Chapter 5
Descriptive Statistics for Summarising Data

The first broad category of statistics we discuss concerns *descriptive statistics*. The purpose of the procedures and fundamental concepts in this category is quite straightforward: to facilitate the description and summarisation of data. By 'describe' we generally mean either the use of some pictorial or graphical representation of the data or the computation of an index or number designed to summarise a specific characteristic of a variable or measurement.

We seldom interpret individual data points or observations primarily because it is too difficult for the human brain to extract or identify the essential nature, patterns, or trends evident in the data, particularly if the sample is large. Rather we utilise procedures and measures which provide a general depiction of how the data are behaving. These statistical procedures are designed to identify or display specific patterns or trends in the data. What remains after their application is simply for us to interpret and tell the story.

Reflect on the QCI research scenario and the associated data set discussed in *Chap. 4*. Consider the following questions that Maree might wish to address with respect to decision accuracy and speed scores:

- What was the typical level of accuracy and decision speed for inspectors in the sample? [see *Procedure 5.4* – Assessing central tendency.]
- What was the most common accuracy and speed score amongst the inspectors? [see *Procedure 5.4* – Assessing central tendency.]
- What was the range of accuracy and speed scores; the lowest and the highest scores? [see *Procedure 5.5* – Assessing variability.]
- How frequently were different levels of inspection accuracy and speed observed? What was the shape of the distribution of inspection accuracy and speed scores? [see *Procedure 5.1* – Frequency tabulation, distributions & crosstabulation.]

(continued)

R. W. Cooksey, *Illustrating Statistical Procedures: Finding Meaning in Quantitative Data*, https://doi.org/10.1007/978-981-15-2537-7_5

- What percentage of inspectors would have 'failed' to 'make the cut' assuming the industry standard for acceptable inspection accuracy and speed combined was set at 95%? [see *Procedure 5.7* – Standard (z) scores.]
- How variable were the inspectors in their accuracy and speed scores? Were all the accuracy and speed levels relatively close to each other in magnitude or were the scores widely spread out over the range of possible test outcomes? [see *Procedure 5.5* – Assessing variability.]
- What patterns might be visually detected when looking at various QCI variables singly and together as a set? [see *Procedure 5.2* – Graphical methods for dispaying data, *Procedure 5.3* – Multivariate graphs & displays, and *Procedure 5.6* – Exploratory data analysis.]

This chapter includes discussions and illustrations of a number of procedures available for answering questions about data like those posed above. In addition, you will find discussions of two fundamental concepts, namely *probability* and the *normal distribution*; concepts that provide building blocks for *Chaps. 6* and *7*.

Procedure 5.1: Frequency Tabulation, Distributions & Crosstabulation

Classification	Univariate (crosstabulations are bivariate); descriptive.
Purpose	To produce an efficient counting summary of a sample of data points for ease of interpretation.
Measurement level	Any level of measurement can be used for a variable summarised in frequency tabulations and crosstabulations.

Frequency Tabulation and Distributions

Frequency tabulation serves to provide a convenient counting summary for a set of data that facilitates interpretation of various aspects of those data. Basically, frequency tabulation occurs in two stages:

- First, the scores in a set of data are rank ordered from the lowest value to the highest value.
- Second, the number of times each specific score occurs in the sample is counted. This count records the *frequency* of occurrence for that specific data value.

Consider the overall job satisfaction variable, **jobsat**, from the QCI data scenario. Performing frequency tabulation across the 112 Quality Control Inspectors on this variable using the *SPSS Frequencies* procedure (Allen et al. 2019, ch. 3; George and Mallery 2019, ch. 6) produces the frequency tabulation shown in Table 5.1. Note that three of the inspectors in the sample did not provide a rating for **jobsat** thereby producing three missing values (= 2.7% of the sample of 112) and leaving 109 inspectors with valid data for the analysis.

Table 5.1 Frequency tabulation of overall job satisfaction scores

jobsat

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Very Low	4	3.6	3.7	3.7
	2	8	7.1	7.3	11.0
	3	8	7.1	7.3	18.3
	4 Neutral	18	16.1	16.5	34.9
	5	19	17.0	17.4	52.3
	6	34	30.4	31.2	83.5
	7 Very High	18	16.1	16.5	100.0
	Total	109	97.3	100.0	
Missing	System	3	2.7		
Total		112	100.0		

The display of frequency tabulation is often referred to as the *frequency distribution* for the sample of scores. For each value of a variable, the frequency of its occurrence in the sample of data is reported. It is possible to compute various percentages and percentile values from a frequency distribution.

Table 5.1 shows the 'Percent' or *relative frequency* of each score (the percentage of the 112 inspectors obtaining each score, including those inspectors who were missing scores, which SPSS labels as 'System' missing). Table 5.1 also shows the 'Valid Percent' which is computed only for those inspectors in the sample who gave a valid or non-missing response.

Finally, it is possible to add up the 'Valid Percent' values, starting at the low score end of the distribution, to form the *cumulative distribution* or 'Cumulative Percent'. A cumulative distribution is useful for finding *percentiles* which reflect what percentage of the sample scored at a specific value or below.

We can see in Table 5.1 that 4 of the 109 valid inspectors (a 'Valid Percent' of 3.7%) indicated the lowest possible level of job satisfaction—a value of 1 (Very Low) – whereas 18 of the 109 valid inspectors (a 'Valid Percent' of 16.5%) indicated the highest possible level of job satisfaction—a value of 7 (Very High). The 'Cumulative Percent' number of 18.3 in the row for the job satisfaction score of 3 can be interpreted as "roughly 18% of the sample of inspectors reported a job satisfaction score of 3 or less"; that is, nearly a fifth of the sample expressed some degree of negative satisfaction with their job as a quality control inspector in their particular company.

If you have a large data set having many different scores for a particular variable, it may be more useful to tabulate frequencies on the basis of intervals of scores.

For the **accuracy** scores in the QCI database, you could count scores occurring in intervals such as 'less than 75% accuracy', 'between 75% but less than 85% accuracy', 'between 85% but less than 95% accuracy', and '95% accuracy or greater', rather than counting the individual scores themselves. This would yield what is termed a 'grouped' frequency distribution since the data have been grouped into intervals or score classes. Producing such an analysis using *SPSS* would involve extra steps to create the new category or 'grouping' system for scores prior to conducting the frequency tabulation.

Crosstabulation

In a *frequency crosstabulation,* we count frequencies on the basis of two variables simultaneously rather than one; thus we have a bivariate situation.

For example, Maree might be interested in the number of male and female inspectors in the sample of 112 who obtained each **jobsat** score. Here there are two variables to consider: inspector's **gender** and inspector's j**obsat** score. Table 5.2 shows such a crosstabulation as compiled by the SPSS *Crosstabs* procedure (George and Mallery 2019, ch. 8). Note that inspectors who did not report a score for **jobsat** and/or **gender** have been omitted as missing values, leaving 106 valid inspectors for the analysis.

(continued)

Table 5.2 Frequency crosstabulation of **jobsat** scores by **gender** category for the QCI data

jobsat * gender Crosstabulation

			gender		
			1 Male	2 Female	Total
jobsat	1 Very Low	Count	2	1	3
		% within jobsat	66.7%	33.3%	100.0%
		% within gender	3.5%	2.0%	2.8%
	2	Count	3	5	8
		% within jobsat	37.5%	62.5%	100.0%
		% within gender	5.3%	10.2%	7.5%
	3	Count	2	6	8
		% within jobsat	25.0%	75.0%	100.0%
		% within gender	3.5%	12.2%	7.5%
	4 Neutral	Count	11	7	18
		% within jobsat	61.1%	38.9%	100.0%
		% within gender	19.3%	14.3%	17.0%
	5	Count	14	5	19
		% within jobsat	73.7%	26.3%	100.0%
		% within gender	24.6%	10.2%	17.9%
	6	Count	17	15	32
		% within jobsat	53.1%	46.9%	100.0%
		% within gender	29.8%	30.6%	30.2%
	7 Very High	Count	8	10	18
		% within jobsat	44.4%	55.6%	100.0%
		% within gender	14.0%	20.4%	17.0%
Total		Count	57	49	106
		% within jobsat	53.8%	46.2%	100.0%
		% within gender	100.0%	100.0%	100.0%

The crosstabulation shown in Table 5.2 gives a composite picture of the distribution of satisfaction levels for male inspectors and for female inspectors. If frequencies or 'Counts' are added across the **gender** categories, we obtain the numbers in the 'Total' column (the percentages or relative frequencies are also shown immediately below each count) for each discrete value of **jobsat** (note this column of statistics differs from that in Table 5.1 because the **gender** variable was missing for certain inspectors). By adding down each **gender** column, we obtain, in the bottom row labelled 'Total', the number of males and the number of females that comprised the sample of 106 valid inspectors.

The totals, either across the rows or down the columns of the crosstabulation, are termed the *marginal distributions* of the table. These

(continued)

marginal distributions are equivalent to frequency tabulations for each of the variables **jobsat** and **gender**. As with frequency tabulation, various percentage measures can be computed in a crosstabulation, including the percentage of the sample associated with a specific count within either a row ('% within **jobsat**') or a column ('% within **gender**'). You can see in Table 5.2 that 18 inspectors indicated a job satisfaction level of 7 (Very High); of these 18 inspectors reported in the 'Total' column, 8 (44.4%) were male and 10 (55.6%) were female. The marginal distribution for **gender** in the 'Total' row shows that 57 inspectors (53.8% of the 106 valid inspectors) were male and 49 inspectors (46.2%) were female. Of the 57 male inspectors in the sample, 8 (14.0%) indicated a job satisfaction level of 7 (Very High). Furthermore, we could generate some additional interpretive information of value by adding the '% within gender' values for job satisfaction levels of 5, 6 and 7 (i.e. differing degrees of positive job satisfaction). Here we would find that 68.4% (= 24.6% + 29.8% + 14.0%) of male inspectors indicated some degree of positive job satisfaction compared to 61.2% (= 10.2% + 30.6% + 20.4%) of female inspectors.

This helps to build a picture of the possible relationship between an inspector's gender and their level of job satisfaction (a relationship that, as we will see later, can be quantified and tested using *Procedure 6.2* and *Procedure 7.1*).

It should be noted that a crosstabulation table such as that shown in Table 5.2 is often referred to as a *contingency table* about which more will be said later (see *Procedure 7.1* and *Procedure 7.18*).

Advantages

Frequency tabulation is useful for providing convenient data summaries which can aid in interpreting trends in a sample, particularly where the number of discrete values for a variable is relatively small. A cumulative percent distribution provides additional interpretive information about the relative positioning of specific scores within the overall distribution for the sample.

Crosstabulation permits the simultaneous examination of the distributions of values for two variables obtained from the same sample of observations. This examination can yield some useful information about the possible relationship between the two variables. More complex crosstabulations can be also done where the values of three or more variables are tracked in a single systematic summary. The use of frequency tabulation or cross-tabulation in conjunction with various other statistical measures, such as measures of central tendency (see *Procedure 5.4*) and measures of variability (see *Procedure 5.5*), can provide a relatively complete descriptive summary of any data set.

Disadvantages

Frequency tabulations can get messy if interval or ratio-level measures are tabulated simply because of the large number of possible data values. Grouped frequency distributions really should be used in such cases. However, certain choices, such as the size of the score interval (group size), must be made, often arbitrarily, and such choices can affect the nature of the final frequency distribution.

Additionally, percentage measures have certain problems associated with them, most notably, the potential for their misinterpretation in small samples. One should be sure to know the sample size on which percentage measures are based in order to obtain an interpretive reference point for the actual percentage values.

For example *In a sample of 10 individuals, 20% represents only two individuals whereas in a sample of 300 individuals, 20% represents 60 individuals. If all that is reported is the 20%, then the mental inference drawn by readers is likely to be that a sizeable number of individuals had a score or scores of a particular value—but what is 'sizeable' depends upon the total number of observations on which the percentage is based.*

Where Is This Procedure Useful?

Frequency tabulation and crosstabulation are very commonly applied procedures used to summarise information from questionnaires, both in terms of tabulating various demographic characteristics (e.g. gender, age, education level, occupation) and in terms of actual responses to questions (e.g. numbers responding 'yes' or 'no' to a particular question). They can be particularly useful in helping to build up the data screening and demographic stories discussed in *Chap. 4*. Categorical data from observational studies can also be analysed with this technique (e.g. the number of times Suzy talks to Frank, to Billy, and to John in a study of children's social interactions).

Certain types of experimental research designs may also be amenable to analysis by crosstabulation with a view to drawing inferences about distribution differences across the sets of categories for the two variables being tracked.

You could employ crosstabulation in conjunction with the tests described in *Procedure 7.1* to see if two different styles of advertising campaign differentially affect the product purchasing patterns of male and female consumers.

In the QCI database, Maree could employ crosstabulation to help her answer the question "do different types of electronic manufacturing firms (**company**) differ in terms of their tendency to employ male versus female quality control inspectors (**gender**)?"

Software Procedures

Application	Procedures
SPSS	*Analyze* → *Descriptive Statistics* → *Frequencies...* or *Crosstabs...* and select the variable(s) you wish to analyse; for the *Crosstabs* procedure, hitting the '*Cells*' button will allow you to choose various types of statistics and percentages to show in each cell of the table.
NCSS	*Analysis* → *Descriptive Statistics* → *Frequency Tables* or *Cross Tabulation* and select the variable(s) you wish to analyse.
SYSTAT	*Analyze* → *One-Way Frequency Tables...* or *Tables* → *Two-Way...* and select the variable(s) you wish to analyse and choose the optional statistics you wish to see.
STATGRAPHICS	*Describe* → *Categorical Data* → *Tabulation* or → *Crosstabulation* and select the variable(s) you wish to analyse; hit '*OK*' and when the 'Tables and Graphs' window opens, choose the Tables and Graphs you wish to see.
R Commander	*Statistics* → *Summaries* → *Frequency Tables...* or *Crosstabulation* → *Two-way table...* and select the variable(s) you wish to analyse and choose the optional statistics you wish to see.

Procedure 5.2: Graphical Methods for Displaying Data

Classification Univariate (scatterplots are bivariate); descriptive.
Purpose To visually summarise characteristics of a data sample for ease of interpretation.
Measurement level Any level of measurement can be accommodated by these graphical methods. Scatterplots are generally used for interval or ratio-level data.

Graphical methods for displaying data include bar and pie charts, histograms and frequency polygons, line graphs and scatterplots. It is important to note that what is presented here is a small but representative sampling of the types of simple graphs one can produce to summarise and display trends in data. Generally speaking, SPSS offers the easiest facility for producing and editing graphs, but with a rather limited range of styles and types. SYSTAT, STATGRAPHICS and NCSS offer a much wider range of graphs (including graphs unique to each package), but with the drawback that it takes somewhat more effort to get the graphs in exactly the form you want.

Bar and Pie Charts

These two types of graphs are useful for summarising the frequency of occurrence of various values (or ranges of values) where the data are categorical (nominal or ordinal level of measurement).

- A *bar chart* uses vertical and horizontal axes to summarise the data. The vertical axis is used to represent frequency (number) of occurrence or the relative frequency (percentage) of occurrence; the horizontal axis is used to indicate the data categories of interest.
- A *pie chart* gives a simpler visual representation of category frequencies by cutting a circular plot into wedges or slices whose sizes are proportional to the relative frequency (percentage) of occurrence of specific data categories. Some pie charts can have a one or more slices emphasised by 'exploding' them out from the rest of the pie.

Consider the **company** variable from the QCI database. This variable depicts the types of manufacturing firms that the quality control inspectors worked for. Figure 5.1 illustrates a bar chart summarising the percentage of female inspectors in the sample coming from each type of firm. Figure 5.2 shows a pie chart representation of the same data, with an 'exploded slice' highlighting the percentage of female inspectors in the sample who worked for large business computer manufacturers – the lowest percentage of the five types of companies. Both graphs were produced using SPSS.

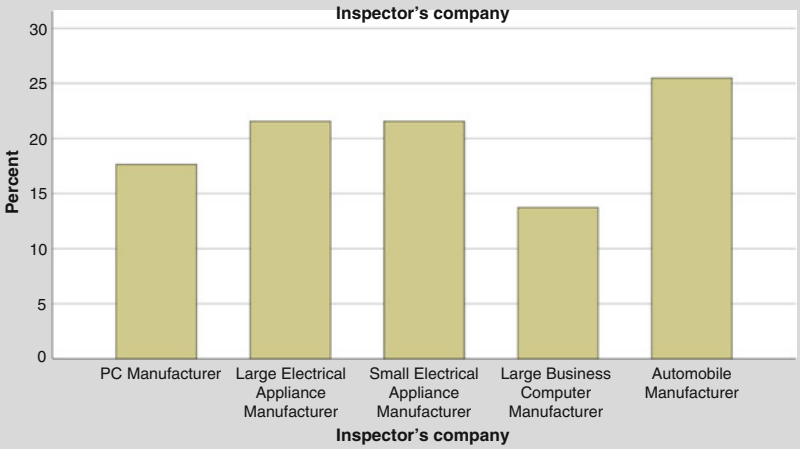

Fig. 5.1 Bar chart: Percentage of female inspectors

(continued)

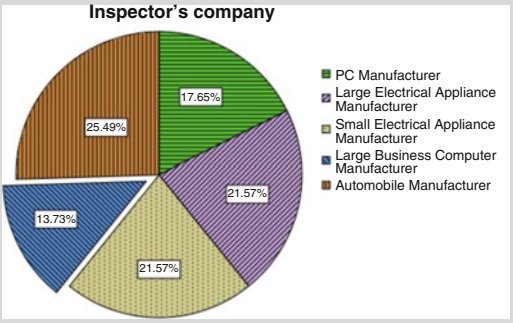

Fig. 5.2 Pie chart: Percentage of female inspectors

The pie chart was modified with an option to show the actual percentage along with the label for each category. The bar chart shows that computer manufacturing firms have relatively fewer female inspectors compared to the automotive and electrical appliance (large and small) firms. This trend is less clear from the pie chart which suggests that pie charts may be less visually interpretable when the data categories occur with rather similar frequencies. However, the 'exploded slice' option can help interpretation in some circumstances.

Certain software programs, such as SPSS, STATGRAPHICS, NCSS and Microsoft Excel, offer the option of generating 3-dimensional bar charts and pie charts and incorporating other 'bells and whistles' that can potentially add visual richness to the graphic representation of the data. However, you should generally be careful with these fancier options as they can produce distortions and create ambiguities in interpretation (e.g. see discussions in Jacoby 1997; Smithson 2000; Wilkinson 2009). Such distortions and ambiguities could ultimately end up providing misinformation to researchers as well as to those who read their research.

Histograms and Frequency Polygons

These two types of graphs are useful for summarising the frequency of occurrence of various values (or ranges of values) where the data are essentially continuous (interval or ratio level of measurement) in nature. Both histograms and frequency polygons use vertical and horizontal axes to summarise the data. The vertical axis is used to represent the frequency (number) of occurrence or the relative frequency (percentage) of occurrences; the horizontal axis is used for the data values or ranges of values of interest. The *histogram* uses bars of varying heights to depict frequency; the *frequency polygon* uses lines and points.

There is a visual difference between a histogram and a bar chart: the bar chart uses bars that do not physically touch, signifying the discrete and categorical nature of the data, whereas the bars in a histogram physically touch to signal the potentially continuous nature of the data.

Suppose Maree wanted to graphically summarise the distribution of **speed** scores for the 112 inspectors in the QCI database. Figure 5.3 (produced using NCSS) illustrates a histogram representation of this variable. Figure 5.3 also illustrates another representational device called the 'density plot' (the solid tracing line overlaying the histogram) which gives a smoothed impression of the overall shape of the distribution of **speed** scores. Figure 5.4 (produced using STATGRAPHICS) illustrates the frequency polygon representation for the same data.

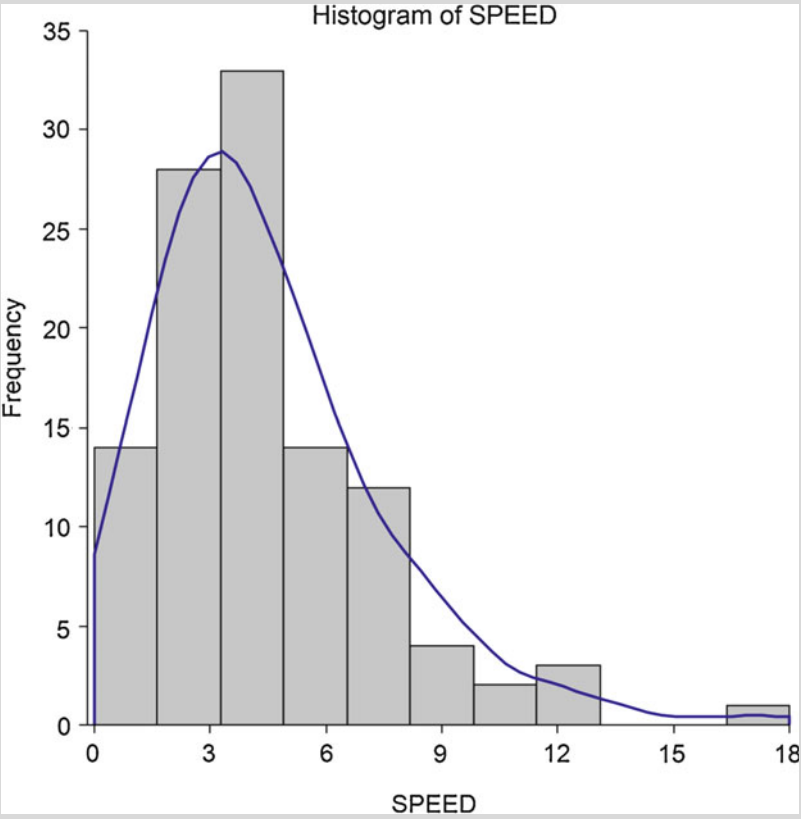

Fig. 5.3 Histogram of the **speed** variable (with density plot overlaid)

These graphs employ a grouped format where **speed** scores which fall within specific intervals are counted as being essentially the same score. The shape of the data distribution is reflected in these plots. Each graph tells us that the inspection **speed** scores are positively skewed with only a few inspectors taking very long times to make their inspection judgments and the majority of inspectors taking rather shorter amounts of time to make their decisions.

(continued)

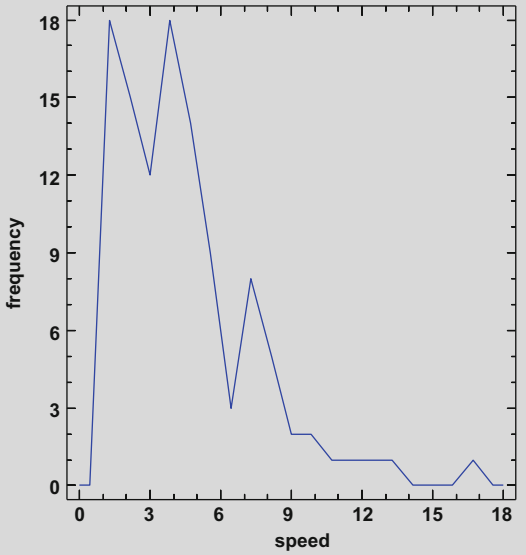

Fig. 5.4 Frequency polygon plot of the **speed** variable

Both representations tell a similar story; the choice between them is largely a matter of personal preference. However, if the number of bars to be plotted in a histogram is potentially very large (and this is usually directly controllable in most statistical software packages), then a frequency polygon would be the preferred representation simply because the amount of visual clutter in the graph will be much reduced.

It is somewhat of an art to choose an appropriate definition for the width of the score grouping intervals (or 'bins' as they are often termed) to be used in the plot: choose too many and the plot may look too lumpy and the overall distributional trend may not be obvious; choose too few and the plot will be too coarse to give a useful depiction. Programs like SPSS, SYSTAT, STATGRAPHICS and NCSS are designed to choose an 'appropriate' number of bins to be used, but the analyst's eye is often a better judge than any statistical rule that a software package would use.

There are several interesting variations of the histogram which can highlight key data features or facilitate interpretation of certain trends in the data. One such variation is a graph is called a **dual histogram** (available in SYSTAT; a variation called a 'comparative histogram' can be created in NCSS) – a graph that facilitates visual comparison of the frequency distributions for a specific variable for participants from two distinct groups.

Suppose Maree wanted to graphically compare the distributions of **speed** scores for inspectors in the two categories of education level (**educlev**) in the QCI database. Figure 5.5 shows a dual histogram (produced using SYSTAT) that accomplishes this goal. This graph still employs the grouped

(continued)

format where **speed** scores falling within particular intervals are counted as being essentially the same score. The shape of the data distribution within each group is also clearly reflected in this plot. However, the story conveyed by the dual histogram is that, while the inspection **speed** scores are positively skewed for inspectors in both categories of **educlev,** the comparison suggests that inspectors with a high school level of education (= 1) tend to take slightly longer to make their inspection decisions than do their colleagues who have a tertiary qualification (= 2).

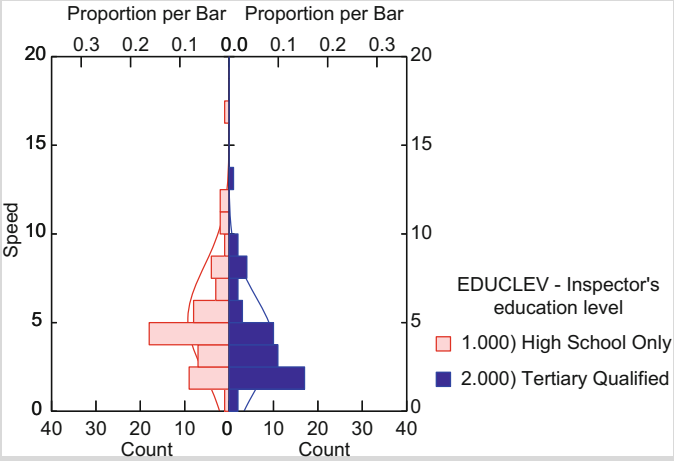

Fig. 5.5 Dual histogram of *speed* for the two categories of **educlev**

Line Graphs

The *line graph* is similar in style to the frequency polygon but is much more general in its potential for summarising data. In a line graph, we seldom deal with percentage or frequency data. Instead we can summarise other types of information about data such as averages or means (see *Procedure 5.4* for a discussion of this measure), often for different groups of participants. Thus, one important use of the line graph is to break down scores on a specific variable according to membership in the categories of a second variable.

In the context of the QCI database, Maree might wish to summarise the average inspection **accuracy** scores for the inspectors from different types of manufacturing companies. Figure 5.6 was produced using SPSS and shows such a line graph.

(continued)

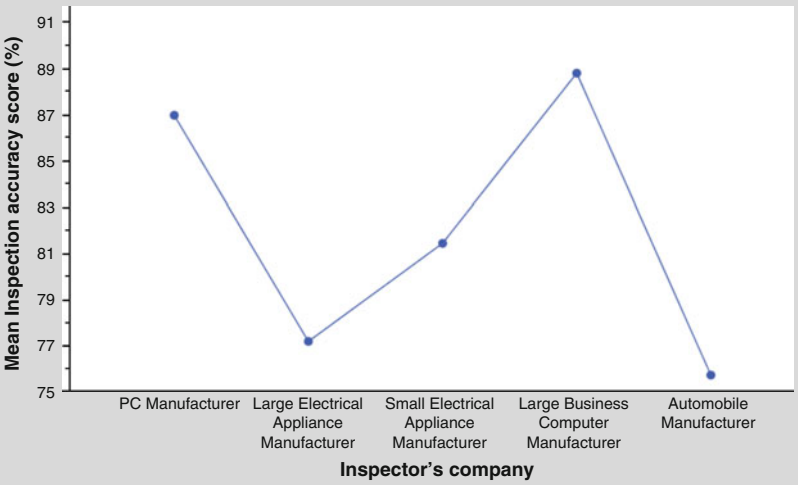

Fig. 5.6 Line graph comparison of companies in terms of average inspection **accuracy**

Note how the trend in performance across the different companies becomes clearer with such a visual representation. It appears that the inspectors from the Large Business Computer and PC manufacturing companies have better average inspection accuracy compared to the inspectors from the remaining three industries.

With many software packages, it is possible to further elaborate a line graph by including error or confidence intervals bars (see *Procedure 8.3*). These give some indication of the precision with which the average level for each category in the population has been estimated (narrow bars signal a more precise estimate; wide bars signal a less precise estimate).

Figure 5.7 shows such an elaborated line graph, using 95% confidence interval bars, which can be used to help make more defensible judgments (compared to Fig. 5.6) about whether the companies are substantively different from each other in average inspection performance. Companies whose confidence interval bars do not overlap each other can be inferred to be substantively different in performance characteristics.

The accuracy confidence interval bars for participants from the Large Business Computer manufacturing firms do not overlap those from the Large or Small Electrical Appliance manufacturers or the Automobile manufacturers.

(continued)

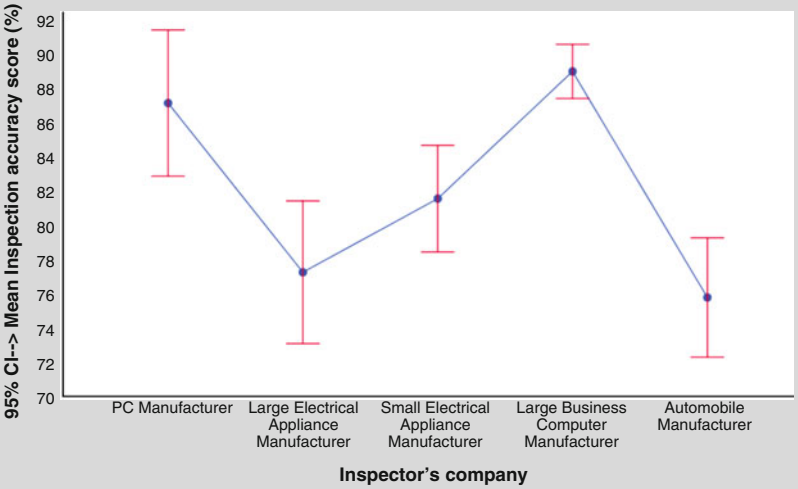

Fig. 5.7 Line graph using confidence interval bars to compare **accuracy** across companies

We might conclude that quality control inspection accuracy is substantially better in the Large Business Computer manufacturing companies than in these other industries but is not substantially better than the PC manufacturing companies. We might also conclude that inspection accuracy in PC manufacturing companies is not substantially different from Small Electrical Appliance manufacturers.

Scatterplots

Scatterplots are useful in displaying the relationship between two interval- or ratio-scaled variables or measures of interest obtained on the same individuals, particularly in correlational research (see *Fundamental Concept III* and *Procedure 6.1*).

In a scatterplot, one variable is chosen to be represented on the horizontal axis; the second variable is represented on the vertical axis. In this type of plot, all data point pairs in the sample are graphed. The shape and tilt of the cloud of points in a scatterplot provide visual information about the strength and direction of the relationship between the two variables. A very compact elliptical cloud of points signals a strong relationship; a very loose or nearly circular cloud signals a weak or non-existent relationship. A cloud of points generally tilted upward toward the right side of the graph signals a positive relationship (higher scores on one variable associated with higher scores on the other and vice-versa). A cloud of points generally tilted downward toward the right side of the graph signals a negative relationship (higher scores on one variable associated with lower scores on the other and vice-versa).

Maree might be interested in displaying the relationship between inspection **accuracy** and inspection **speed** in the QCI database. Figure 5.8, produced using SPSS, shows what such a scatterplot might look like. Several characteristics of the data for these two variables can be noted in Fig. 5.8. The shape of the distribution of data points is evident. The plot has a fan-shaped characteristic to it which indicates that accuracy scores are highly variable (exhibit a very wide range of possible scores) at very fast inspection speeds but get much less variable and tend to be somewhat higher as inspection speed increases (where inspectors take longer to make their quality control decisions). Thus, there does appear to be some relationship between inspection **accuracy** and inspection **speed** (a weak positive relationship since the cloud of points tends to be very loose but tilted generally upward toward the right side of the graph – slower speeds tend to be slightly associated with higher accuracy.

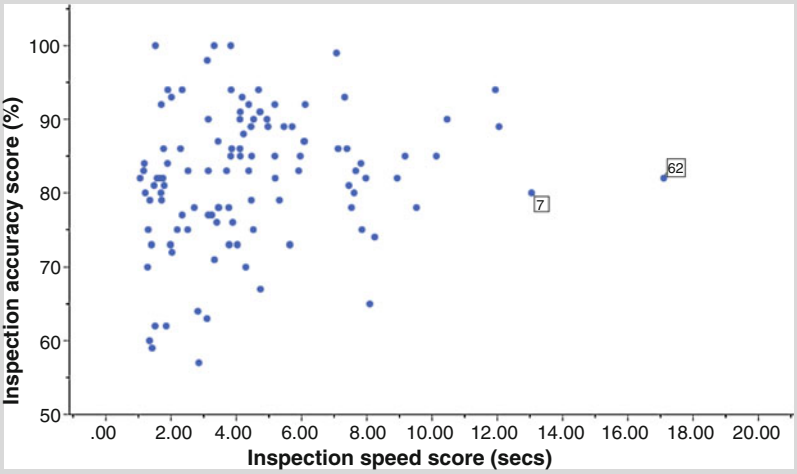

Fig. 5.8 Scatterplot relating inspection **accuracy** to inspection **speed**

However, it is not the case that the inspection decisions which take longest to make are necessarily the most accurate (see the labelled points for inspectors 7 and 62 in Fig. 5.8). Thus, Fig. 5.8 does not show a simple relationship that can be unambiguously summarised by a statement like "the longer an inspector takes to make a quality control decision, the more accurate that decision is likely to be". The story is more complicated.

Some software packages, such as SPSS, STATGRAPHICS and SYSTAT, offer the option of using different plotting symbols or *markers* to represent the members of different groups so that the relationship between the two focal variables (the ones anchoring the X and Y axes) can be clarified with reference to a third categorical measure.

Maree might want to see if the relationship depicted in Fig. 5.8 changes depending upon whether the inspector was tertiary-qualified or not (this information is represented in the **educlev** variable of the QCI database).

Figure 5.9 shows what such a modified scatterplot might look like; the legend in the upper corner of the figure defines the marker symbols for each category of the **educlev** variable. Note that for both High School only-educated inspectors and Tertiary-qualified inspectors, the general fan-shaped relationship between **accuracy** and **speed** is the same. However, it appears that the distribution of points for the High School only-educated inspectors is shifted somewhat upward and toward the right of the plot suggesting that these inspectors tend to be somewhat more accurate as well as slower in their decision processes.

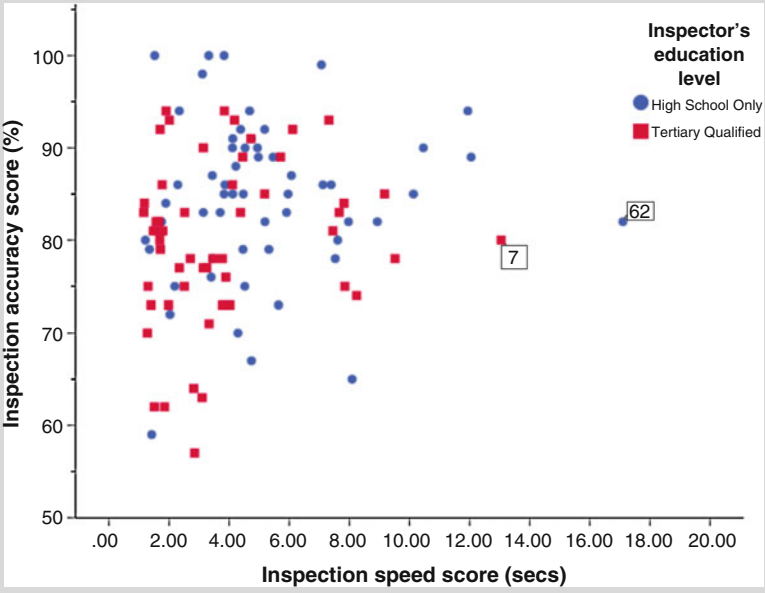

Fig. 5.9 Scatterplot displaying **accuracy** vs **speed** conditional on **educlev** group

There are many other styles of graphs available, often dependent upon the specific statistical package you are using. Interestingly, NCSS and, particularly, SYSTAT and STATGRAPHICS, appear to offer the most variety in terms of types of graphs available for visually representing data. A reading of the user's manuals for these programs (see the Useful additional readings) would expose you to the great diversity of plotting techniques available to researchers. Many of these techniques go by rather interesting names such as: Chernoff's faces, radar plots, sunflower plots, violin plots, star plots, Fourier blobs, and dot plots.

Advantages

These graphical methods provide summary techniques for visually presenting certain characteristics of a set of data. Visual representations are generally easier to understand than a tabular representation and when these plots are combined with available numerical statistics, they can give a very complete picture of a sample of data. Newer methods have become available which permit more complex representations to be depicted, opening possibilities for creatively visually representing more aspects and features of the data (leading to a style of visual data storytelling called *infographics*; see, for example, McCandless 2014; Toseland and Toseland 2012). Many of these newer methods can display data patterns from multiple variables in the same graph (several of these newer graphical methods are illustrated and discussed in *Procedure 5.3*).

Disadvantages

Graphs tend to be cumbersome and space consuming if a great many variables need to be summarised. In such cases, using numerical summary statistics (such as means or correlations) in tabular form alone will provide a more economical and efficient summary. Also, it can be very easy to give a misleading picture of data trends using graphical methods by simply choosing the 'correct' scaling for maximum effect or choosing a display option (such as a 3-D effect) that 'looks' presentable but which actually obscures a clear interpretation (see Smithson 2000; Wilkinson 2009).

Thus, you must be careful in creating and interpreting visual representations so that the influence of aesthetic choices for sake of appearance do not become more important than obtaining a faithful and valid representation of the data—a very real danger with many of today's statistical packages where 'default' drawing options have been pre-programmed in. No single plot can completely summarise all possible characteristics of a sample of data. Thus, choosing a specific method of graphical display may, of necessity, force a behavioural researcher to represent certain data characteristics (such as frequency) at the expense of others (such as averages).

Where Is This Procedure Useful?

Virtually any research design which produces quantitative data and statistics (even to the extent of just counting the number of occurrences of several events) provides opportunities for graphical data display which may help to clarify or illustrate important data characteristics or relationships. Remember, graphical displays are communication tools just like numbers—which tool to choose depends upon the message to be conveyed. Visual representations of data are generally more useful in

communicating to lay persons who are unfamiliar with statistics. Care must be taken though as these same lay people are precisely the people most likely to misinterpret a graph if it has been incorrectly drawn or scaled.

Software Procedures

Application	Procedures
SPSS	*Graphs → Chart Builder*... and choose from a range of gallery chart types: *Bar*, *Pie/Polar, Histogram, Line, Scatter/Dot*; drag the chart type into the working area and customise the chart with desired variables, labels, etc. many elements of a chart, including error bars, can be controlled.
NCSS	*Graphics → Bar Charts → Bar Charts* or *Graphics → Pie Charts* or *Graphics → Histograms → Histograms* or *Histograms – Comparative* or *Graphics → Error Bar Charts → Error Bar Charts* or *Graphics → Scatter Plots → Scatter Plots; w*hichever type of chart you choose, you can control many features of the chart from the dialog box that pops open upon selection.
STATGRAPHICS	*Plot → Business Charts → Barchart...* or *Piechart...* or *Plot → Exploratory Plots → Frequency Histogram...* or *Plot → Scatterplots → X – Y Plot...; w*hichever type of chart you choose, you can control a number of features of the chart from the series of dialog boxes that pops open upon selection.
SYSTAT	*Graph → Bar*... or *Pie*... or *Histogram*... or *Line*... or *Scatterplot*... or *Graph Gallery...* (which offers a range of other more novel graphical displays, including the dual histogram). For each choice, a dialog box opens which allows you to control almost every characteristic of the graph you want.
R Commander	*Graphs → Bar graph* or *Pie chart* or *Histogram* or *Scatterplot* or *Plot of Means*; for some graphs (*Scatterplot* being the exception), there is minimal control offered by **R** Commander over the appearance of the graph (you need to use full **R** commands to control more aspects; e.g. see Chang 2019).

Procedure 5.3: Multivariate Graphs & Displays

Classification Multivariate; descriptive.

Purpose To simultaneously and visually summarise characteristics of many variables obtained on the same entities for ease of interpretation.

Measurement level Multivariate graphs and displays are generally produced using interval or ratio-level data. However, such graphs may be grouped according to a nominal or ordinal categorical variable for comparison purposes.

Graphical methods for displaying multivariate data (i.e. many variables at once) include scatterplot matrices, radar (or spider) plots, multiplots, parallel coordinate displays, and icon plots. Multivariate graphs are useful for visualising broad trends and patterns across many variables (Cleveland 1995; Jacoby 1998). Such graphs typically sacrifice precision in representation in favour of a snapshot pictorial summary that can help you form general impressions of data patterns.

It is important to note that what is presented here is a small but reasonably representative sampling of the types of graphs one can produce to summarise and display trends in multivariate data. Generally speaking, SYSTAT offers the best facilities for producing multivariate graphs, followed by STATGRAPHICS, but with the drawback that it is somewhat tricky to get the graphs in exactly the form you want. SYSTAT also has excellent facilities for creating new forms and combinations of graphs – essentially allowing graphs to be tailor-made for a specific communication purpose. Both SPSS and NCSS offer a more limited range of multivariate graphs, generally restricted to scatterplot matrices and variations of multiplots. Microsoft Excel or STATGRAPHICS are the packages to use if radar or spider plots are desired.

Scatterplot Matrices

A *scatterplot matrix* is a useful multivariate graph designed to show relationships between pairs of many variables in the same display.

Figure 5.10 illustrates a scatterplot matrix, produced using SYSTAT, for the **mentabil**, **accuracy**, **speed**, **jobsat** and **workcond** variables in the QCI database. It is easy to see that all the scatterplot matrix does is stack all pairs of scatterplots into a format where it is easy to pick out the graph for any 'row' variable that intersects a column 'variable'.

In those plots where a 'row' variable intersects itself in a column of the matrix (along the so-called 'diagonal'), SYSTAT permits a range of univariate displays to be shown. Figure 5.10 shows univariate histograms for each variable (recall *Procedure 5.2*). One obvious drawback of the scatterplot matrix is that, if many variables are to be displayed (say ten or more); the graph gets very crowded and becomes very hard to visually appreciate.

Looking at the first column of graphs in Fig. 5.10, we can see the scatterplot relationships between **mentabil** and each of the other variables. We can get a

(continued)

visual impression that **mentabil** seems to be slightly negatively related to **accuracy** (the cloud of scatter points tends to angle downward to the right, suggesting, very slightly, that higher **mentabil** scores are associated with lower levels of **accuracy**).

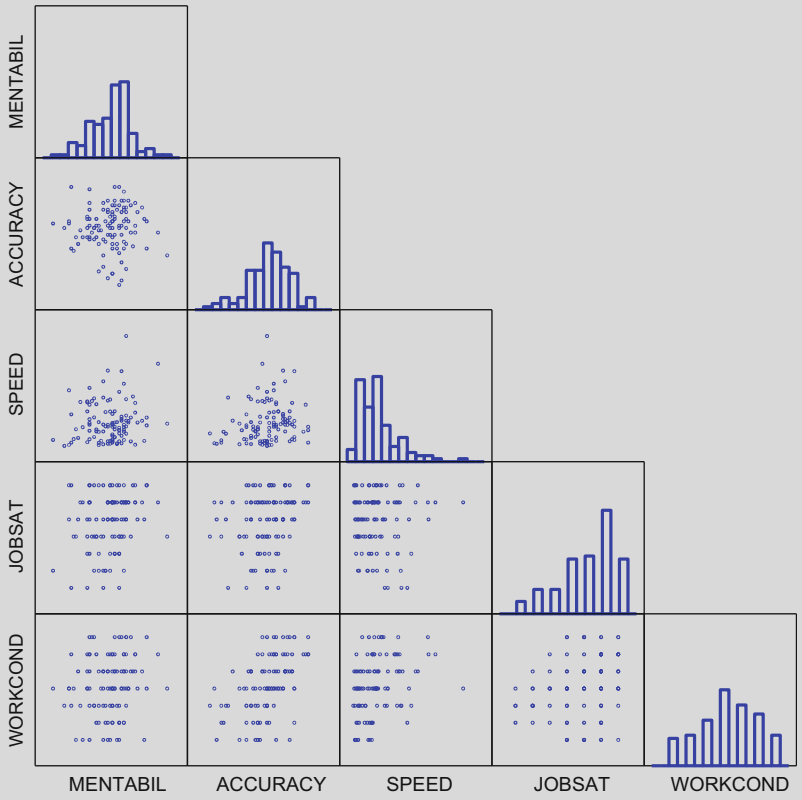

Fig. 5.10 Scatterplot matrix relating **mentabil**, **accuracy**, **speed**, **jobsat** & **workcond**

Conversely, the visual impression of the relationship between **mentabil** and **speed** is that the relationship is slightly positive (higher **mentabil** scores tend to be associated with higher **speed** scores = longer inspection times). Similar types of visual impressions can be formed for other parts of Fig. 5.10. Notice that the histogram plots along the diagonal give a clear impression of the shape of the distribution for each variable.

Radar Plots

The *radar plot* (also known as a *spider graph* for obvious reasons) is a simple and effective device for displaying scores on many variables. Microsoft Excel offers a range of options and capabilities for producing radar plots, such as the plot shown in Fig. 5.11. Radar plots are generally easy to interpret and provide a good visual basis for comparing plots from different individuals or groups, even if a fairly large number of variables (say, up to about 25) are being displayed. Like a clock face, variables are evenly spaced around the centre of the plot in clockwise order starting at the 12 o'clock position. Visual interpretation of a radar plot primarily relies on shape comparisons, i.e. the rise and fall of peaks and valleys along the spokes around the plot. Valleys near the centre display low scores on specific variables, peaks near the outside of the plot display high scores on specific variables. [Note that, technically, radar plots employ polar coordinates.] SYSTAT can draw graphs using polar coordinates but not as easily as Excel can, from the user's perspective. Radar plots work best if all the variables represented are measured on the same scale (e.g. a 1 to 7 Likert-type scale or 0% to 100% scale). Individuals who are missing any scores on the variables being plotted are typically omitted.

The radar plot in Fig. 5.11, produced using Excel, compares two specific inspectors, 66 and 104, on the nine attitude rating scales. Inspector 66 gave the highest rating (= 7) on the **cultqual** variable and inspector 104 gave the lowest rating (= 1). The plot shows that inspector 104 tended to provide very low ratings on all nine attitude variables, whereas inspector 66 tended to give very high ratings on all variables except **acctrain** and **trainapp**, where the scores were similar to those for inspector 104. Thus, in general, inspector 66 tended to show much more positive attitudes toward their workplace compared to inspector 104.

While Fig. 5.11 was generated to compare the scores for two individuals in the QCI database, it would be just as easy to produce a radar plot that compared the five types of companies in terms of their average ratings on the nine variables, as shown in Fig. 5.12.

Here we can form the visual impression that the five types of companies differ most in their average ratings of **mgmtcomm** and least in the average ratings of **polsatis**. Overall, the average ratings from inspectors from PC manufacturers (black diamonds with solid lines) seem to be generally the most positive as their scores lie on or near the outer ring of scores and those

(continued)

from Automobile manufacturers tend to be least positive on many variables (except the training-related variables).

Extrapolating from Fig. 5.12, you may rightly conclude that including too many groups and/or too many variables in a radar plot comparison can lead to so much clutter that any visual comparison would be severely degraded. You may have to experiment with using colour-coded lines to represent different groups versus line and marker shape variations (as used in Fig. 5.12), because choice of coding method for groups can influence the interpretability of a radar plot.

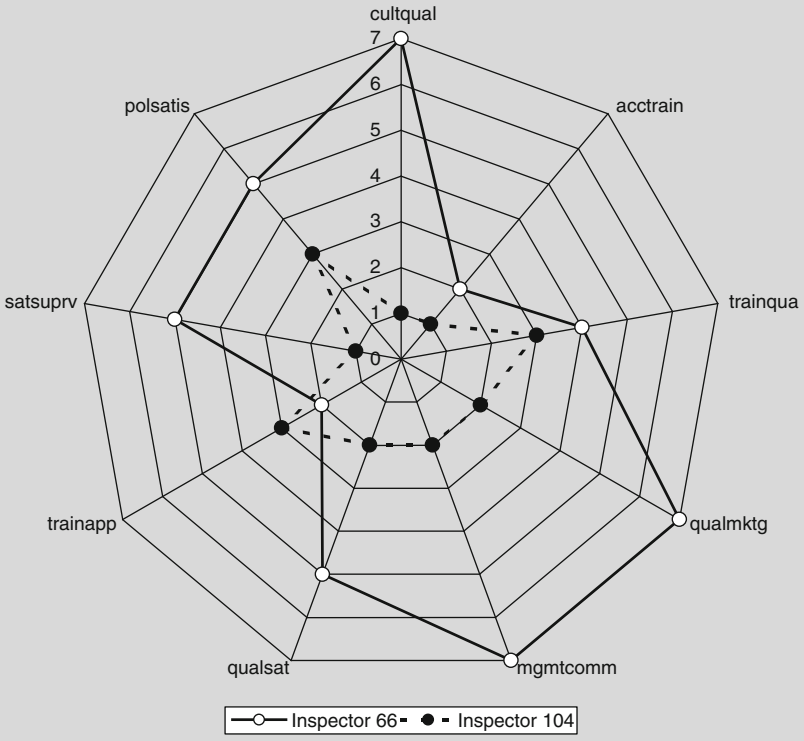

Fig. 5.11 Radar plot comparing attitude ratings for inspectors 66 and 104

(continued)

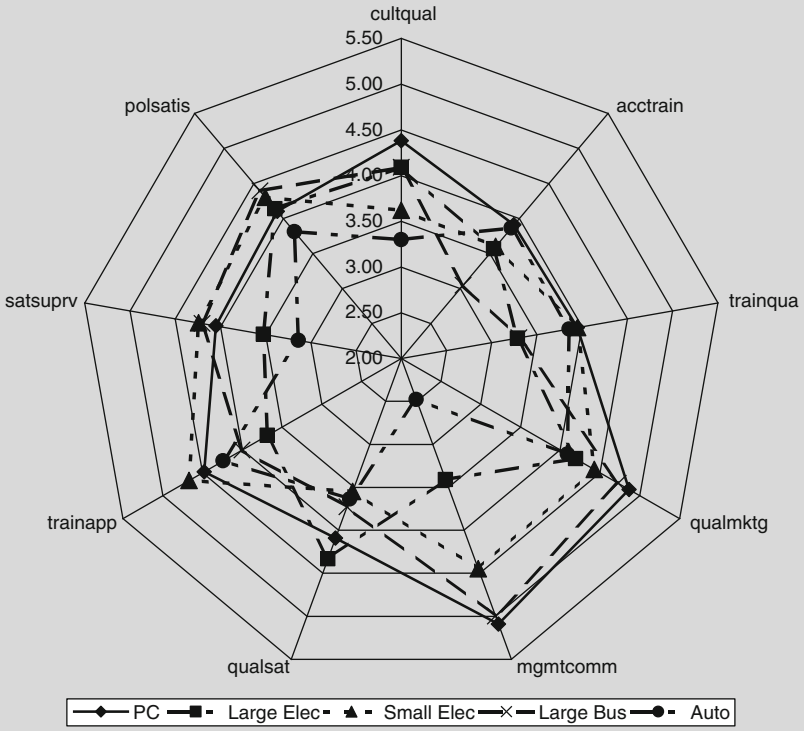

Fig. 5.12 Radar plot comparing average attitude ratings for five types of company

Multiplots

A *multiplot* is simply a hybrid style of graph that can display group comparisons across a number of variables. There are a wide variety of possible multiplots one could potentially design (SYSTAT offers great capabilities with respect to multiplots). Figure 5.13 shows a multiplot comprising a side-by-side series of profile-based line graphs – one graph for each type of company in the QCI database.

The multiplot in Fig. 5.13, produced using SYSTAT, graphs the profile of average attitude ratings for all inspectors within a specific type of company. This multiplot shows the same story as the radar plot in Fig. 5.12, but in a different graphical format. It is still fairly clear that the average ratings from inspectors from PC manufacturers tend to be higher than for the other types of companies and the profile for inspectors from automobile manufacturers tends to be lower than for the other types of companies.

(continued)

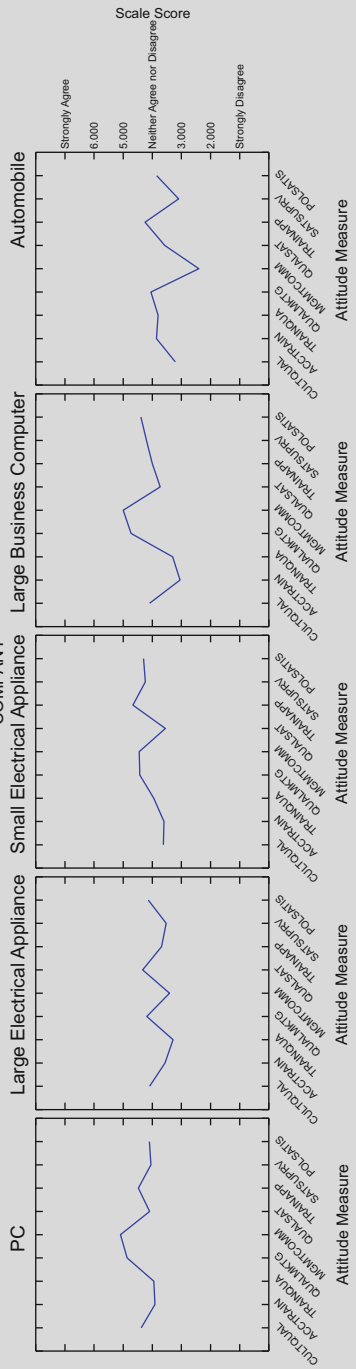

Fig. 5.13 Multiplot comparing profiles of average attitude ratings for five company types

(continued)

The profile for inspectors from large electrical appliance manufacturers is the flattest, meaning that their average attitude ratings were less variable than for other types of companies. Comparing the ease with which you can glean the visual impressions from Figs. 5.12 and 5.13 may lead you to prefer one style of graph over another. If you have such preferences, chances are others will also, which may mean you need to carefully consider your options when deciding how best to display data for effect.

Frequently, choice of graph is less a matter of which style is right or wrong, but more a matter of which style will suit specific purposes or convey a specific story, i.e. the choice is often strategic.

Parallel Coordinate Displays

A *parallel coordinate display* is useful for displaying individual scores on a range of variables, all measured using the same scale. Furthermore, such graphs can be combined side-by-side to facilitate very broad visual comparisons among groups, while retaining individual profile variability in scores. Each line in a parallel coordinate display represents one individual, e.g. an inspector.

The interpretation of a parallel coordinate display, such as the two shown in Fig. 5.14, depends on visual impressions of the peaks and valleys (highs and lows) in the profiles as well as on the density of similar profile lines. The graph is called 'parallel coordinate' simply because it assumes that all variables are measured on the same scale and that scores for each variable can therefore be located along vertical axes that are parallel to each other (imagine vertical lines on Fig. 5.14 running from bottom to top for each variable on the X-axis). The main drawback of this method of data display is that only those individuals in the sample who provided legitimate scores on all of the variables being plotted (i.e. who have no missing scores) can be displayed.

The parallel coordinate display in Fig. 5.14, produced using SYSTAT, graphs the profile of average attitude ratings for all inspectors within two specific types of company: the left graph for inspectors from PC manufacturers and the right graph for automobile manufacturers.

There are fewer lines in each display than the number of inspectors from each type of company simply because several inspectors from each type of company were missing a rating on at least one of the nine attitude variables. The graphs show great variability in scores amongst inspectors within a company type, but there are some overall patterns evident.

(continued)

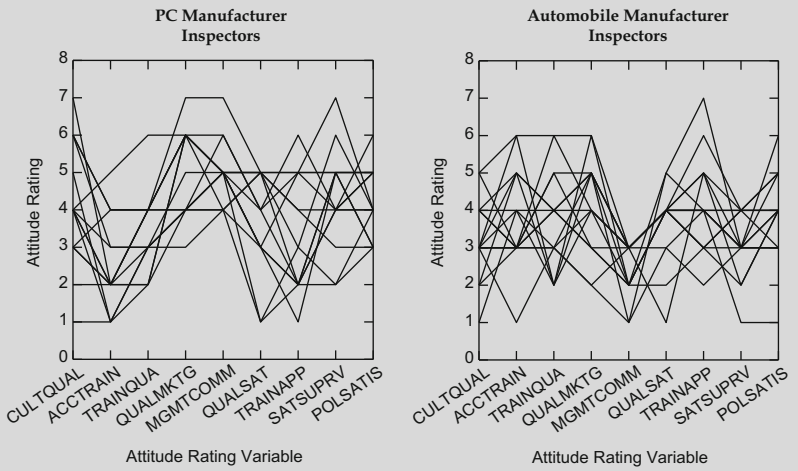

Fig. 5.14 Parallel coordinate displays comparing profiles of average attitude ratings for five company types

For example, inspectors from automobile companies clearly and fairly uniformly rated **mgmtcomm** toward the low end of the scale, whereas the reverse was generally true for that variable for inspectors from PC manufacturers. Conversely, inspectors from automobile companies tend to rate **acctrain** and **trainapp** more toward the middle to high end of the scale, whereas the reverse is generally true for those variables for inspectors from PC manufacturers.

Icon Plots

Perhaps the most creative types of multivariate displays are the so-called *icon plots*. SYSTAT and STATGRAPHICS offer an impressive array of different types of icon plots, including, amongst others, Chernoff's faces, profile plots, histogram plots, star glyphs and sunray plots (Jacoby 1998 provides a detailed discussion of icon plots).

Icon plots generally use a specific visual construction to represent variables scores obtained by each individual within a sample or group. All icon plots are thus methods for displaying the response patterns for individual members of a sample, as long as those individuals are not missing any scores on the variables to be displayed (note that this is the same limitation as for radar plots and parallel coordinate displays). To illustrate icon plots, without generating too many icons to focus on, Figs. 5.15, 5.16, 5.17 and 5.18 present four different icon plots for QCI inspectors classified, using a new variable called **BEST_WORST**, as either the

worst performers (= 1 where their accuracy scores were less than 70%) or the best performers (= 2 where their accuracy scores were 90% or greater).

The *Chernoff's faces plot* gets its name from the visual icon used to represent variable scores – a cartoon-type face. This icon tries to capitalise on our natural human ability to recognise and differentiate faces. Each feature of the face is controlled by the scores on a single variable. In SYSTAT, up to 20 facial features are controllable; the first five being curvature of mouth, angle of brow, width of nose, length of nose and length of mouth (SYSTAT Software Inc., 2009, p. 259). The theory behind Chernoff's faces is that similar patterns of variable scores will produce similar looking faces, thereby making similarities and differences between individuals more apparent.

The *profile plot* and *histogram plot* are actually two variants of the same type of icon plot. A profile plot represents individuals' scores for a set of variables using simplified line graphs, one per individual. The profile is scaled so that the vertical height of the peaks and valleys correspond to actual values for variables where the variables anchor the X-axis in a fashion similar to the parallel coordinate display. So, as you examine a profile from left to right across the X-axis of each graph, you are looking across the set of variables. A histogram plot represents the same information in the same way as for the profile plot but using histogram bars instead.

Figure 5.15, produced using SYSTAT, shows a Chernoff's faces plot for the best and worst performing inspectors using their ratings of job satisfaction, working conditions and the nine general attitude statements.

Each face is labelled with the inspector number it represents. The gaps indicate where an inspector had missing data on at least one of the variables, meaning a face could not be generated for them. The worst performers are drawn using red lines; the best using blue lines. The first variable is **jobsat** and this variable controls mouth curvature; the second variable is **workcond** and this controls angle of brow, and so on. It seems clear that there are differences in the faces between the best and worst performers with, for example, best performers tending to be more satisfied (smiling) and with higher ratings for working conditions (brow angle).

Beyond a broad visual impression, there is little in terms of precise inferences you can draw from a Chernoff's faces plot. It really provides a visual sketch, nothing more. The fact that there is no obvious link between facial features, variables and score levels means that the Chernoff's faces icon plot is difficult to interpret at the level of individual variables – a holistic impression of similarity and difference is what this type of plot facilitates.

Figure 5.16 produced using SYSTAT, shows a profile plot for the best and worst performing inspectors using their ratings of job satisfaction, working conditions and the nine attitude variables.

(continued)

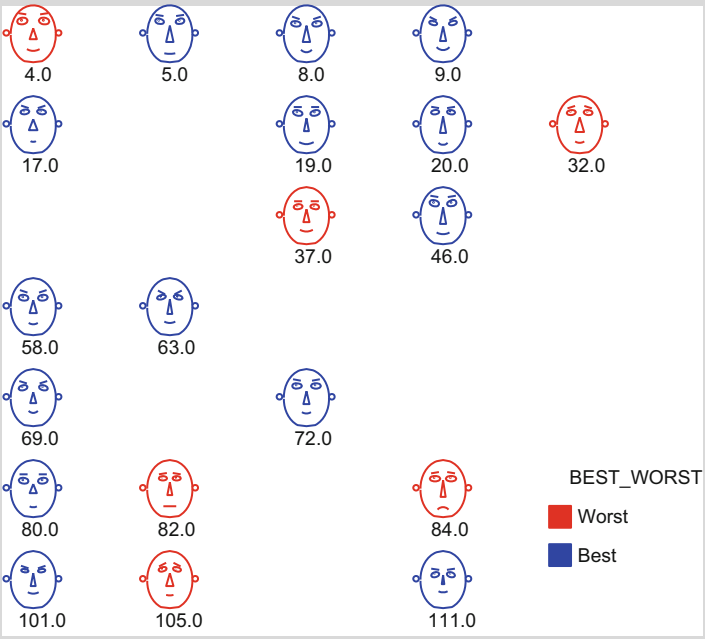

Fig. 5.15 Chernoff's faces icon plot comparing individual attitude ratings for best and worst performing inspectors

Like the Chernoff's faces plot (Fig. 5.15), as you read across the rows of the plot from left to right, each plot corresponds respectively to a inspector in the sample who was either in the worst performer (red) or best performer (blue) category. The first attitude variable is **jobsat** and anchors the left end of each line graph; the last variable is **polsatis** and anchors the right end of the line graph. The remaining variables are represented in order from left to right across the X-axis of each graph. Figure 5.16 shows that these inspectors are rather different in their attitude profiles, with best performers tending to show taller profiles on the first two variables, for example.

Figure 5.17 produced using SYSTAT, shows a histogram plot for the best and worst performing inspectors based on their ratings of job satisfaction, working conditions and the nine attitude variables. This plot tells the same story as the profile plot, only using histogram bars. Some people would prefer the histogram icon plot to the profile plot because each histogram bar corresponds to one variable, making the visual linking of a specific bar to a specific variable much easier than visually linking a specific position along the profile line to a specific variable.

(continued)

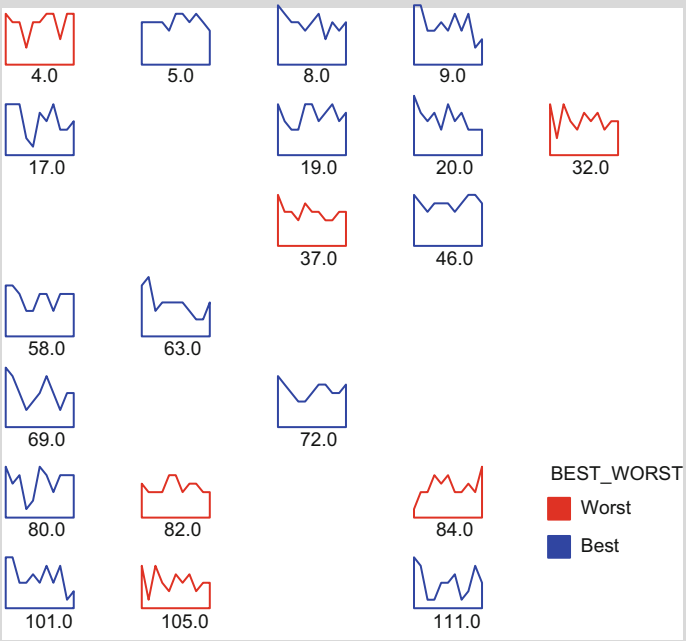

Fig. 5.16 Profile plot comparing individual attitude ratings for best and worst performing inspectors

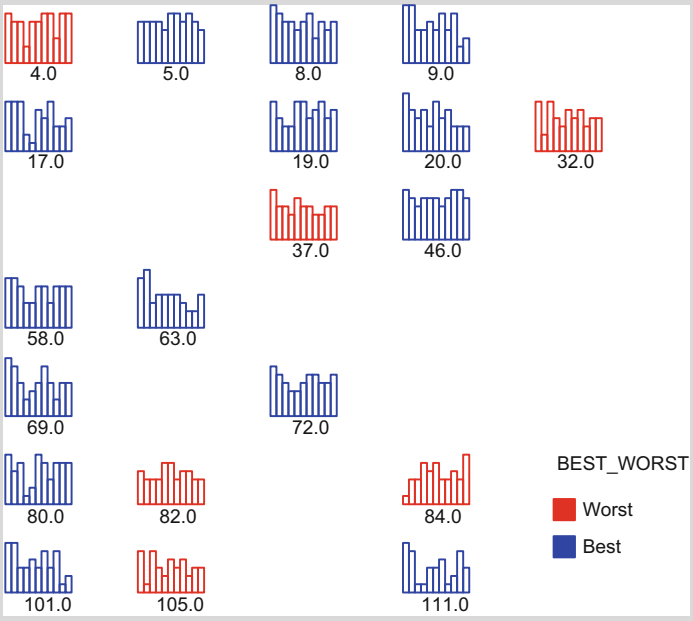

Fig. 5.17 Histogram plot comparing individual attitude ratings for best and worst performing inspectors

The *sunray plot* is actually a simplified adaptation of the radar plot (called a "star glyph") used to represent scores on a set of variables for each individual within a sample or group. Remember that a radar plot basically arranges the variables around a central point like a clock face; the first variable is represented at the 12 o'clock position and the remaining variables follow around the plot in a clockwise direction.

Unlike a radar plot, while the spokes (the actual 'star' of the glyph's name) of the plot are visible, no interpretive scale is evident. A variable's score is visually represented by its distance from the central point. Thus, the star glyphs in a sunray plot are designed, like Chernoff's faces, to provide a general visual impression, based on icon shape. A wide diameter well-rounded plot indicates an individual with high scores on all variables and a small diameter well-rounded plot vice-versa. Jagged plots represent individuals with highly variable scores across the variables. 'Stars' of similar size, shape and orientation represent similar individuals.

Figure 5.18, produced using STATGRAPHICS, shows a sunray plot for the best and worst performing inspectors. An interpretation glyph is also shown in the lower right corner of Fig. 5.18, where variables are aligned with the spokes of a star (e.g. **jobsat** is at the 12 o'clock position). This sunray plot could lead you to form the visual impression that the worst performing inspectors (group 1) have rather less rounded rating profiles than do the best performing inspectors (group 2) and that the **jobsat** and **workcond** spokes are generally lower for the worst performing inspectors.

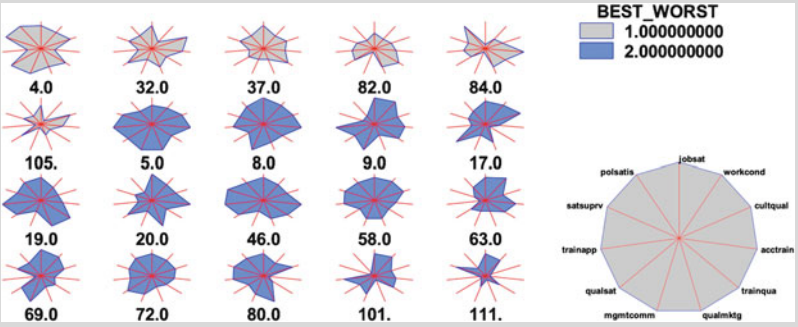

Fig. 5.18 Sunray plot comparing individual attitude ratings for best and worst performing inspectors

Comparatively speaking, the sunray plot makes identifying similar individuals a bit easier (perhaps even easier than Chernoff's faces) and, when ordered as STATGRAPHICS showed in Fig. 5.18, permits easier visual comparisons between groups of individuals, but at the expense of precise knowledge about variable scores. Remember, a holistic impression is the goal pursued using a sunray plot.

Advantages

Multivariate graphical methods provide summary techniques for visually presenting certain characteristics of a complex array of data on variables. Such visual representations are generally better at helping us to form holistic impressions of multivariate data rather than any sort of tabular representation or numerical index. They also allow us to compress many numerical measures into a finite representation that is generally easy to understand. Multivariate graphical displays can add interest to an otherwise dry statistical reporting of numerical data. They are designed to appeal to our pattern recognition skills, focusing our attention on features of the data such as shape, level, variability and orientation. Some multivariate graphs (e.g. radar plots, sunray plots and multiplots) are useful not only for representing score patterns for individuals but also providing summaries of score patterns across groups of individuals.

Disadvantages

Multivariate graphs tend to get very busy-looking and are hard to interpret if a great many variables or a large number of individuals need to be displayed (imagine any of the icon plots, for a sample of 200 questionnaire participants, displayed on a A4 page – each icon would be so small that its features could not be easily distinguished, thereby defeating the purpose of the display). In such cases, using numerical summary statistics (such as averages or correlations) in tabular form alone will provide a more economical and efficient summary. Also, some multivariate displays will work better for conveying certain types of information than others.

For example *Information about variable relationships may be better displayed using a scatterplot matrix. Information about individual similarities and difference on a set of variables may be better conveyed using a histogram or sunray plot. Multiplots may be better suited to displaying information about group differences across a set of variables. Information about the overall similarity of individual entities in a sample might best be displayed using Chernoff's faces.*

Because people differ greatly in their visual capacities and preferences, certain types of multivariate displays will work for some people and not others. Sometimes, people will not see what you see in the plots. Some plots, such as Chernoff's faces, may not strike a reader as a serious statistical procedure and this could adversely influence how convinced they will be by the story the plot conveys. None of the multivariate displays described here provide sufficiently precise information for solid inferences or interpretations; all are designed to simply facilitate the formation of holistic visual impressions. In fact, you may have noticed that some displays (scatterplot matrices and the icon plots, for example) provide no numerical scaling information that would help make precise interpretations. If precision in summary information is desired, the types of multivariate displays discussed here would not be the best strategic choices.

Where Is This Procedure Useful?

Virtually any research design which produces quantitative data/statistics for multiple variables provides opportunities for multivariate graphical data display which may help to clarify or illustrate important data characteristics or relationships. Thus, for survey research involving many identically-scaled attitudinal questions, a multivariate display may be just the device needed to communicate something about patterns in the data. Multivariate graphical displays are simply specialised communication tools designed to compress a lot of information into a meaningful and efficient format for interpretation—which tool to choose depends upon the message to be conveyed.

Generally speaking, visual representations of multivariate data could prove more useful in communicating to lay persons who are unfamiliar with statistics or who prefer visual as opposed to numerical information. However, these displays would probably require some interpretive discussion so that the reader clearly understands their intent.

Software Procedures

Application	Procedures
SPSS	*Graphs → Chart Builder*. . . and choose *Scatter/Dot* from the gallery; drag the *Scatterplot Matrix* chart type into the working area and customise the chart with desired variables, labels, etc. Only a few elements of each chart can be configured and altered.
NCSS	*Graphics → Scatter Plots → Scatter Plot Matrix.* Only a few elements of this plot are customisable in NCSS.
SYSTAT	*Graph → Scatterplot Matrix (SPLOM)* . . . (and you can select what type of plot you want to appear in the diagonal boxes) or *Graph → Line Chart* . . . (*Multiplots* can be selected by choosing a *Grouping* variable. e.g. **company**) or *Graph → Multivariate Display → Parallel Coordinate Display* . . . or *Icon Plot* . . . (for icon plots, you can choose from a range of icons including Chernoff's faces, histogram, star, sun or profile amongst others). A large number of elements of each type of plot are easily customisable, although it may take some trial and error to get exactly the look you want.
STATGRAPHICS	*Plot → Multivariate Visualization → Scatterplot Matrix...* or *Parallel Coordinates Plot...* or *Chernoff's Faces* or *Star Glyphs and Sunray Plots...* Several elements of each type of plot are easily customisable, although it may take some trial and error to get exactly the look you want.
R commander	*Graphs → Scatterplot Matrix...* You can select what type of plot you want to appear in the diagonal boxes, and you can control some other features of the plot. Other multivariate data displays are available via various **R** packages (e.g. the *lattice* or *car* package), but not through **R** commander.

Procedure 5.4: Assessing Central Tendency

Classification	Univariate; descriptive.
Purpose	To provide numerical summary measures that give an indication of the central, average or typical score in a distribution of scores for a variable.
Measurement level	*Mean* – variables should be measured at the interval or ratio-level. *Median* – variables should be measured at least at the ordinal-level. *Mode* – variables can be measured at any of the four levels.

The three most commonly reported measures of central tendency are the mean, median and mode. Each measure reflects a specific way of defining central tendency in a distribution of scores on a variable and each has its own advantages and disadvantages.

Mean

The *mean* is the most widely used measure of central tendency (also called the arithmetic average). Very simply, a mean is the sum of all the scores for a specific variable in a sample divided by the number of scores used in obtaining the sum. The resulting number reflects the average score for the sample of individuals on which the scores were obtained. If one were asked to predict the score that any single individual in the sample would obtain, the best prediction, in the absence of any other relevant information, would be the sample mean. Many parametric statistical methods (such as *Procedures 7.2, 7.4, 7.6* and *7.10*) deal with sample means in one way or another. For any sample of data, there is one and only one possible value for the mean in a specific distribution. For most purposes, the mean is the preferred measure of central tendency because it utilises all the available information in a sample.

In the context of the QCI database, Maree could quite reasonably ask what inspectors scored on the average in terms of mental ability (**mentabil**), inspection accuracy (**accuracy**), inspection speed (**speed**), overall job satisfaction (**jobsat**), and perceived quality of their working conditions (**workcond**). Table 5.3 shows the mean scores for the sample of 112 quality control inspectors on each of these variables. The statistics shown in Table 5.3 were computed using the SPSS *Frequencies*... procedure. Notice that the table indicates how many of the 112 inspectors had a valid score for each variable

(continued)

and how many were missing a score (e.g. 109 inspectors provided a valid rating for **jobsat;** 3 inspectors did not).

Each mean needs to be interpreted in terms of the original units of measurement for each variable. Thus, the inspectors in the sample showed an average mental ability score of 109.84 (higher than the general population mean of 100 for the test), an average inspection **accuracy** of 82.14%, and an average **speed** for making quality control decisions of 4.48 s. Furthermore, in terms of their work context, inspectors reported an average overall job satisfaction of 4.96 (on the 7-point scale, or a level of satisfaction nearly one full scale point above the Neutral point of 4—indicating a generally positive but not strong level of job satisfaction, and an average perceived quality of work conditions of 4.21 (on the 7-point scale which is just about at the level of Stressful but Tolerable.

Table 5.3 Measures of central tendency for specific QCI variables

Statistics

		mentabil	accuracy	speed	jobsat	workcond
N	Valid	111	111	111	109	106
	Missing	1	1	1	3	6
Mean		109.84	82.14	4.4801	4.96	4.21
Median		111.00	83.00	3.8900	5.00	4.00
Mode		111	82	3.14	6	4
Percentiles	25	104.00	77.00	2.1900	4.00	3.00
	50	111.00	83.00	3.8900	5.00	4.00
	75	116.00	89.00	5.7100	6.00	6.00

The mean is sensitive to the presence of extreme values, which can distort its value, giving a biased indication of central tendency. As we will see below, the median is an alternative statistic to use in such circumstances. However, it is also possible to compute what is called a *trimmed mean* where the mean is calculated after a certain percentage (say, 5% or 10%) of the lowest and highest scores in a distribution have been ignored (a process called 'trimming'; see, for example, the discussion in Field 2018, pp. 262–264). This yields a statistic less influenced by extreme scores. The drawbacks are that the decision as to what percentage to trim can be somewhat subjective and trimming necessarily sacrifices information (i.e. the extreme scores) in order to achieve a less biased measure. Some software packages, such as SPSS, SYSTAT or NCSS, can report a specific percentage trimmed mean, if that option is selected for descriptive statistics or exploratory data analysis (see *Procedure 5.6*) procedures. Comparing the original mean with a trimmed mean can provide an indication of the degree to which the original mean has been biased by extreme values.

Median

Very simply, the *median* is the centre or middle score of a set of scores. By 'centre' or 'middle' is meant that 50% of the data values are smaller than or equal to the median and 50% of the data values are larger when the entire distribution of scores is rank ordered from the lowest to highest value. Thus, we can say that the median is that score in the sample which occurs at the 50th percentile. [Note that a 'percentile' is attached to a specific score that a specific percentage of the sample scored at or below. Thus, a score at the 25th percentile means that 25% of the sample achieved this score or a lower score.] Table 5.3 shows the 25th, 50th and 75th percentile scores for each variable – *note how the 50th percentile score is exactly equal to the median in each case*.

The median is reported somewhat less frequently than the mean but does have some advantages over the mean in certain circumstances. One such circumstance is when the sample of data has a few extreme values in one direction (either very large or very small relative to all other scores). In this case, the mean would be influenced (biased) to a much greater degree than would the median since all of the data are used to calculate the mean (including the extreme scores) whereas only the single centre score is needed for the median. For this reason, many nonparametric statistical procedures (such as *Procedures 7.3, 7.5* and *7.9*) focus on the median as the comparison statistic rather than on the mean.

A discrepancy between the values for the mean and median of a variable provides some insight to the degree to which the mean is being influenced by the presence of extreme data values. In a distribution where there are no extreme values on either side of the distribution (or where extreme values balance each other out on either side of the distribution, as happens in a normal distribution – see *Fundamental Concept II*), the mean and the median will coincide at the same value and the mean will not be biased.

For highly skewed distributions, however, the value of the mean will be pulled toward the long tail of the distribution because that is where the extreme values lie. However, in such skewed distributions, the median will be insensitive (statisticians call this property 'robustness') to extreme values in the long tail. For this reason, the direction of the discrepancy between the mean and median can give a very rough indication of the direction of skew in a distribution ('mean larger than median' signals possible positive skewness; 'mean smaller than median' signals possible negative skewness). Like the mean, there is one and only one possible value for the median in a specific distribution.

In Fig. 5.19, the left graph shows the distribution of **speed** scores and the right-hand graph shows the distribution of **accuracy** scores. The **speed** distribution clearly shows the mean being pulled toward the right tail of the distribution whereas the **accuracy** distribution shows the mean being just slightly pulled toward the left tail. The effect on the mean is stronger in the **speed** distribution indicating a greater biasing effect due to some very long inspection decision times.

(continued)

If we refer to Table 5.3, we can see that the median score for each of the five variables has also been computed. Like the mean, the median must be interpreted in the original units of measurement for the variable. We can see that for **mentabil**, **accuracy**, and **workcond**, the value of the median is very close to the value of the mean, suggesting that these distributions are not strongly influenced by extreme data values in either the high or low direction. However, note that the median **speed** was 3.89 s compared to the mean of 4.48 s, suggesting that the distribution of **speed** scores is positively skewed (the mean is larger than the median—refer to Fig. 5.19). Conversely, the median **jobsat** score was 5.00 whereas the mean score was 4.96 suggesting very little substantive skewness in the distribution (mean and median are nearly equal).

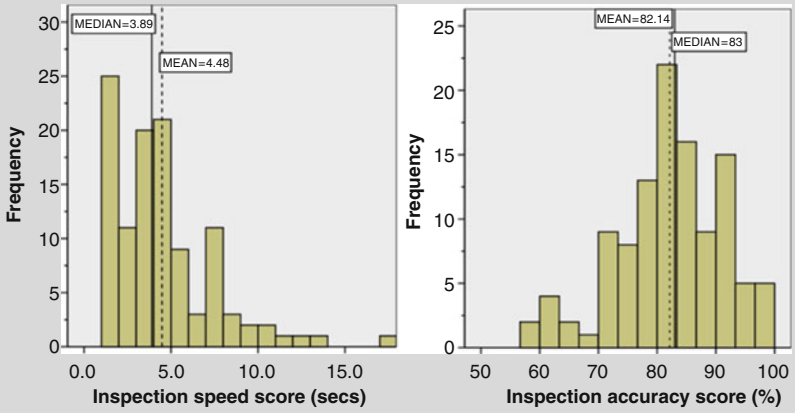

Fig. 5.19 Effects of skewness in a distribution on the values for the mean and median

Mode

The *mode* is the simplest measure of central tendency. It is defined as the most frequently occurring score in a distribution. Put another way, it is the score that more individuals in the sample obtain than any other score. An interesting problem associated with the mode is that there may be more than one in a specific distribution. In the case where multiple modes exist, the issue becomes which value do you report? The answer is that you must report all of them. In a 'normal' bell-shaped distribution, there is only one mode and it is indeed at the centre of the distribution, coinciding with both the mean and the median.

Table 5.3 also shows the mode for each of the five variables. For example, more inspectors achieved a **mentabil** score of 111 more often than any other score and inspectors reported a **jobsat** rating of 6 more often than any other rating. SPSS only ever reports one mode even if several are present, so one must be careful and look at a histogram plot for each variable to make a final determination of the mode(s) for that variable.

Advantages

All three measures of central tendency yield information about what is going on in the centre of a distribution of scores. The mean and median provide a single number which can summarise the central tendency in the entire distribution. The mode can yield one or multiple indices. With many measurements on individuals in a sample, it is advantageous to have single number indices which can describe the distributions in summary fashion. In a normal or near-normal distribution of sample data, the mean, the median, and the mode will all generally coincide at the one point. In this instance, all three statistics will provide approximately the same indication of central tendency. Note however that it is seldom the case that all three statistics would yield exactly the same number for any particular distribution. The mean is the most useful statistic, unless the data distribution is skewed by extreme scores, in which case the median should be reported.

Disadvantages

While measures of central tendency are useful descriptors of distributions, summarising data using a single numerical index necessarily reduces the amount of information available about the sample. Not only do we need to know what is going on in the centre of a distribution, we also need to know what is going on around the centre of the distribution. For this reason, most social and behavioural researchers report not only measures of central tendency, but also measures of variability (see *Procedure 5.5*). The mode is the least informative of the three statistics because of its potential for producing multiple values.

Where Is This Procedure Useful?

Measures of central tendency are useful in almost any type of experimental design, survey or interview study, and in any observational studies where quantitative data are available and must be summarised. The decision as to whether the mean or median should be reported depends upon the nature of the data which should ideally be ascertained by visual inspection of the data distribution. Some researchers opt to report both measures routinely. Computation of means is a prelude to many parametric statistical methods (see, for example, *Procedure 7.2, 7.4, 7.6, 7.8, 7.10, 7.11* and *7.16*); comparison of medians is associated with many nonparametric statistical methods (see, for example, *Procedure 7.3, 7.5, 7.9* and *7.12*).

Software Procedures

Application	Procedures
SPSS	*Analyze* → *Descriptive Statistics* → *Frequencies* . . . then press the '*Statistics*' button and choose mean, median and mode. To see trimmed means, you must use the *Analyze* → *Descriptive Statistics* → *Explore* . . . Exploratory Data Analysis procedure; see *Procedure 5.6*.
NCSS	*Analysis* → *Descriptive Statistics* → *Descriptive Statistics* then select the reports and plots that you want to see; make sure you indicate that you want to see the 'Means Section' of the Report. If you want to see trimmed means, tick the 'Trimmed Section' of the Report.
SYSTAT	*Analyze* → *Basic Statistics* . . . then select the mean, median and mode (as well as any other statistics you might wish to see). If you want to see trimmed means, tick the 'Trimmed mean' section of the dialog box and set the percentage to trim in the box labelled 'Two-sided'.
STATGRAPHICS	*Describe* → *Numerical Data* → *One-Variable Analysis...* or *Multiple-Variable Analysis...* then choose the variable(s) you want to describe and select Summary Statistics (you don't get any options for statistics to report – measures of central tendency and variability are automatically produced). STATGRAPHICS will not report modes and you will need to use *One-Variable Analysis...* and request 'Percentiles' in order to see the 50%ile score which will be the median; however, it won't be labelled as the median.
R Commander	*Statistics* → *Summaries* → *Numerical summaries...* then select the central tendency statistics you want to see. **R** Commander will not produce modes and to see the median, make sure that the 'Quantiles' box is ticked – the .5 quantile score (= 50%ile) score is the median; however, it won't be labelled as the median.

Procedure 5.5: Assessing Variability

Classification	Univariate; descriptive.
Purpose	To give an indication of the degree of spread in a sample of scores; that is, how different the scores tend to be from each other with respect to a specific measure of central tendency.
Measurement level	For the *variance* and *standard deviation*, interval or ratio-level measures are needed if these measures of variability are to have any interpretable meaning. At least an ordinal-level of measurement is required for the *range* and *interquartile range* to be meaningful.

There are a variety of measures of variability to choose from including the range, interquartile range, variance and standard deviation. Each measure reflects a specific way of defining variability in a distribution of scores on a variable and each has its own advantages and disadvantages. Most measures of variability are associated with a specific measure of central tendency so that researchers are now commonly expected to report both a measure of central tendency and its associated measure of variability whenever they display numerical descriptive statistics on continuous or ranked-ordered variables.

Range

This is the simplest measure of variability for a sample of data scores. The *range* is merely the largest score in the sample minus the smallest score in the sample. The range is the one measure of variability not explicitly associated with any measure of central tendency. It gives a very rough indication as to the extent of spread in the scores. However, since the range uses only two of the total available scores in the sample, the rest of the scores are ignored, which means that a lot of potentially useful information is being sacrificed. There are also problems if either the highest or lowest (or both) scores are atypical or too extreme in their value (as in highly skewed distributions). When this happens, the range gives a very inflated picture of the typical variability in the scores. Thus, the range tends not be a frequently reported measure of variability.

Table 5.4 shows a set of descriptive statistics, produced by the SPSS *Frequencies* procedure, for the **mentabil, accuracy, speed, jobsat** and **workcond** measures in the QCI database. In the table, you will find three rows labelled 'Range', 'Minimum' and 'Maximum'.

(continued)

Table 5.4 Measures of central tendency and variability for specific QCI variables

Statistics

		mentabil	accuracy	speed	jobsat	workcond
N	Valid	111	111	111	109	106
	Missing	1	1	1	3	6
Mean		109.84	82.14	4.4801	4.96	4.21
Median		111.00	83.00	3.8900	5.00	4.00
Mode		111	82	3.14	6	4
Std. Deviation		8.764	9.172	2.88751	1.644	1.717
Variance		76.810	84.118	8.338	2.702	2.947
Range		50	43	16.05	6	6
Minimum		85	57	1.05	1	1
Maximum		135	100	17.10	7	7
Percentiles	25	104.00	77.00	2.1900	4.00	3.00
	50	111.00	83.00	3.8900	5.00	4.00
	75	116.00	89.00	5.7100	6.00	6.00

Using the data from these three rows, we can draw the following descriptive picture. **Mentabil** scores spanned a range of 50 (from a minimum score of 85 to a maximum score of 135). **Speed** scores had a range of 16.05 s (from 1.05 s – the fastest quality decision to 17.10 – the slowest quality decision). **Accuracy** scores had a range of 43 (from 57% – the least accurate inspector to 100% – the most accurate inspector). Both work context measures (**jobsat** and **workcond**) exhibited a range of 6 – the largest possible range given the 1 to 7 scale of measurement for these two variables.

Interquartile Range

The *Interquartile Range* (*IQR*) is a measure of variability that is specifically designed to be used in conjunction with the median. The IQR also takes care of the extreme data problem which typically plagues the range measure. The IQR is defined as the range that is covered by the middle 50% of scores in a distribution once the scores have been ranked in order from lowest value to highest value. It is found by locating the value in the distribution at or below which 25% of the sample scored and subtracting this number from the value in the distribution at or below which 75% of the sample scored. The IQR can also be thought of as the range one would compute after the bottom 25% of scores and the top 25% of scores in the distribution have been 'chopped off' (or 'trimmed' as statisticians call it).

The IQR gives a much more stable picture of the variability of scores and, like the median, is relatively insensitive to the biasing effects of extreme data values. Some behavioural researchers prefer to divide the IQR in half which gives a measure called the *Semi-Interquartile Range* (*S-IQR*). The S-IQR can be interpreted as the distance one must travel away from the median, in either direction, to reach the value which separates the top (or bottom) 25% of scores in the distribution from the remaining 75%.

The IQR or S-IQR is typically not produced by descriptive statistics procedures by default in many computer software packages; however, it can usually be requested as an optional statistic to report or it can easily be computed by hand using percentile scores. Both the median and the IQR figure prominently in Exploratory Data Analysis, particularly in the production of *boxplots* (see *Procedure 5.6*).

Figure 5.20 illustrates the conceptual nature of the IQR and S-IQR compared to that of the range. Assume that 100% of data values are covered by the distribution curve in the figure. It is clear that these three measures would provide very different values for a measure of variability. Your choice would depend on your purpose. If you simply want to signal the overall span of scores between the minimum and maximum, the range is the measure of choice. But if you want to signal the variability around the median, the IQR or S-IQR would be the measure of choice.

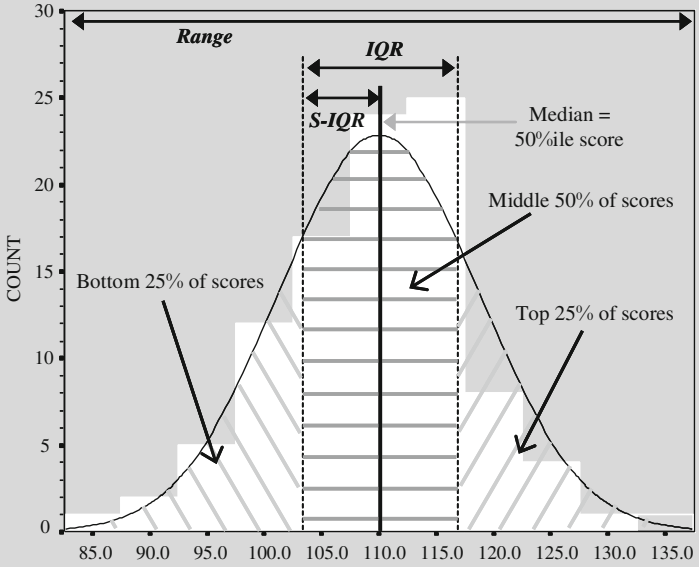

Fig. 5.20 How the range, IQR and S-IQR measures of variability conceptually differ

Note: Some behavioural researchers refer to the IQR as the *hinge-spread* (or *H-spread*) because of its use in the production of boxplots:

- the 25th percentile data value is referred to as the ‘lower hinge’;
- the 75th percentile data value is referred to as the ‘upper hinge’; and
- their difference gives the H-spread.

Midspread is another term you may see used as a synonym for interquartile range.

Referring back to Table 5.4, we can find statistics reported for the median and for the ‘quartiles’ (25th, 50th and 75th percentile scores) for each of the five variables of interest. The ‘quartile’ values are useful for finding the IQR or S-IQR because SPSS does not report these measures directly. The median clearly equals the 50th percentile data value in the table.

If we focus, for example, on the **speed** variable, we could find its IQR by subtracting the 25th percentile score of 2.19 s from the 75th percentile score of 5.71 s to give a value for the IQR of 3.52 s (the S-IQR would simply be 3.52 divided by 2 or 1.76 s). Thus, we could report that the median decision speed for inspectors was 3.89 s and that the middle 50% of inspectors showed scores spanning a range of 3.52 s. Alternatively, we could report that the median decision speed for inspectors was 3.89 s and that the middle 50% of inspectors showed scores which ranged 1.76 s either side of the median value.

Note: We could compare the ‘Minimum’ or ‘Maximum’ *scores* to the 25th percentile score and 75th percentile score respectively to get a feeling for whether the minimum or maximum might be considered extreme or uncharacteristic data values.

Variance

The *variance* uses information from every individual in the sample to assess the variability of scores relative to the sample mean. Variance assesses the average squared deviation of each score from the mean of the sample. *Deviation* refers to the difference between an observed score value and the mean of the sample—they are squared simply because adding them up in their naturally occurring unsquared form (where some differences are positive and others are negative) always gives a total of zero, which is useless for an index purporting to measure something.

If many scores are quite different from the mean, we would expect the variance to be large. If all the scores lie fairly close to the sample mean, we would expect a small variance. If all scores exactly equal the mean (i.e. all the scores in the sample have the same value), then we would expect the variance to be zero.

Figure 5.21 illustrates some possibilities regarding variance of a distribution of scores having a mean of 100. The very tall curve illustrates a distribution with small variance. The distribution of medium height illustrates a distribution with medium variance and the flattest distribution ia a distribution with large variance.

If we had a distribution with no variance, the curve would simply be a vertical line at a score of 100 (meaning that all scores were equal to the mean). You can see that as variance increases, the tails of the distribution extend further outward and the concentration of scores around the mean decreases. You may have noticed that variance and range (as well as the IQR) will be related, since the range focuses on the difference between the ends of the two tails in the distribution and larger variances extend the tails. So, a larger variance will generally be associated with a larger range and IQR compared to a smaller variance.

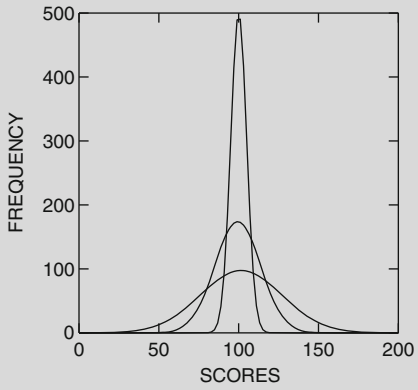

Fig. 5.21 The concept of variance

It is generally difficult to descriptively interpret the variance measure in a meaningful fashion since it involves squared deviations around the sample mean. [Note: If you look back at Table 5.4, you will see the variance listed for each of the variables (e.g. the variance of **accuracy** scores is 84.118), but the numbers themselves make little sense and do not relate to the original measurement scale for the variables (which, for the **accuracy** variable, went from 0% to 100% accuracy).] Instead, we use the variance as a steppingstone for obtaining a measure of variability that we can clearly interpret, namely the *standard deviation*. However, you should know that variance is an important concept in its own right simply because it provides the statistical foundation for many of the correlational procedures and statistical inference procedures described in *Chaps. 6, 7* and *8*.

For example *When considering either correlations or tests of statistical hypotheses, we frequently speak of one variable explaining or sharing variance with another (see* Procedure 6.4 *and* 7.7*). In doing so, we are invoking the concept of variance as set out here—what we are saying is that variability in the behaviour of scores on one*

particular variable may be associated with or predictive of variability in scores on another variable of interest (e.g. it could explain why those scores have a non-zero variance).

Standard Deviation

The *standard deviation* (often abbreviated as SD, sd or Std. Dev.) is the most commonly reported measure of variability because it has a meaningful interpretation and is used in conjunction with reports of sample means. Variance and standard deviation are closely related measures in that the standard deviation is found by taking the square root of the variance. The standard deviation, very simply, is a summary number that reflects the 'average distance of each score from the mean of the sample'. In many parametric statistical methods, both the sample mean and sample standard deviation are employed in some form. Thus, the standard deviation is a very important measure, not only for data description, but also for hypothesis testing and the establishment of relationships as well.

Referring again back to Table 5.4, we'll focus on the results for the **speed** variable for discussion purposes. Table 5.4 shows that the mean inspection **speed** for the QCI sample was 4.48 s. We can also see that the standard deviation (in the row labelled 'Std Deviation') for **speed** was 2.89 s.

This standard deviation has a straightforward interpretation: we would say that 'on the average, an inspector's quality inspection decision speed differed from the mean of the sample by about 2.89 s in either direction'. In a normal distribution of scores (see *Fundamental Concept II*), we would expect to see about 68% of all inspectors having decision speeds between 1.59 s (the mean minus one amount of the standard deviation) and 7.37 s (the mean plus one amount of the standard deviation).

We noted earlier that the range of the **speed** scores was 16.05 s. However, the fact that the maximum **speed** score was 17.1 s compared to the 75th percentile score of just 5.71 s seems to suggest that this maximum speed might be rather atypically large compared to the bulk of **speed** scores. This means that the range is likely to be giving us a false impression of the overall variability of the inspectors' decision speeds.

Furthermore, given that the mean **speed** score was higher than the median **speed** score, suggesting that **speed** scores were positively skewed (this was confirmed by the histogram for **speed** shown in Fig. 5.19 in *Procedure 5.4*), we might consider emphasising the median and its associated IQR or S-IQR rather than the mean and standard deviation. Of course, similar diagnostic and interpretive work could be done for each of the other four variables in Table 5.4.

Advantages

Measures of variability (particularly the standard deviation) provide a summary measure that gives an indication of how variable (spread out) a particular sample of scores is. When used in conjunction with a relevant measure of central tendency (particularly the mean), a reasonable yet economical description of a set of data emerges. When there are extreme data values or severe skewness is present in the data, the IQR (or S-IQR) becomes the preferred measure of variability to be reported in conjunction with the sample median (or 50th percentile value). These latter measures are much more resistant ('robust') to influence by data anomalies than are the mean and standard deviation.

Disadvantages

As mentioned above, the range is a very cursory index of variability, thus, it is not as useful as variance or standard deviation. Variance has little meaningful interpretation as a descriptive index; hence, standard deviation is most often reported. However, the standard deviation (or IQR) has little meaning if the sample mean (or median) is not reported along with it.

For example *Knowing that the standard deviation for* ***accuracy*** *is 9.17 tells you little unless you know the mean* ***accuracy*** *(82.14) that it is the standard deviation from.*

Like the sample mean, the standard deviation can be strongly biased by the presence of extreme data values or severe skewness in a distribution in which case the median and IQR (or S-IQR) become the preferred measures. The biasing effect will be most noticeable in samples which are small in size (say, less than 30 individuals) and far less noticeable in large samples (say, in excess of 200 or 300 individuals). [Note that, in a manner similar to a trimmed mean, it is possible to compute a **trimmed standard deviation** to reduce the biasing effect of extreme data values, see Field 2018, p. 263.]

It is important to realise that the resistance of the median and IQR (or S-IQR) to extreme values is only gained by deliberately sacrificing a good deal of the information available in the sample (nothing is obtained without a cost in statistics). What is sacrificed is information from all other members of the sample other than those members who scored at the median and 25th and 75th percentile points on a variable of interest; information from all members of the sample would automatically be incorporated in mean and standard deviation for that variable.

Where Is This Procedure Useful?

Any investigation where you might report on or read about measures of central tendency on certain variables should also report measures of variability. This is particularly true for data from experiments, quasi-experiments, observational studies and questionnaires. It is important to consider measures of central tendency and measures of variability to be inextricably linked—one should never report one without the other if an adequate descriptive summary of a variable is to be communicated.

Other descriptive measures, such as those for skewness and kurtosis[1] may also be of interest if a more complete description of any variable is desired. Most good statistical packages can be instructed to report these additional descriptive measures as well.

Of all the statistics you are likely to encounter in the business, behavioural and social science research literature, means and standard deviations will dominate as measures for describing data. Additionally, these statistics will usually be reported when any parametric tests of statistical hypotheses are presented as the mean and standard deviation provide an appropriate basis for summarising and evaluating group differences.

Software Procedures

Application	Procedures
SPSS	*Analyze → Descriptive Statistics → Frequencies* . . . then press the '*Statistics*' button and choose Std. Deviation, Variance, Range, Minimum and/or Maximum as appropriate. SPSS does not produce or have an option to produce either the IQR or S-IQR, however, if your request 'Quantiles' you will see the 25th and 75th %ile scores, which can then be used to quickly compute either variability measure. Remember to select appropriate central tendency measures as well.
NCSS	*Analysis → Descriptive Statistics → Descriptive Statistics* then select the reports and plots that you want to see; make sure you indicate that you want to see the Variance Section of the Report. Remember to select appropriate central tendency measures as well (by opting to see the Means Section of the Report).
SYSTAT	*Analyze → Basic Statistics* . . . then select SD, Variance, Range, Interquartile range, Minimum and/or Maximum as appropriate. Remember to select appropriate central tendency measures as well.

(continued)

[1]For more information, see *Chap. 1 – The language of statistics*.

Application	Procedures
STATGRAPHICS	*Describe* → *Numerical Data* → *One-Variable Analysis...* or *Multiple-Variable Analysis...* then choose the variable(s) you want to describe and select Summary Statistics (you don't get any options for statistics to report – measures of central tendency and variability are automatically produced). STATGRAPHICS does not produce either the IQR or S-IQR, however, if you use *One-Variable Analysis...* 'Percentiles' can be requested in order to see the 25th and 75th %ile scores, which can then be used to quickly compute either variability measure.
R Commander	*Statistics* → *Summaries* → *Numerical summaries...* then select either the Standard Deviation or Interquartile Range as appropriate. **R** Commander will not produce the range statistic or report minimum or maximum scores. Remember to select appropriate central tendency measures as well.

Fundamental Concept I: Basic Concepts in Probability

The Concept of Simple Probability

In *Procedures* 5.1 and 5.2, you encountered the idea of the frequency of occurrence of specific events such as particular scores within a sample distribution. Furthermore, it is a simple operation to convert the frequency of occurrence of a specific event into a number representing the relative frequency of that event. The relative frequency of an observed event is merely the number of times the event is observed divided by the total number of times one makes an observation. The resulting number ranges between 0 and 1 but we typically re-express this number as a percentage by multiplying it by 100%.

In the QCI database, Maree Lakota observed data from 112 quality control inspectors of which 58 were male and 51 were female (gender indications were missing for three inspectors). The statistics 58 and 51 are thus the frequencies of occurrence for two specific types of research participant, a male inspector or a female inspector.

If she divided each frequency by the total number of observations (i.e. 112), whe would obtain .52 for males and .46 for females (leaving .02 of observations with unknown gender). These statistics are relative frequencies which indicate the proportion of times that Maree obtained data from a male or female inspector. Multiplying each relative frequency by 100% would yield 52% and 46% which she could interpret as indicating that 52% of her sample was male and 46% was female (leaving 2% of the sample with unknown gender).

It does not take much of a leap in logic to move from the concept of 'relative frequency' to the concept of 'probability'. In our discussion above, we focused on relative frequency as indicating the proportion or percentage of times a specific category of participant was obtained in a sample. The emphasis here is on data from a sample.

Imagine now that Maree had infinite resources and research time and was able to obtain ever larger samples of quality control inspectors for her study. She could still compute the relative frequencies for obtaining data from males and females in her sample but as her sample size grew larger and larger, she would notice these relative frequencies converging toward some fixed values.

If, by some miracle, Maree could observe all of the quality control inspectors on the planet today, she would have measured the entire population and her computations of relative frequency for males and females would yield two precise numbers, each indicating the proportion of the population of inspectors that was male and the proportion that was female.

If Maree were then to list all of these inspectors and randomly choose one from the list, the chances that she would choose a male inspector would be equal to the proportion of the population of inspectors that was male and this logic extends to choosing a female inspector. The number used to quantify this notion of 'chances' is called a probability. Maree would therefore have established the probability of randomly observing a male or a female inspector in the population on any specific occasion.

Probability is expressed on a 0.0 (the observation or event will certainly *not* be seen) to 1.0 (the observation or event will certainly be seen) scale where values close to 0.0 indicate observations that are less certain to be seen and values close to 1.0 indicate observations that are more certain to be seen (a value of .5 indicates an even chance that an observation or event will or will not be seen – a state of maximum uncertainty). Statisticians often interpret a probability as the *likelihood* of observing an event or type of individual in the population.

In the QCI database, we noted that the relative frequency of observing males was .52 and for females was .46. If we take these relative frequencies as estimates of the proportions of each gender in the population of inspectors, then .52 and .46 represent the probability of observing a male or female inspector, respectively.

Statisticians would state this as "the probability of observing a male quality control inspector is .52" or in a more commonly used shorthand code, the likelihood of observing a male quality control inspector is $p = .52$ (p for probability). For some, probabilities make more sense if they are converted to percentages (by multiplying by 100%). Thus, $p = .52$ can also understood as a 52% chance of observing a male quality control inspector.

We have seen that relative frequency is a sample statistic that can be used to estimate the population probability. Our estimate will get more precise as we use larger and larger samples (technically, as the size of our samples more closely approximates the size of our population). In most behavioural research, we never have access to entire populations so we must always estimate our probabilities.

In some very special populations, having a known number of fixed possible outcomes, such as results of coin tosses or rolls of a die, we can analytically establish event probabilities without doing an infinite number of observations; all we must do is assume that we have a fair coin or die. Thus, with a fair coin, the probability of observing a H or a T on any single coin toss is ½ or .5 or 50%; the probability of observing a 6 on any single throw of a die is 1/6 or .16667 or 16.667%. With behavioural data, though, we can never measure all possible behavioural outcomes, which thereby forces researchers to depend on samples of observations in order to make estimates of population values.

The concept of probability is central to much of what is done in the statistical analysis of behavioural data. Whenever a behavioural scientist wishes to establish whether a particular relationship exists between variables or whether two groups, treated differently, actually show different behaviours, he/she is playing a probability game. Given a sample of observations, the behavioural scientist must decide whether what he/she has observed is providing sufficient information to conclude something about the population from which the sample was drawn.

This decision always has a non-zero probability of being in error simply because in samples that are much smaller than the population, there is always the chance or probability that we are observing something rare and atypical instead of something which is indicative of a consistent population trend. Thus, the concept of probability forms the cornerstone for statistical inference about which we will have more to say later (see *Fundamental Concept VI*). Probability also plays an important role in helping us to understand theoretical statistical distributions (e.g. the normal distribution) and what they can tell us about our observations. We will explore this idea further in *Fundamental Concept II*.

The Concept of Conditional Probability

It is important to understand that the concept of probability as described above focuses upon the likelihood or chances of observing a specific event or type of observation for a specific variable relative to a population or sample of observations. However, many important behavioural research issues may focus on the question of the probability of observing a specific event given that the researcher has knowledge that some other event has occurred or been observed (this latter event is usually measured by a second variable). Here, the focus is on the potential relationship or link between two variables or two events.

With respect to the QCI database, Maree could ask the quite reasonable question "what is the probability (estimated in the QCI sample by a relative frequency) of observing an inspector being female *given* that she knows that an inspector works for a Large Business Computer manufacturer.

To address this question, all she needs to know is:

- how many inspectors from Large Business Computer manufacturers are in the sample (**22**); and
- how many of those inspectors were female (**7**) (inspectors who were missing a score for either **company** or **gender** have been ignored here).

If she divides 7 by 22, she would obtain the probability that an inspector is female *given* that they work for a Large Business Computer manufacturer – that is, $p = .32$.

This type of question points to the important concept of *conditional probability* ('conditional' because we are asking "what is the probability of observing one event conditional upon our knowledge of some other event").

Continuing with the previous example, Maree would say that the conditional probability of observing a female inspector working for a Large Business Computer manufacturer is .32 or, equivalently, a 32% chance. Compare this conditional probability of $p = .32$ to the overall probability of observing a female inspector in the entire sample ($p = .46$ as shown above).

This means that there is evidence for a connection or relationship between gender and the type of company an inspector works for. That is, the chances are lower for observing a female inspector from a Large Business Computer manufacturer than they are for simply observing a female inspector at all.

Maree therefore has evidence suggesting that females may be relatively underrepresented in Large Business Computer manufacturing companies compared to the overall population. Knowing something about the company an inspector works for therefore can help us make a better prediction about their likely gender.

Suppose, however, that Maree's conditional probability had been exactly equal to $p = .46$. This would mean that there was exactly the same chance of observing a female inspector working for a Large Business Computer manufacturer as there was of observing a female inspector in the general population. Here, knowing something about the company an inspector works doesn't help Maree make any better prediction about their likely gender. This would mean that the two variables are statistically independent of each other.

A classic case of events that are statistically independent is two successive throws of a fair die: rolling a six on the first throw gives us no information for predicting how likely it will be that we would roll a six on the second throw. The conditional probability of observing a six on the second throw given that I have observed a six on the first throw is 0.16667 (= 1 divided by 6) which is the same as the simple probability of observing a six

on any specific throw. This statistical independence also means that if we wanted to know what the probability of throwing two sixes on two successive throws of a fair die, we would just multiply the probabilities for each independent event (i.e., throw) together; that is, .16667 × .16667 = .02789 (this is known as the *multiplication rule* of probability, see, for example, Smithson 2000, p. 114).

Finally, you should know that conditional probabilities are often asymmetric. This means that for many types of behavioural variables, reversing the conditional arrangement will change the story about the relationship. Bayesian statistics (see *Fundamental Concept IX*) relies heavily upon this asymmetric relationship between conditional probabilities.

Maree has already learned that the conditional probability that an inspector is female *given* that they worked for a Large Business Computer manufacturer is $p = .32$. She could easily turn the conditional relationship around and ask what is the conditional probability that an inspector works for a Large Business Computer manufacturer *given* that the inspector is female?

From the QCI database, she can find that 51 inspectors in her total sample were female and of those 51, 7 worked for a Large Business Computer manufacturer. If she divided 7 by 51, she would get $p = .14$ (did you notice that all that changed was the number she divided by?). Thus, there is only a 14% chance of observing an inspector working for a Large Business Computer manufacturer *given* that the inspector is female – a rather different probability from $p = .32$, which tells a different story.

As you will see in *Procedures 6.2* and *7.1*, conditional relationships between categorical variables are precisely what crosstabulation contingency tables are designed to reveal.

Procedure 5.6: Exploratory Data Analysis

Classification	Univariate; descriptive.
Purpose	To visually summarise data, displaying some key characteristics of their distribution, while maintaining as much of their original integrity as possible.
Measurement level	Exploratory Data Analysis (EDA) procedures are most usefully employed to explore data measured at the ordinal, interval or ratio-level.

There are a variety of visual display methods for EDA, including stem & leaf displays, boxplots and violin plots. Each method reflects a specific way of displaying features of a distribution of scores or measurements and, of course, each has its own advantages and disadvantages. In addition, EDA displays are surprisingly flexible and can combine features in various ways to enhance the story conveyed by the plot.

Stem & Leaf Displays

The *stem & leaf display* is a simple data summary technique which not only rank orders the data points in a sample but presents them visually so that the shape of the data distribution is reflected. Stem & leaf displays are formed from data scores by splitting each score into two parts: the first part of each score serving as the 'stem', the second part as the 'leaf' (e.g. for 2-digit data values, the 'stem' is the number in the tens position; the 'leaf' is the number in the ones position). Each stem is then listed vertically, in ascending order, followed horizontally by all the leaves in ascending order associated with it. The resulting display thus shows all of the scores in the sample, but reorganised so that a rough idea of the shape of the distribution emerges. As well, extreme scores can be easily identified in a stem & leaf display.

Consider the **accuracy** and **speed** scores for the 112 quality control inspectors in the QCI sample. Figure 5.22 (produced by the **R** Commander *Stem-and-leaf display* . . . procedure) shows the stem & leaf displays for inspection **accuracy** (left display) and **speed** (right display) data.

[The first six lines reflect information from **R** Commander about each display: lines 1 and 2 show the actual **R** command used to produce the plot (the variable name has been highlighted in bold); line 3 gives a warning indicating that inspectors with missing values (= NA in **R**) on the variable have been omitted from the display; line 4 shows how the stems and leaves have been defined; line 5 indicates what a leaf unit represents in value; and line 6 indicates the total number (n) of inspectors included in the display).] In Fig. 5.22, for the **accuracy** display on the left-hand side, the 'stems' have been split into 'half-stems'—one (which is starred) associated with the 'leaves' 0 through 4 and the other associated with the 'leaves' 5 through 9—a strategy that gives the display better balance and visual appeal.

```
> stem.leaf(statbook_2ed_R$accuracy, trim.outliers=FALSE,
na.rm=TRUE)
[1] "Warning: NA elements have been removed!!"
1 | 2: represents 12
 leaf unit: 1
           n: 111
    2      5. | 79
    7      6* | 02234
    9      6. | 57
   19      7* | 0012333334
   39      7. | 55555667778888889999
  (25)     8* | 00001112222222233333333444
   47      8. | 55555556666667778899999
   25      9* | 00000111222233344444
    5      9. | 89
    3     10* | 000
```

```
> stem.leaf(statbook_2ed_R$speed, trim.outliers=FALSE,
na.rm=TRUE)
[1] "Warning: NA elements have been removed!!"
1 | 2: represents 1.2
 leaf unit: 0.1
           n: 111
   25      1 | 01122333444555667777778899
   36      2 | 00123355788
  (20)     3 | 11111233444477788888
   55      4 | 01111122334445567779 9
   34      5 | 111346799
   25      6 | 001
   22      7 | 01334566889
   11      8 | 029
    8      9 | 15
    6     10 | 14
    4     11 | 9
    3     12 | 0
    2     13 | 0
          14 |
          15 |
          16 |
    1     17 | 1
```

Fig. 5.22 Stem & leaf displays produced by **R** Commander

(continued)

Fig. 5.23 Stem & leaf display, produced by SYSTAT, of the **accuracy** QCI variable

```
Stem and Leaf Plot of Variable: ACCURACY, N = 111

Minimum       :            57.000
Lower Hinge   :            77.000
Median        :            83.000
Upper Hinge   :            89.000
Maximum       :            100.000

                5   7

                5   9
                6   02234
                6   57
                7   0012333334
                7 H 55555667778888889999
                8 M 000011122222223333333444
                8 H 5555555666666777899999
                9   00000111222233344444
                9   89
               10   000

1 Cases with missing values excluded from plot.
```

Notice how the left stem & leaf display conveys a fairly clear (yet sideways) picture of the shape of the distribution of **accuracy** scores. It has a rather symmetrical bell-shape to it with only a slight suggestion of negative skewness (toward the extreme score at the top). The right stem & leaf display clearly depicts the highly positively skewed nature of the distribution of **speed** scores. Importantly, we could reconstruct the entire sample of scores for each variable using its display, which means that unlike most other graphical procedures, we didn't have to sacrifice any information to produce the visual summary.

Some programs, such as SYSTAT, embellish their stem & leaf displays by indicating in which stem or half-stem the 'median' (50th percentile), the 'upper hinge score' (75th percentile), and 'lower hinge score' (25th percentile) occur in the distribution (recall the discussion of *interquartile range* in *Procedure 5.5*). This is shown in Fig. 5.23, produced by SYSTAT, where M and H indicate the stem locations for the median and hinge points, respectively. This stem & leaf display labels a single extreme **accuracy** score as an 'outside value' and clearly shows that this actual score was 57.

Boxplots

Another important EDA technique is the *boxplot* or, as it is sometimes known, the *box-and-whisker plot*. This plot provides a symbolic representation that preserves less of the original nature of the data (compared to a stem & leaf display) but typically gives a better picture of the distributional characteristics. The basic boxplot, shown in Fig. 5.24, utilises information about the median (50th percentile score) and

the upper (75th percentile score) and lower (25th percentile score) hinge points in the construction of the 'box' portion of the graph (the 'median' defines the centre line in the box; the 'upper' and 'lower hinge values' define the end boundaries of the box—thus the box encompasses the middle 50% of data values).

Additionally, the boxplot utilises the IQR (recall *Procedure 5.5*) as a way of defining what are called 'fences' which are used to indicate score boundaries beyond which we would consider a score in a distribution to be an 'outlier' (or an extreme or unusual value). In SPSS, the inner fence is typically defined as 1.5 times the IQR in each direction and a 'far' outlier or extreme case is typically defined as 3 times the IQR in either direction (Field 2018, p. 193). The 'whiskers' in a boxplot extend out to the data values which are closest to the upper and lower inner fences (in most cases, the vast majority of data values will be contained within the fences). Outliers beyond these 'whiskers' are then individually listed. 'Near' outliers are those lying just beyond the inner fences and 'far' outliers lie well beyond the inner fences.

Figure 5.24 shows two simple boxplots (produced using SPSS), one for the **accuracy** QCI variable and one for the **speed** QCI variable. The **accuracy** plot shows a median value of about 83, roughly 50% of the data fall between about 77 and 89 and there is one outlier, inspector 83, in the lower 'tail' of the distribution. The **accuracy** boxplot illustrates data that are relatively symmetrically distributed without substantial skewness. Such data will tend to have their median in the middle of the box, whiskers of roughly equal length extending out from the box and few or no outliers.

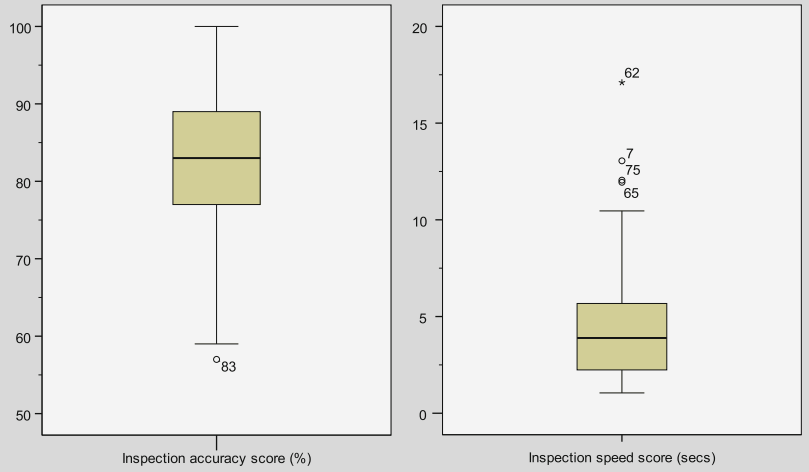

Fig. 5.24 Boxplots for the **accuracy** and **speed** QCI variables

The **speed** plot shows a median value of about 4 s, roughly 50% of the data fall between 2 s and 6 s and there are four outliers, inspectors 7, 62, 65 and 75 (although inspectors 65 and 75 fall at the same place and are rather difficult

(continued)

to read), all falling in the slow speed 'tail' of the distribution. Inspectors 65, 75 and 7 are shown as 'near' outliers (open circles) whereas inspector 62 is shown as a 'far' outlier (asterisk). The **speed** boxplot illustrates data which are asymmetrically distributed because of skewness in one direction. Such data may have their median offset from the middle of the box and/or whiskers of unequal length extending out from the box and outliers in the direction of the longer whisker. In the **speed** boxplot, the data are clearly positively skewed (the longer whisker and extreme values are in the slow speed 'tail').

Boxplots are very versatile representations in that side-by-side displays for sub-groups of data within a sample can permit easy visual comparisons of groups with respect to central tendency and variability. Boxplots can also be modified to incorporate information about error bands associated with the median producing what is called a 'notched boxplot'. This helps in the visual detection of meaningful subgroup differences, where boxplot 'notches' don't overlap.

Figure 5.25 (produced using NCSS), compares the distributions of **accuracy** and **speed** scores for QCI inspectors from the five types of companies, plotted side-by-side

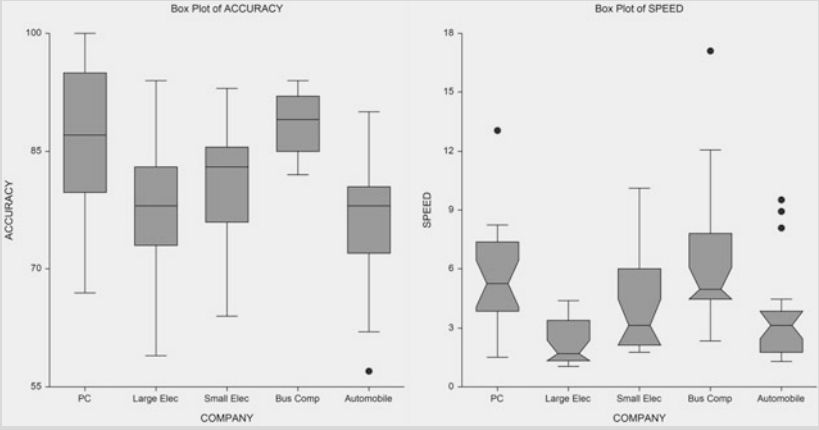

Fig. 5.25 Comparisons of the accuracy (regular boxplots) and speed (notched boxplots) QCI variables for different types of companies

.

Focus first on the left graph in Fig. 5.25 which plots the distribution of **accuracy** scores broken down by **company** using regular boxplots. This plot clearly shows the differing degree of skewness in each type of company (indicated by one or more outliers in one 'tail', whiskers which are not the same length and/or the median line being offset from the centre of a box), the differing variability of scores within each type of company (indicated by the

(continued)

overall length of each plot—box and whiskers), and the differing central tendency in each type of company (the median lines do not all fall at the same level of **accuracy** score). From the left graph in Fig. 5.25, we could conclude that: inspection **accuracy** scores are most variable in PC and Large Electrical Appliance manufacturing companies and least variable in the Large Business Computer manufacturing companies; Large Business Computer and PC manufacturing companies have the highest median level of inspection accuracy; and inspection accuracy scores tend to be negatively skewed (many inspectors toward higher levels, relatively fewer who are poorer in inspection performance) in the Automotive manufacturing companies. One inspector, working for an Automotive manufacturing company, shows extremely poor inspection accuracy performance.

The right display compares types of companies in terms of their inspection **speed** scores, using' notched' boxplots. The notches define upper and lower error limits around each median. Aside from the very obvious positive skewness for **speed** scores (with a number of slow speed outliers) in every type of company (least so for Large Electrical Appliance manufacturing companies), the story conveyed by this comparison is that inspectors from Large Electrical Appliance and Automotive manufacturing companies have substantially faster median decision speeds compared to inspectors from Large Business Computer and PC manufacturing companies (i.e. their 'notches' do not overlap, in terms of **speed** scores, on the display).

Boxplots can also add interpretive value to other graphical display methods through the creation of hybrid displays. Such displays might combine a standard histogram with a boxplot along the X-axis to provide an enhanced picture of the data distribution as illustrated for the **mentabil** variable in Fig. 5.26 (produced using NCSS). This hybrid

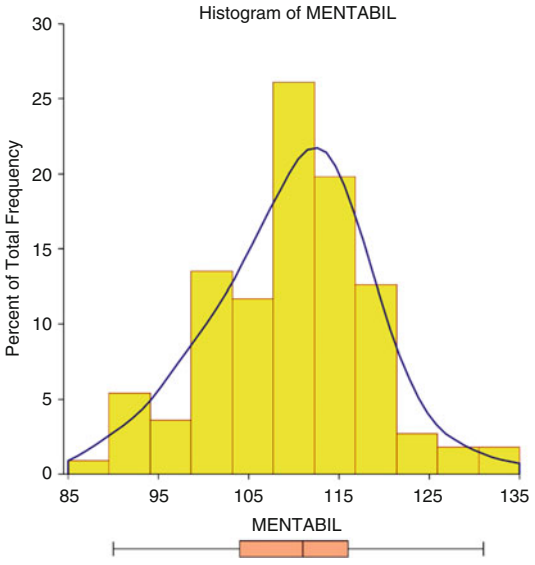

Fig. 5.26 A hybrid histogram-density-boxplot of the **mentabil** QCI variable

plot also employs a data 'smoothing' method called a **density trace** to outline an approximate overall shape for the data distribution. Any one graphical method would tell some of the story, but combined in the hybrid display, the story of a relatively symmetrical set of **mentabil** scores becomes quite visually compelling.

Violin Plots

Violin plots are a more recent and interesting EDA innovation, implemented in the NCSS software package (Hintze 2012). The violin plot gets its name from the rough shape that the plots tend to take on. Violin plots are another type of hybrid plot, this time combining density traces (mirror-imaged right and left so that the plots have a sense of symmetry and visual balance) with boxplot-type information (median, IQR and upper and lower inner 'fences', but not outliers). The goal of the violin plot is to provide a quick visual impression of the shape, central tendency and variability of a distribution (the length of the violin conveys a sense of the overall variability whereas the width of the violin conveys a sense of the frequency of scores occurring in a specific region).

Figure 5.27 (produced using NCSS), compares the distributions of **speed** scores for QCI inspectors across the five types of companies, plotted side-by-side. The violin plot conveys a similar story to the boxplot comparison for **speed** in the right graph of Fig. 5.25. However, notice that with the violin plot, unlike with a boxplot, you also get a sense of distributions that have 'clumps' of scores in specific areas. Some violin plots, like that for Automobile manufacturing companies in Fig. 5.27, have a shape suggesting a multi-modal distribution (recall *Procedure 5.4* and the discussion of the fact that a distribution may have multiple modes). The violin plot in Fig. 5.27 has also been produced to show where the median (solid line) and mean (dashed line) would fall within each violin. This facilitates two interpretations: (1) a relative comparison of central tendency across the five companies and (2) relative degree of skewness in the distribution for each company (indicated by the separation of the two lines within a violin; skewness is particularly bad for the Large Business Computer manufacturing companies).

(continued)

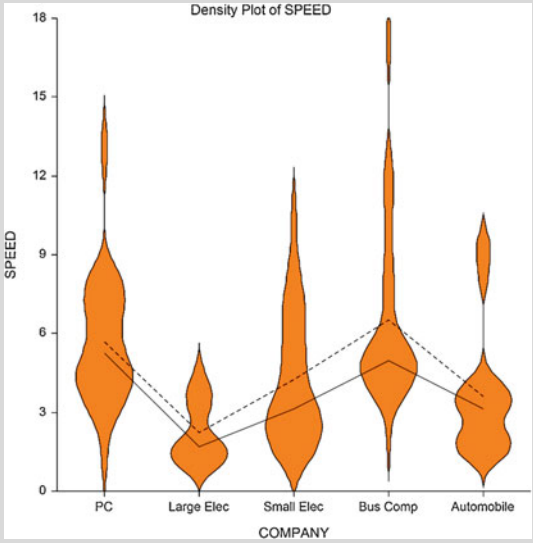

Fig. 5.27 Violin plot comparisons of the speed QCI variable for different types of companies

Advantages

EDA methods (of which we have illustrated only a small subset; we have not reviewed dot density diagrams, for example) provide summary techniques for visually displaying certain characteristics of a set of data. The advantage of the EDA methods over more traditional graphing techniques such as those described in *Procedure 5.2* is that as much of the original integrity of the data is maintained as possible while maximising the amount of summary information available about distributional characteristics.

Stem & leaf displays maintain the data in as close to their original form as possible whereas boxplots and violin plots provide more symbolic and flexible representations. EDA methods are best thought of as communication devices designed to facilitate quick visual impressions and they can add interest to any statistical story being conveyed about a sample of data. NCSS, SYSTAT, STATGRAPHICS and **R** Commander generally offer more options and flexibility in the generation of EDA displays than SPSS.

Disadvantages

EDA methods tend to get cumbersome if a great many variables or groups need to be summarised. In such cases, using numerical summary statistics (such as means and standard deviations) will provide a more economical and efficient summary.

Boxplots or violin plots are generally more space efficient summary techniques than stem & leaf displays.

Often, EDA techniques are used as data screening devices, which are typically not reported in actual write-ups of research (we will discuss data screening in more detail in *Procedure 8.2*). This is a perfectly legitimate use for the methods although there is an argument for researchers to put these techniques to greater use in published literature.

Software packages may use different rules for constructing EDA plots which means that you might get rather different looking plots and different information from different programs (you saw some evidence of this in Figs. 5.22 and 5.23). It is important to understand what the programs are using as decision rules for locating fences and outliers so that you are clear on how best to interpret the resulting plot—such information is generally contained in the user's guides or manuals for NCSS (Hintze 2012), SYSTAT (SYSTAT Inc. 2009a, b), STATGRAPHICS (StatPoint Technologies Inc. 2010) and SPSS (Norušis 2012).

Where Is This Procedure Useful?

Virtually any research design which produces numerical measures (even to the extent of just counting the number of occurrences of several events) provides opportunities for employing EDA displays which may help to clarify data characteristics or relationships. One extremely important use of EDA methods is as data screening devices for detecting outliers and other data anomalies, such as non-normality and skewness, before proceeding to parametric statistical analyses. In some cases, EDA methods can help the researcher to decide whether parametric or nonparametric statistical tests would be best to apply to his or her data because critical data characteristics such as distributional shape and spread are directly reflected.

Software Procedures

Application	Procedures
SPSS	*Analyze* → *Descriptive Statistics* → *Explore* . . . produces stem-and-leaf displays and boxplots by default; variables may be explored on a whole-of-sample basis or broken down by the categories of a specific variable (called a 'factor' in the procedure). Cases can also be labelled with a variable (like **inspector** in the QCI database), so that outlier points in the boxplot are identifiable. *Graphs* → *Chart Builder*. . . can also be used to custom build different types of boxplots.
NCSS	*Analysis* → *Descriptive Statistics* → *Descriptive Statistics* produces a stem-and-leaf display by default. *Graphics* → *Box Plots* → *Box Plots* can be used to produce box plots with different features (such as 'notches' and connecting lines). *Graphics* → *Density Plots* → *Density Plots* can be configured to produce violin plots (by selecting the plot shape as 'density with reflection').

(continued)

Application	Procedures
SYSTAT	*Analyze* → *Stem-and-Leaf*. . . can be used to produce stem-and-leaf displays for variables; however, you cannot really control any features of these displays. *Graph* → *Box Plot*. . . can be used to produce boxplots of many types, with a number of features being controllable.
STATGRAPHICS	*Describe* → *Numerical Data* → *One-Variable Analysis*... allows you to do a complete exploration of a single variable, including stem-and-leaf display (you need to select this option) and boxplot (produced by default). Some features of the boxplot can be controlled, but not features of the stem-and-leaf diagram. *Plot* → *Exploratory Plots* → *Box-and-Whisker Plots*... and select either *One Sample*... or *Multiple Samples*... which can produce not only descriptive statistics but also boxplots with some controllable features.
R Commander	*Graphs* → *Stem-and-leaf display*... or *Boxplots*... the dialog box for each procedure offers some features of the display or plot that can be controlled; whole-of-sample boxplots or boxplots by groups are possible.

Procedure 5.7: Standard (z) Scores

Classification Univariate; descriptive.

Purpose To transform raw scores from a sample of data to a standardised form which permits comparisons with other scores within the same sample or with scores from other samples of data.

Measurement level Generally, standard scores are computed from interval or ratio-level data.

In certain practical situations in behavioural research, it may be desirable to know where a specific individual's score lies relative to all other scores in a distribution. A convenient measure is to observe how many standard deviations (see *Procedure 5.5*) above or below the sample mean a specific score lies. This measure is called a *standard score* or z-score. Very simply, any raw score can be converted to a z-score by subtracting the sample mean from the raw score and dividing that result by the sample's standard deviation. z-scores can be positive or negative and their sign simply indicates whether the score lies above (+) or below (−) the mean in value. A z-score has a very simple interpretation: it measures the number of standard deviations above or below the sample mean a specific raw score lies.

In the QCI database, we have a sample mean for **speed** scores of 4.48 s, a standard deviation for **speed** scores of 2.89 s (recall Table 5.4 in *Procedure 5.5*). If we are interested in the z-score for Inspector 65's raw **speed** score of 11.94 s, we would obtain a z-score of +2.58 using the method described above (subtract 4.48 from 11.94 and divide the result by 2.89). The interpretation of this number is that a raw decision **speed** score of 11.94 s lies about 2.9 standard deviations above the mean decision **speed** for the sample.

z-scores have some interesting properties. First, if one converts (statisticians would say 'transforms') every available raw score in a sample to z-scores, the mean of these z-scores will always be zero and the standard deviation of these z-scores will always be 1.0. These two facts about z-scores (mean $= 0$; standard deviation $= 1$) will be true no matter what sample you are dealing with and no matter what the original units of measurement are (e.g. seconds, percentages, number of widgets assembled, amount of preference for a product, attitude rating, amount of money spent). This is because transforming raw scores to z-scores automatically changes the measurement units from whatever they originally were to a new system of measurements expressed in standard deviation units.

Suppose Maree was interested in the performance statistics for the top 25% most accurate quality control inspectors in the sample. Given a sample size of 112, this would mean finding the top 28 inspectors in terms of their **accuracy** scores. Since Maree is interested in performance statistics, **speed** scores would also be of interest. Table 5.5 (generated using the SPSS *Descriptives* ... procedure, listed using the *Case Summaries* ... procedure and formatted for presentation using Excel) shows **accuracy** and **speed** scores for the top 28 inspectors in descending order of **accuracy** scores. The z-score transformation for each of these scores is also shown (last two columns) as are the type of company, education level and gender for each inspector.

There are three inspectors (8, 9 and 14) who scored maximum accuracy of 100%. Such accuracy converts to a z-score of +1.95. Thus 100% accuracy is 1.95 standard deviations above the sample's mean accuracy level. Interestingly, all three inspectors worked for PC manufacturers and all three had only high school-level education. The least accurate inspector in the top 25% had a z-score for **accuracy** that was .75 standard deviations above the sample mean.

Interestingly, the top three inspectors in terms of accuracy had decision speeds that fell below the sample's mean speed; inspector 8 was the fastest inspector of the three with a speed just over 1 standard deviation ($z = -1.03$) below the sample mean. The slowest inspector in the top 25% was inspector 75 (case #28 in the list) with a **speed** z-score of +2.62; i.e., he was over two and a half standard deviations slower in making inspection decisions relative to the sample's mean speed.

(continued)

Table 5.5 Listing of the 28 (top 25%) most accurate QCI inspectors' **accuracy** and **speed** scores as well as standard (z) score transformations for each score

Case number	Inspector	company	educlev	gender	accuracy	speed	Zaccuracy	Zspeed
1	8	PC Manufacturer	High School Only	Male	100	1.52	1.95	−1.03
2	9	PC Manufacturer	High School Only	Female	100	3.32	1.95	−0.40
3	14	PC Manufacturer	High School Only	Male	100	3.83	1.95	−0.23
4	17	PC Manufacturer	High School Only	Female	99	7.07	1.84	0.90
5	101	PC Manufacturer	High School Only		98	3.11	1.73	−0.47
6	19	PC Manufacturer	Tertiary Qualified	Female	94	3.84	1.29	−0.22
7	34	Large Electrical Appliance Manufacturer	Tertiary Qualified	Male	94	1.90	1.29	−0.89
8	63	Large Business Computer Manufacturer	High School Only	Male	94	11.94	1.29	2.58
9	67	Large Business Computer Manufacturer	High School Only	Male	94	2.34	1.29	−0.74
10	80	Large Business Computer Manufacturer	High School Only	Female	94	4.68	1.29	0.07
11	5	PC Manufacturer	Tertiary Qualified	Male	93	4.18	1.18	−0.10
12	18	PC Manufacturer	Tertiary Qualified	Male	93	7.32	1.18	0.98
13	46	Small Electrical Appliance Manufacturer	Tertiary Qualified	Female	93	2.01	1.18	−0.86
14	64	Large Business Computer Manufacturer	High School Only	Female	92	5.18	1.08	0.24
15	77	Large Business Computer Manufacturer	Tertiary Qualified	Female	92	6.11	1.08	0.56
16	79	Large Business Computer Manufacturer	High School Only	Male	92	4.38	1.08	−0.03
17	106	Large Electrical Appliance Manufacturer	Tertiary Qualified	Male	92	1.70	1.08	−0.96
18	58	Small Electrical Appliance Manufacturer	High School Only	Male	91	4.12	0.97	−0.12
19	63	Large Business Computer Manufacturer	High School Only	Male	91	4.73	0.97	0.09

(continued)

(continued)

Table 5.5 (continued)

Case number	Inspector	company	educlev	gender	accuracy	speed	Zaccuracy	Zspeed
20	72	Large Business Computer Manufacturer	Tertiary Qualified	Male	91	4.72	0.97	0.08
21	20	PC Manufacturer	High School Only	Male	90	4.53	0.86	0.02
22	69	Large Business Computer Manufacturer	High School Only	Male	90	4.94	0.86	0.16
23	71	Large Business Computer Manufacturer	High School Only	Female	90	10.46	0.86	2.07
24	85	Automobile Manufacturer	Tertiary Qualified	Female	90	3.14	0.86	−0.46
25	111	Large Business Computer Manufacturer	High School Only	Male	90	4.11	0.86	−0.13
26	6	PC Manufacturer	High School Only	Male	89	5.46	0.75	0.34
27	61	Large Business Computer Manufacturer	Tertiary Qualified	Male	89	5.71	0.75	0.43
28	75	Large Business Computer Manufacturer	High School Only	Male	89	12.05	0.75	2.62

(continued)

The fact that z-scores always have a common measurement scale having a mean of 0 and a standard deviation of 1.0 leads to an interesting application of standard scores. Suppose we focus on inspector number 65 (case #8 in the list) in Table 5.5. It might be of interest to compare this inspector's quality control performance in terms of both his decision accuracy and decision speed. Such a comparison is impossible using raw scores since the inspector's **accuracy** score and **speed** scores are different measures which have differing means and standard deviations expressed in fundamentally different units of measurement (percentages and seconds). However, if we are willing to assume that the score distributions for both variables are approximately the same shape and that both **accuracy** and **speed** are measured with about the same level of reliability or consistency (see *Procedure 8.1*), we can compare the inspector's two scores by first converting them to z-scores within their own respective distributions as shown in Table 5.5.

Inspector 65 looks rather anomalous in that he demonstrated a relatively high level of **accuracy** (raw score = 94%; $z = +1.29$) but took a very long time to make those accurate decisions (raw score = 11.94 s; $z = +2.58$). Contrast this with inspector 106 (case #17 in the list) who demonstrated a similar level of **accuracy** (raw score = 92%; $z = +1.08$) but took a much shorter time to make those accurate decisions (raw score = 1.70 s; $z = -.96$). In terms of evaluating performance, from a company perspective, we might conclude that inspector 106 is performing at an overall higher level than inspector 65 because he can achieve a very high level of accuracy but much more quickly; accurate and fast is more cost effective and efficient than accurate and slow.

Note: We should be cautious here since we know from our previous explorations of the **speed** variable in *Procedure 5.6*, that **accuracy** scores look fairly symmetrical and **speed** scores are positively skewed, so assuming that the two variables have the same distribution shape, so that z-score comparisons are permitted, would be problematic.

You might have noticed that as you scanned down the two columns of z-scores in Table 5.5, there was a suggestion of a pattern between the signs attached to the respective z-scores for each person. There seems to be a very slight preponderance of pairs of z-scores where the signs are reversed (12 out of 22 pairs). This observation provides some very preliminary evidence to suggest that there may be a relationship between inspection accuracy and decision speed, namely that a more accurate decision tends to be associated with a faster decision speed. Of course, this pattern would be better verified using the entire sample rather than the top 25% of inspectors. However, you may find it interesting to learn that it is precisely this sort of suggestive evidence (about agreement or disagreement between z-score signs for pairs of variable scores throughout a sample) that is captured and summarised by a single statistical indicator called a 'correlation coefficient' (see *Fundamental Concept III* and *Procedure 6.1*).

z-scores are not the only type of standard score that is commonly used. Three other types of standard scores are: *stanines* (standard nines), *IQ scores* and *T-scores* (not to be confused with the *t*-test described in *Procedure 7.2*). These other types of scores have the advantage of producing only positive integer scores rather than positive and negative decimal scores. This makes interpretation somewhat easier for certain applications. However, you should know that almost all other types of standard scores come from a specific transformation of *z*-scores. This is because once you have converted raw scores into *z*-scores, they can then be quite readily transformed into any other system of measurement by simply multiplying a person's *z*-score by the new desired standard deviation for the measure and adding to that product the new desired mean for the measure.

For example *T-scores are simply z-scores transformed to have a mean of 50.0 and a standard deviation of 10.0; IQ scores are simply z-scores transformed to have a mean of 100 and a standard deviation of 15 (or 16 in some systems). For more information, see* Fundamental Concept II.

Advantages

Standard scores are useful for representing the position of each raw score within a sample distribution relative to the mean of that distribution. The unit of measurement becomes the number of standard deviations a specific score is away from the sample mean. As such, *z*-scores can permit cautious comparisons across samples or across different variables having vastly differing means and standard deviations within the constraints of the comparison samples having similarly shaped distributions and roughly equivalent levels of measurement reliability. *z*-scores also form the basis for establishing the degree of correlation between two variables. Transforming raw scores into *z*-scores does not change the shape of a distribution or rank ordering of individuals within that distribution. For this reason, a *z*-score is referred to as a *linear transformation* of a raw score. Interestingly, *z*-scores provide an important foundational element for more complex analytical procedures such as factor analysis (*Procedure 6.5*), cluster analysis (*Procedure 6.6*) and multiple regression analysis (see, for example, *Procedure 6.4 and 7.13*).

Disadvantages

While standard scores are useful indices, they are subject to restrictions if used to compare scores across samples or across different variables. The samples must have similar distribution shapes for the comparisons to be meaningful and the measures must have similar levels of reliability in each sample. The groups used to generate the *z*-scores should also be similar in composition (with respect to age, gender

distribution, and so on). Because z-scores are not an intuitively meaningful way of presenting scores to lay-persons, many other types of standard score schemes have been devised to improve interpretability. However, most of these schemes produce scores that run a greater risk of facilitating lay-person misinterpretations simply because their connection with z-scores is hidden or because the resulting numbers 'look' like a more familiar type of score which people do intuitively understand.

For example *It is extremely rare for a T-score to exceed 100 or go below 0 because this would mean that the raw score was in excess of 5 standard deviations away from the sample mean. This unfortunately means that T-scores are often misinterpreted as percentages because they typically range between 0 and 100 and therefore 'look' like percentages. However, T-scores are definitely not percentages.*

Finally, a common misunderstanding of z-scores is that transforming raw scores into z-scores makes them follow a normal distribution (see *Fundamental Concept II*). This is not the case. The distribution of z-scores will have exactly the same shape as that for the raw scores; if the raw scores are positively skewed, then the corresponding z-scores will also be positively skewed.

Where Is This Procedure Useful?

z-scores are particularly useful in evaluative studies where relative performance indices are of interest. Whenever you compute a correlation coefficient (*Procedure 6.1*), you are implicitly transforming the two variables involved into z-scores (which equates the variables in terms of mean and standard deviation), so that only the patterning in the relationship between the variables is represented. z-scores are also useful as a preliminary step to more advanced parametric statistical methods when variables differing in scale, range and/or measurement units must be equated for means and standard deviations prior to analysis.

Software Procedures

Application	Procedures
SPSS	*Analyze* → *Descriptive Statistics* → *Descriptives...* and tick the box labelled 'Save standardized values as variables'. z-scores are saved as new variables (labelled as Z followed by the original variable name as shown in Table 5.5) which can then be listed or analysed further.
NCSS	*Data* → *Transformations* → *Transformation* and select a new variable to hold the z-scores, then select the 'STANDARDIZE' transformation from the list of available functions. z-scores are saved as new variables which can then be listed or analysed further.

(continued)

Application	Procedures
SYSTAT	*Data* → *Standardize* . . . where *z*-scores are saved as new variables which can then be listed or analysed further.
STATGRAPHICS	Open the *Databook* window, and select an empty column in the database, then *Edit* → *Generate Data...* and choose the 'STANDARDIZE' transformation, choose the variable you want to transform and give the new variable a name.
R Commander	*Data* → *Manage variables in active data set* → *Standardize variables...* and select the variables you want to standardize; **R** Commander automatically saves the transformed variable to the data base, appending Z. to the front of each variable's name.

Fundamental Concept II: The Normal Distribution

Arguably the most fundamental distribution used in the statistical analysis of quantitative data in the behavioural and social sciences is the *normal distribution* (also known as the *Gaussian* or *bell-shaped distribution*). Many behavioural phenomena, if measured on a large enough sample of people, tend to produce 'normally distributed' variable scores. This includes most measures of ability, performance and productivity, personality characteristics and attitudes. The normal distribution is important because it is the one form of distribution that you must assume describes the scores of a variable in the population when parametric tests of statistical inference are undertaken. The standard normal distribution is defined as having a population mean of 0.0 and a population standard deviation of 1.0. The normal distribution is also important as a means of interpreting various types of scoring systems.

Figure 5.28 displays the standard normal distribution (mean = 0; standard deviation = 1.0) and shows that there is a clear link between *z*-scores and the normal distribution. Statisticians have analytically calculated the probability (also expressed as percentages or percentiles) that observations will fall above or below any specific *z*-score in the theoretical standard normal distribution. Thus, a *z*-score of +1.0 in the standard normal distribution will have 84.13% (equals a probability of .8413) of observations in the population falling at or below one standard deviation above the mean and 15.87% falling above that point. A *z*-score of −2.0 will have 2.28% of observations falling at that point or below and 97.72% of observations falling above that point. It is clear then that, in a standard normal distribution, *z*-scores have a direct relationship with percentiles.

(continued)

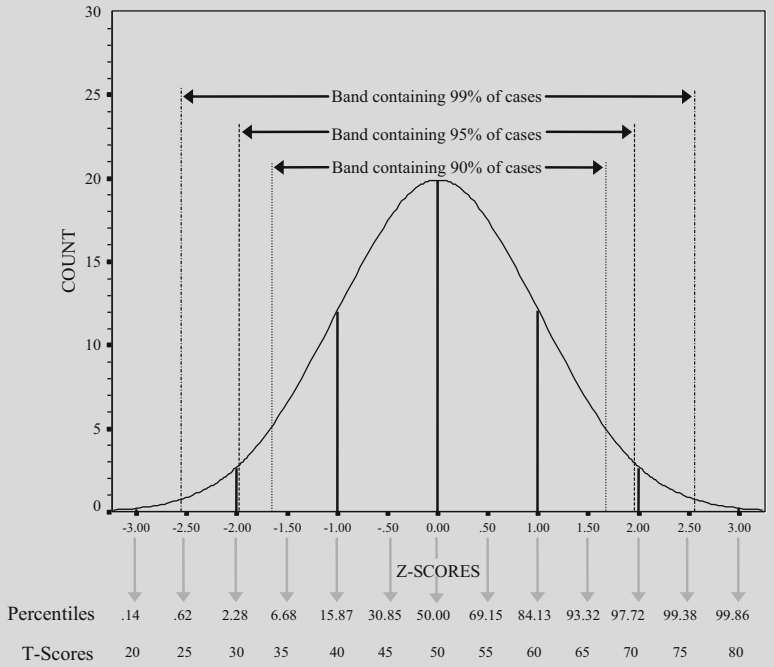

Fig. 5.28 The normal (bell-shaped or Gaussian) distribution

Figure 5.28 also shows how T-scores relate to the standard normal distribution and to z-scores. The mean T-score falls at 50 and each increment or decrement of 10 T-score units means a movement of another standard deviation away from this mean of 50. Thus, a T-score of 80 corresponds to a z-score of +3.0—a score 3 standard deviations higher than the mean of 50.

Of special interest to behavioural researchers are the values for z-scores in a standard normal distribution that encompass 90% of observations ($z = \pm 1.645$—isolating 5% of the distribution in each tail), 95% of observations ($z = \pm 1.96$—isolating 2.5% of the distribution in each tail), and 99% of observations ($z = \pm 2.58$—isolating 0.5% of the distribution in each tail).

Depending upon the degree of certainty required by the researcher, these bands describe regions outside of which one might define an observation as being atypical or as perhaps not belonging to a distribution being centred at a mean of 0.0. Most often, what is taken as atypical or rare in the standard normal distribution is a score at least two standard deviations away from the mean, in either direction. Why choose two standard deviations? Since in the standard normal distribution, only about 5% of

observations will fall outside a band defined by z-scores of ± 1.96 (rounded to 2 for simplicity), this equates to data values that are 2 standard deviations away from their mean. This can give us a defensible way to identify outliers or extreme values in a distribution.

Thinking ahead to what you will encounter in *Chap. 7*, this 'banding' logic can be extended into the world of statistics (like means and percentages) as opposed to just the world of observations. You will frequently hear researchers speak of some statistic estimating a specific value (a *parameter*) in a population, plus or minus some other value.

For example *A survey organisation might report political polling results in terms of a percentage and an error band, e.g. 59% of Australians indicated that they would vote Labour at the next federal election, plus or minus 2%.*

Most commonly, this error band ($\pm 2\%$) is defined by possible values for the population parameter that are about two standard deviations (or two standard errors—a concept discussed further in *Fundamental Concept VIII*) away from the reported or estimated statistical value. In effect, the researcher is saying that on 95% of the occasions he/she would theoretically conduct his/her study, the population value estimated by the statistic being reported would fall between the limits imposed by the endpoints of the error band (the official name for this error band is a *confidence interval*; see *Procedure 8.3*). The well-understood mathematical properties of the standard normal distribution are what make such precise statements about levels of error in statistical estimates possible.

Checking for Normality

It is important to understand that transforming the raw scores for a variable to z-scores (recall *Procedure 5.7*) does not produce z-scores which follow a normal distribution; rather they will have the same distributional shape as the original scores. However, if you are willing to assume that the normal distribution is the correct reference distribution in the population, then you are justified is interpreting z-scores in light of the known characteristics of the normal distribution.

In order to justify this assumption, not only to enhance the interpretability of z-scores but more generally to enhance the integrity of parametric statistical analyses, it is helpful to actually look at the sample frequency distributions for variables (using a *histogram* (illustrated in *Procedure 5.2)* or a *boxplot* (illustrated in *Procedure 5.6*), for example), since non-normality can often be visually detected. It is important to note that in the social and behavioural sciences as well as in economics and finance, certain variables tend to be non-normal by their very nature. This includes variables that measure time taken to complete a task, achieve a goal or make decisions and variables that measure, for example, income, occurrence of rare or extreme events or organisational size. Such variables tend to be positively skewed in the population, a pattern that can often be confirmed by graphing the distribution.

If you cannot justify an assumption of 'normality', you may be able to force the data to be normally distributed by using what is called a 'normalising transformation'. Such transformations will usually involve a nonlinear mathematical conversion (such as computing the logarithm, square root or reciprocal) of the raw scores. Such transformations will force the data to take on a more normal appearance so that the assumption of 'normality' can be reasonably justified, but at the cost of creating a new variable whose units of measurement and interpretation are more complicated. [For some non-normal variables, such as the occurrence of rare, extreme or catastrophic events (e.g. a 100-year flood or forest fire, coronavirus pandemic, the Global Financial Crisis or other type of financial crisis, man-made or natural disaster), the distributions cannot be 'normalised'. In such cases, the researcher needs to model the distribution as it stands. For such events, *extreme value theory* (e.g. see Diebold et al. 2000) has proven very useful in recent years. This theory uses a variation of the Pareto or Weibull distribution as a reference, rather than the normal distribution, when making predictions.]

Figure 5.29 displays before and after pictures of the effects of a logarithmic transformation on the positively skewed **speed** variable from the QCI database. Each graph, produced using NCSS, is of the hybrid histogram-density trace-boxplot type first illustrated in *Procedure 5.6*. The left graph clearly shows the strong positive skew in the **speed** scores and the right graph shows the result of taking the $\log_{10}$ of each raw score.

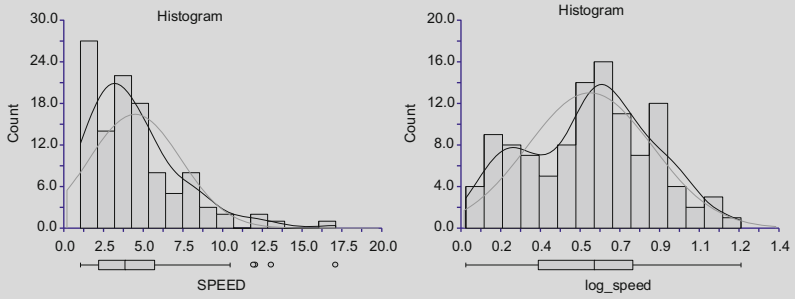

Fig. 5.29 Combined histogram-density trace-boxplot graphs displaying the before and after effects of a 'normalising' $\log_{10}$ transformation of the **speed** variable

Notice how the long tail toward slow **speed** scores is pulled in toward the mean and the very short tail toward fast **speed** scores is extended away from the mean. The result is a more 'normal' appearing distribution. The assumption would then be that we could assume normality of **speed** scores, but only in a $\log_{10}$ format (i.e. it is the log of **speed** scores that we assume is normally distributed in the population). In general, taking the logarithm of raw scores provides a satisfactory remedy for positively skewed distributions (but not for negatively skewed ones). Furthermore, anything we do with the transformed **speed** scores now has to be interpreted in units of $\log_{10}$ (seconds) which is a more complex interpretation to make.

Another visual method for detecting non-normality is to graph what is called a *normal Q-Q plot* (the Q-Q stands for Quantile-Quantile). This plots the percentiles for the observed data against the percentiles for the standard normal distribution (see Cleveland 1995 for more detailed discussion; also see Lane 2007, http://onlinestatbook.com/2/advanced_graphs/ q-q_plots.html). If the pattern for the observed data follows a normal distribution, then all the points on the graph will fall approximately along a diagonal line.

Figure 5.30 shows the normal Q-Q plots for the original **speed** variable and the transformed **log-speed** variable, produced using the SPSS *Explore...* procedure. The diagnostic diagonal line is shown on each graph. In the left-hand plot, for **speed**, the plot points clearly deviate from the diagonal in a way that signals positive skewness. The right-hand plot, for **log_speed,** shows the plot points generally falling along the diagonal line thereby conforming much more closely to what is expected in a normal distribution.

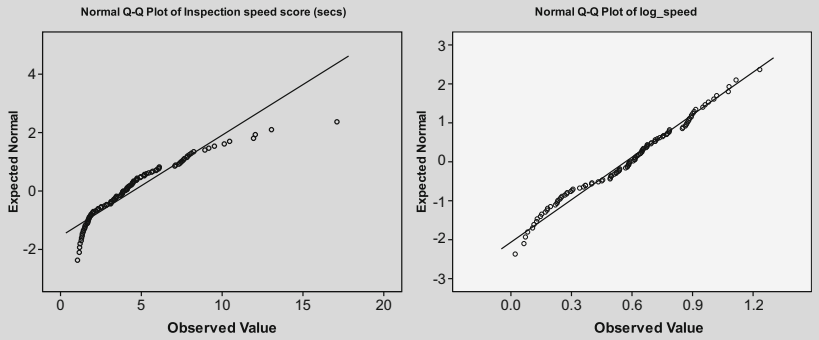

Fig. 5.30 Normal Q-Q plots for the original **speed** variable and the new **log_speed** variable

In addition to visual ways of detecting non-normality, there are also numerical ways. As highlighted in Chap. 1, there are two additional characteristics of any distribution, namely *skewness* (asymmetric distribution tails) and *kurtosis* (peakedness of the distribution). Both have an associated statistic that provides a measure of that characteristic, similar to the mean and standard deviation statistics. In a normal distribution, the values for the skewness and kurtosis statistics are both zero (skewness = 0 means a symmetric distribution; kurtosis = 0 means a mesokurtic distribution). The further away each statistic is from zero, the more the distribution deviates from a normal shape. Both the skewness statistic and the kurtosis statistic have *standard errors* (see *Fundamental Concept VIII*) associated with them (which work very much like the standard deviation, only for a statistic rather than for observations); these can be routinely computed by almost any statistical package when you request a descriptive analysis. Without going into the logic right now (this will come in *Fundamental Concept V*), a rough rule of thumb you can use to check for normality using the skewness and kurtosis statistics is to do the following:

- Prepare: Take the standard error for the statistic and multiply it by 2 (or 3 if you want to be more conservative).
- Interval: Add the result from the Prepare step to the value of the statistic and subtract the result from the value of the statistic. You will end up with two numbers, one low - one high, that define the ends of an interval (what you have just created approximates what is called a 'confidence interval', see *Procedure 8.3*).
- Check: If zero falls inside of this interval (i.e. between the low and high endpoints from the Interval step), then there is likely to be no significant issue with that characteristic of the distribution. If zero falls outside of the interval (i.e. lower than the low value endpoint or higher than the high value endpoint), then you likely have an issue with non-normality with respect to that characteristic.

Visually, we saw in the left graph in Fig. 5.29 that the **speed** variable was highly positively skewed. What if Maree wanted to check some numbers to support this judgment? She could ask SPSS to produce the skewness and kurtosis statistics for both the original **speed** variable and the new **log_speed** variable using the *Frequencies...* or the *Explore...* procedure. Table 5.6 shows what SPSS would produce if the *Frequencies...* procedure were used.

Table 5.6 Skewness and kurtosis statistics and their standard errors for both the original **speed** variable and the new **log_speed** variable

Statistics

		speed	log_speed
N	Valid	111	111
	Missing	1	1
Mean		4.4801	.5676
Std. Deviation		2.88751	.27491
Skewness		1.487	-.050
Std. Error of Skewness		.229	.229
Kurtosis		3.071	-.672
Std. Error of Kurtosis		.455	.455

Using the 3-step check rule described above, Maree could roughly evaluate the normality of the two variables as follows:

For **speed**:

- *skewness*: [Prepare] $2 \times .229 = .458 \rightarrow$ [Interval] $1.487 - .458 = 1.029$ and $1.487 + .458 = 1.945 \rightarrow$ [Check] zero does not fall inside the interval bounded by 1.029 and 1.945, so there appears to be a significant problem with skewness. Since the value for the skewness statistic (1.487) is positive, this means the problem is positive skewness, confirming what the left graph in Fig. 5.29 showed.

(continued)

- *kurtosis*: [Prepare] $2 \times .455 = .91 \rightarrow$ [Interval] $3.071 - .91 = 2.161$ and $3.071 + .91 = 3.981 \rightarrow$ [Check] zero does not fall in interval bounded by 2.161 and 3.981, so there appears to be a significant problem with kurtosis. Since the value for the kurtosis statistic (1.487) is positive, this means the problem is leptokurtosis—the peakedness of the distribution is too tall relative to what is expected in a normal distribution.

For **log_speed:**

- *skewness*: [Prepare] $2 \times .229 = .458 \rightarrow$ [Interval] $-.050 - .458 = -.508$ and $-.050 + .458 = .408 \rightarrow$ [Check] zero falls within interval bounded by $-.508$ and .408, so there appears to be no problem with skewness. The log transform appears to have corrected the problem, confirming what the right graph in Fig. 5.29 showed.
- *kurtosis*: [Prepare] $2 \times .455 = .91 \rightarrow$ [Interval] $-.672 - .91 = -1.582$ and $-.672 + .91 = .238 \rightarrow$ [Check] zero falls within interval bounded by -1.582 and .238, so there appears to be no problem with kurtosis. The log transform appears to have corrected this problem as well, rendering the distribution more approximately mesokurtic (i.e. normal) in shape.

There are also more formal tests of significance (see *Fundamental Concept V*) that one can use to numerically evaluate normality, such as the *Kolmogorov-Smirnov test* and the *Shapiro-Wilk's test*. Each of these tests, for example, can be produced by SPSS on request, via the *Explore...* procedure.

References

References for Procedure 5.1

Allen, P., Bennett, K., & Heritage, B. (2019). *SPSS statistics: A practical guide* (4th ed.). South Melbourne, VIC: Cengage Learning Australia Pty. ch. 3.

George, D., & Mallery, P. (2019). *IBM SPSS statistics 25 step by step: A simple guide and reference* (15th ed.). New York: Routledge. ch. 6 and 8.

Useful Additional Readings for Procedure 5.1

Agresti, A. (2018). *Statistical methods for the social sciences* (5th ed.). Boston: Pearson. ch. 3.

Argyrous, G. (2011). *Statistics for research: With a guide to SPSS* (3rd ed.). London: Sage. ch. 4–5.

De Vaus, D. (2002). *Analyzing social science data: 50 key problems in data analysis*. London: Sage. ch. 28, 32.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Upper Saddle River, NJ: Pearson. ch. 3.
Gravetter, F. J., & Wallnau, L. B. (2017). *Statistics for the behavioural sciences* (10th ed.). Belmont, CA: Wadsworth Cengage. ch. 2.
Steinberg, W. J. (2011). *Statistics alive* (2nd ed.). Los Angeles: Sage. ch. 3.

References for Procedure 5.2

Chang, W. (2019). *R graphics cookbook: Practical recipes for visualizing data* (2nd ed.). Sebastopol, CA: O'Reilly Media. ch. 2–5.
Jacoby, W. G. (1997). *Statistical graphics for univariate and bivariate data*. Thousand Oaks, CA: Sage.
McCandless, D. (2014). *Knowledge is beautiful*. London: William Collins.
Smithson, M. J. (2000). *Statistics with confidence*. London: Sage. ch. 3.
Toseland, M., & Toseland, S. (2012). *Infographica: The world as you have never seen it before*. London: Quercus Books.
Wilkinson, L. (2009). Cognitive science and graphic design. In SYSTAT Software Inc (Ed.), *SYSTAT 13: Graphics* (pp. 1–21). Chicago, IL: SYSTAT Software Inc.

Useful Additional Readings for Procedure 5.2

Argyrous, G. (2011). *Statistics for research: With a guide to SPSS* (3rd ed.). London: Sage. ch. 4–5.
De Vaus, D. (2002). *Analyzing social science data: 50 key problems in data analysis*. London: Sage. ch. 29, 33.
Field, A. (2018). *Discovering statistics using SPSS for windows* (5th ed.). Los Angeles: Sage. ch. 5.
George, D., & Mallery, P. (2019). *IBM SPSS statistics 25 step by step: A simple guide and reference* (15th ed.). Boston, MA: Pearson Education. ch. 5.
Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Upper Saddle River, NJ: Pearson. ch. 3.
Hintze, J. L. (2012). *NCSS 8 help system: Graphics*. Kaysville, UT: Number Cruncher Statistical Systems. ch. 141–143, 155, 161.
StatPoint Technologies, Inc. (2010). *STATGRAPHICS Centurion XVI user manual*. Warrenton, VA: StatPoint Technologies Inc.. ch. 4.
Steinberg, W. J. (2011). *Statistics alive* (2nd ed.). Los Angeles: Sage. ch. 4.
SYSTAT Software Inc. (2009). *SYSTAT 13: Graphics*. Chicago, IL: SYSTAT Software Inc. ch. 2, 5.

References for Procedure 5.3

Cleveland, W. R. (1995). *Visualizing data*. Summit, NJ: Hobart Press.
Jacoby, W. J. (1998). *Statistical graphics for visualizing multivariate data*. Thousand Oaks, CA: Sage.
SYSTAT Software Inc. (2009). *SYSTAT 13: Graphics*. Chicago, IL: SYSTAT Software Inc. ch. 6.

Useful Additional Readings for Procedure 5.3

Hintze, J. L. (2012). *NCSS 8 help system: Graphics*. Kaysville, UT: Number Cruncher Statistical Systems. ch. 162.
Kirk, A. (2016). *Data visualisation: A handbook for data driven design*. Los Angeles: Sage.
Knaflic, C. N. (2015). *Storytelling with data: A data visualization guide for business professionals*. Hoboken, NJ: Wiley.
Tufte, E. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CN: Graphics Press.

Reference for Procedure 5.4

Field, A. (2018). *Discovering statistics using SPSS for windows* (5th ed.). Los Angeles: Sage. ch. 6.

Useful Additional Readings for Procedure 5.4

Agresti, A. (2018). *Statistical methods for the social sciences* (5th ed.). Boston: Pearson. ch. 3.
Argyrous, G. (2011). *Statistics for research: With a guide to SPSS* (3rd ed.). London: Sage. ch. 9.
De Vaus, D. (2002). *Analyzing social science data: 50 key problems in data analysis*. London: Sage. ch. 30.
George, D., & Mallery, P. (2019). *IBM SPSS statistics 25 step by step: A simple guide and reference* (15th ed.). New York: Routledge. ch. 7.
Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Upper Saddle River, NJ: Pearson. ch. 4.
Gravetter, F. J., & Wallnau, L. B. (2017). *Statistics for the behavioural sciences* (10th ed.). Belmont, CA: Wadsworth Cengage. ch. 3.
Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill Inc. ch. 13.
Steinberg, W. J. (2011). *Statistics alive* (2nd ed.). Los Angeles: Sage. ch. 5.

References for Procedure 5.5

Field, A. (2018). *Discovering statistics using SPSS for windows* (5th ed.). Los Angeles: Sage. ch. 6.

Useful Additional Readings for Procedure 5.5

Agresti, A. (2018). *Statistical methods for the social sciences* (5th ed.). Boston: Pearson. ch. 3.
Argyrous, G. (2011). *Statistics for research: With a guide to SPSS* (3rd ed.). London: Sage. ch. 11.
De Vaus, D. (2002). *Analyzing social science data: 50 key problems in data analysis*. London: Sage. ch 15.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Upper Saddle River, NJ: Pearson. ch. 6.
Gravetter, F. J., & Wallnau, L. B. (2012). *Statistics for the behavioural sciences* (9th ed.). Belmont, CA: Wadsworth Cengage. ch. 5.
Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill Inc. ch. 13.
Steinberg, W. J. (2011). *Statistics alive* (2nd ed.). Los Angeles: Sage. ch. 8.

References for Fundamental Concept I

Smithson, M. J. (2000). *Statistics with confidence*. London: Sage. ch. 4.

Useful Additional Readings for Fundamental Concept I

Agresti, A. (2018). *Statistical methods for the social sciences* (5th ed.). Boston: Pearson. ch. 4.
Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Upper Saddle River, NJ: Pearson. ch. 9.
Gravetter, F. J., & Wallnau, L. B. (2017). *Statistics for the behavioural sciences* (10th ed.). Belmont, CA: Wadsworth Cengage. ch. 6.
Howell, D. C. (2013). *Statistical methods for psychology* (8th ed.). Belmont, CA: Cengage Wadsworth. ch. 5.
Steinberg, W. J. (2011). *Statistics alive* (2nd ed.). Los Angeles: Sage. ch. 10.

References for Procedure 5.6

Norušis, M. J. (2012). *IBM SPSS statistics 19 guide to data analysis*. Upper Saddle River, NJ: Prentice Hall. ch. 7.
Field, A. (2018). *Discovering statistics using SPSS for Windows* (5th ed.). Los Angeles: Sage. ch. 5, section 5.5.
Hintze, J. L. (2012). *NCSS 8 help system: Introduction*. Kaysville, UT: Number Cruncher Statistical System. ch. 152, 200.
StatPoint Technologies, Inc. (2010). *STATGRAPHICS Centurion XVI user manual*. Warrenton, VA: StatPoint Technologies Inc..
SYSTAT Software Inc. (2009a). *SYSTAT 13: Graphics*. Chicago, IL: SYSTAT Software Inc. ch. 3.
SYSTAT Software Inc. (2009b). *SYSTAT 13: Statistics - I*. Chicago, IL: SYSTAT Software Inc. ch. 1 and 9.

Useful Additional Readings for Procedure 5.6

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Upper Saddle River, NJ: Pearson. ch. 3.

Hartwig, F., & Dearing, B. E. (1979). *Exploratory data analysis*. Beverly Hills, CA: Sage.
Howell, D. C. (2013). *Statistical methods for psychology* (8th ed.). Belmont, CA: Cengage Wadsworth. ch. 2.
Leinhardt, G., & Leinhardt, L. (1997). Exploratory data analysis. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: An international handbook* (2nd ed., pp. 519–528). Oxford: Pergamon Press.
Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill, Inc.. ch. 13.
Smithson, M. J. (2000). *Statistics with confidence*. London: Sage. ch. 3.
Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley Publishing. [This is the classic text in the area, written by the statistician who developed most of the standard EDA techniques in use today].
Velleman, P. F., & Hoaglin, D. C. (1981). *ABC's of EDA*. Boston: Duxbury Press.

Useful Additional Readings for Procedure 5.7

Agresti, A. (2018). *Statistical methods for the social sciences* (5th ed.). Boston: Pearson. ch. 3.
Argyrous, G. (2011). *Statistics for research: With a guide to SPSS* (3rd ed.). London: Sage. ch. 10.
De Vaus, D. (2002). *Analyzing social science data: 50 key problems in data analysis*. London: Sage. ch. 30.
George, D., & Mallery, P. (2019). *IBM SPSS statistics 25 step by step: A simple guide and reference* (15th ed.). New York: Routledge. ch. 7.
Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Upper Saddle River, NJ: Pearson. ch. 5.
Gravetter, F. J., & Wallnau, L. B. (2017). *Statistics for the behavioural sciences* (10th ed.). Belmont, CA: Wadsworth Cengage. ch. 4.
Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill Inc. ch. 13.
Steinberg, W. J. (2011). *Statistics alive* (2nd ed.). Los Angeles: Sage. ch. 6.

References for Fundemental Concept II

Cleveland, W. R. (1995). *Visualizing data*. Summit, NJ: Hobart Press.
Diebold, F. X., Schuermann, T., & Stroughair, D. (2000). Pitfalls and opportunities in the use of extreme value theory in risk management. *The Journal of Risk Finance, 1*(2), 30–35.
Lane, D. (2007). *Online statistics education: A multimedia course of study*. Houston, TX: Rice University. http://onlinestatbook.com/.

Useful Additional Readings for Fundemental Concept II

Argyrous, G. (2011). *Statistics for research: With a guide to SPSS* (3rd ed.). London: Sage. ch. 11.
De Vaus, D. (2002). *Analyzing social science data: 50 key problems in data analysis*. London: Sage. ch. 11.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Upper Saddle River, NJ: Pearson. ch. 6.

Howell, D. C. (2013). *Statistical methods for psychology* (8th ed.). Belmont, CA: Cengage Wadsworth. ch. 3.

Keller, D. K. (2006). *The tao of statistics: A path to understanding (with no math)*. Thousand Oaks, CA: Sage. ch. 10.

Steinberg, W. J. (2011). *Statistics alive* (2nd ed.). Los Angeles: Sage. ch. 7, 8, 9.