10
14	20 747.00	17 424.00	7	39	9 116.00	8 588.00	4
15	25 826.00	13 996.00	7	40	26 325.00	25 922.00	2
16	15 423.00	32 416.00	4	41	8 249.00	16 103.00	2
17	15 263.00	14 150.00	8	42	8 407.00	14 083.00	3
18	18 146.00	17 600.00	1	43	18 537.00	11 990.00	10
19	18 739.00	15 351.00	4	44	23 866.00	29 443.00	4
20	7 875.00	22 605.00	2	45	6 872.00	19 532.00	7
21	8 122.00	37 970.00	5	46	4 319.00	10 018.00	5
22	18 072.00	11 557.00	5	47	9 505.00	12 937.00	7
23	6 404.00	11 449.00	7	48	3 891.00	15 654.00	8
24	16 056.00	20 054.00	8	49	8 090.00	7 492.00	4
25	16 056.00	13 211.00	7	50	1 119.00	12 345.00	7

7 Suppose you were preparing two-way tables of percentages for the following pairs of variables. How would you run the percentages?
- **a** Age and consumption of breakfast cereal.
- **b** Family income and confidence about the family's future.
- **c** Marital status and sports participation.
- **d** Crime rate and unemployment rate.

8 You study the attrition between students who enter college as first years (or 'freshers') and those who stay to graduate. You find the following relationships between attrition, aid and distance of home from school. What is your interpretation? Consider all variables and relationships.

	Aid		Home Near School		Home Far School	
	Yes	No	Yes	No	Yes	No
Drop out	25%	20%	5%	15%	30%	40%
Stay	75%	80%	95%	85%	70%	60%

9 A local health agency is experimenting with two appeal letters, A and B, with which to raise funds. It sends out 400 of the A appeal and 400 of the B appeal (divided equally among working-class and middle-class neighbourhoods). The agency secures the results shown in the table below.
- **a** Which appeal is the best?
- **b** Which class responded better?
- **c** Is appeal or social class a more powerful independent variable?

	Appeal A		Appeal B	
	Middle Class	Working Class	Middle Class	Working Class
Contribution	20%	40%	15%	30%
No contribution	80%	60%	85%	70%
	100%	100%	100%	100%

10 Assume you have collected data on employees of a large organization in a major metropolitan area. You analyse the data by type of work classification, education level, and whether the workers were raised in a rural or urban setting. The results are shown below. How would you interpret them?

Annual employee turnover per 100 employees

	Part A		Part B			
			High Education		Low Education	
	Salaried	Wage	Salaried	Wage	Salaried	Wage
Rural	8	16	6	14	18	18
Urban	12	16	10	12	19	20

Recommended further reading

Evans, James R. and William M. Lindsay, *Management and the Control of Quality*. Mason, OH: South-Western Publishing, 2002. Technically detailed coverage of quality improvement techniques.

Hoaglin, David C., Frederick Mosteller and John W. Tuckey (eds), *Exploring Data Tables, Trends and Shapes* (revised edition). New York: Wiley-Blackwell, 2006. An edited volume from pioneers in exploratory data analysis offering a good and complete overview on the topic.

Myatt, Glenn J., *Making Sense of Data. A Practical Guide to Exploratory Data Analysis and Data Mining*. New York: Wiley-Interscience, 2006. An introduction to first steps in understanding data.

Tufte, Edward, *The Visual Display of Quantitative Information* (2nd edn), Cheshire, CT: Graphics Press, 2001. A superb book on how graphics add to understanding data.

Get started with understanding statistical techniques!

When you have read this chapter, log on to the Online Learning Centre website at *www.mcgraw-hill.co.uk/textbooks/blumberg* to explore chapter-by-chapter test questions, additional case studies, a glossary and more online study tools for *Business Research Methods*.

Notes

1 John W. Tukey, *Exploratory Data Analysis*. Reading, MA: Addison-Wesley, 1977.

2 David C. Hoaglin, Frederick Mosteller and John W. Tukey (eds), *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley & Sons, 1983, p. 2.

3 Tukey, *Exploratory Data Analysis*, pp. 2–3.

4 Frederick Hartwig with Brian E. Dearing, *Exploratory Data Analysis*. Beverly Hills, CA: Sage Publications, 1979, pp. 9–12.

5 The exhibits in this section were created with statistical and graphic programs particularly suited to exploratory data analysis. The authors acknowledge the following vendors for evaluation and use of their products: SPSS, Inc., 233 Dr. S. Wacker, Chicago, IL, 60606; and Data Description, PO Box 4555, Ithaca, NY, 14852.

6 Paul F. Velleman and David C. Hoaglin, *Applications, Basics, and Computing of Exploratory Data Analysis*. Boston: Duxbury Press, 1981, p. 13.

7 John Hanke, Eastern Washington University, contributed to this section. For further references to stem-and-leaf displays, see John D. Emerson and David C. Hoaglin, 'Stem-and-leaf displays', in *Understanding Robust and Exploratory Data Analysis*, pp. 7–31; and Velleman and Hoaglin, *Applications*, pp. 1–13.

8 This section is adapted from the following excellent discussions of boxplots: Velleman and Hoaglin, *Applications*, pp. 65–76; Hartwig, *Exploratory Data Analysis*, pp. 19–25; John D. Emerson and Judith Strenio, 'Boxplots and batch comparison', in *Understanding Robust and Exploratory Data Analysis*, pp. 59–93; and Amir D. Aczel, *Complete Business Statistics*. Homewood, IL: Irwin, 1989, pp. 723–8.

9 Tukey, *Exploratory Data Analysis*, pp. 27–55.

10 Hoaglin et al., *Understanding Robust and Exploratory Data Analysis*, p. 2.

11 Several robust estimators that are suitable replacements for the mean and standard deviation are not discussed here – for example, the trimmed mean, trimean, the M-estimators (such as Huber's, Tukey's, Hampel's and Andrew's estimators), and the median absolute deviation (MAD). See Hoaglin et al., *Understanding Robust and Exploratory Data Analysis*, Chapter 10; and SPSS, Inc., SPSS Base 9.0 User's Guide. Chicago: SPSS, Inc., 1999, Chapter 13.

12 The difference between the definition of a hinge and a quartile is based on variations in their calculation. We use Q1, 25th percentile and lower hinge synonymously; and Q3, 75th percentile, and upper hinge, similarly. There are technical differences, although they are not significant in this context.

13 Here, we use the market value of the EUROSTOXX50 firms on 16 May 2004.

14 R. McGill, J.W. Tukey and W.A. Larsen, 'Variations of box plots', *The American Statistician* 32 (1978), pp. 12–16.

15 See J. Chambers, W. Cleveland, B. Kleiner and John W. Tukey, *Graphical Methods for Data Analysis*. Boston, MA: Duxbury Press, 1983.

16 This section is based on the discussion of transformation, in John D. Emerson and Michael A. Stoto, 'Transforming data', *Understanding Robust and Exploratory Data Analysis*, pp. 97–127; and Velleman and Hoaglin, *Applications*, pp. 48–53.

17 Hoaglin et al., *Understanding Robust and Exploratory Data Analysis*, p. 77.

18 Ibid., p. 125.

19 Harper W. Boyd Jr. and Ralph Westfall, *Marketing Research* (3rd edn). Homewood, IL: Irwin, 1972, p. 540.

20 SPSS, Inc., *SPSS Tables 8.0*. Chicago: SPSS, Inc., 1998 (with its system file: Bank Data).