

Analyzing data quantitatively

Prepared by Associate Professor:
Ahmed Rageh Ismail

ENT 636
Research Project

Lecture 8



Chapter

08

Learning outcomes

By the end of this chapter you should:

- identify the main issues that you need to consider when preparing data for quantitative analysis;
- recognize different types of data and understand the implications of data type for subsequent analyses;
- select the most appropriate tables and graphs to explore and illustrate different aspects of your data;
- select the most appropriate statistics to describe individual variables and to examine relationships between variables and trends in your data;
- interpret the tables, graphs and statistics that you use correctly.

Preparing data for quantitative analysis

When preparing data for quantitative analysis you need to be clear about the:

- definition and selection of cases;
- data type or types (scale of measurement);
- numerical codes used to classify data to ensure they will enable your research questions to be answered.

We now consider each of these.

Definition and selection of cases

A case is defined as an individual unit for which data have been collected. For example, a case might be a respondent who had completed a questionnaire, an individual organization or country for which secondary data had already been compiled, a magazine advertisement, a television commercial, or an organization's tweets.

The data set would comprise the data collected from the respondents, organizations or countries, magazine advertisements, television commercials or organization's tweets you intend to analyze.

Data types

Data for quantitative analysis can be divided into two distinct groups: **categorical** and **numerical**. Categorical data refer to data whose values cannot be measured numerically but can be either classified into sets (categories) according to the characteristics that identify or describe the variable or placed in rank order.

They can be further subdivided into descriptive and ranked. An auto manufacturer might categorize the types of vehicles it produces as hatchback, saloon, estate and SUV.

You might classify aspects of an image in terms of the gender of the person depicted and whether or not she or he is smiling.

Data types

The verbal responses to an open-ended interview question asking participants to describe their journey to work could, once transcribed into text, be used to generate data about their main mode of travel to work. These could be categorized as 'bicycle', 'bus', 'rail', 'car' or 'walk'. Alternatively, you may be looking at particular concepts in illustrations in annual reports such as whether the central figure in each is male or female.

Although the sources of these data differ, they are all known as **nominal data** or descriptive data as it is impossible to define such a category numerically or rank it. Rather, these data simply count the number of occurrences in each category of a variable.

Chapter

08

Data types

For virtually all analyses the categories should be unambiguous and discrete; in other words, having one particular feature, such as a vehicle being an SUV, excludes it from being in all other vehicle categories. This prevents questions arising regarding which category an individual case belongs to. Although these data are purely descriptive, you can count them to establish which category has the most and whether cases are spread evenly between categories.

Some statisticians (and statistics) also separate descriptive data where there are only two categories. These are known as **dichotomous data**, as the variable is divided into two categories, such as the variable 'result' being divided into 'pass' and 'fail'.

Ordinal (or **rank**) data are a more precise form of categorical data. In such instances you know the relative position of each case within your data set, although the actual numerical measures (such as scores) on which the position is based are not recorded.

A researcher exploring an organization's online communication may rank each of that organization's tweets over a three-month period as positive, neutral or negative. Similarly, a questionnaire asking a rating or scale question, such as how strongly a respondent agrees with a statement, also collects ranked (ordinal) data. Despite this, some researchers argue that, where such data are likely to have similar size gaps between data values, they can be analyzed as if they were numerical interval data.

Numerical data are those whose values are measured or counted numerically as quantities. This means that numerical data are more precise than categorical as you can assign each data value a position on a numerical scale. It also means that you can analyze these data using a far wider range of statistics. There are two possible ways of subdividing numerical data: into **interval** or **ratio data** and, alternatively, into **continuous** or **discrete data**.

The Celsius temperature scale is a good example of an interval scale. Although the difference between, say, 20°C and 30°C is 10°C it does not mean that 30°C is one and a half times as warm. This is because 0°C does not represent a true zero. When it is 0°C outside, there is still some warmth, rather than none at all!

In contrast, for ratio data , you can also calculate the relative difference or ratio between any two data values for a variable. Consequently, if a multinational company makes a profit of \$1,000,000,000 in one year and \$2,000,000,000 the following year, we can say that profits have doubled.

Similarly, if you are estimating the number of people attending events such as political rallies using aerial photographs you might estimate the number of people at one event is half as many as at another.

Continuous data are those whose values can theoretically take any value (sometimes within a restricted range) provided that you can measure them accurately enough. Data such as furnace temperature, delivery distance and length of service are therefore continuous data.

Discrete data can, by contrast, be measured precisely. Each case takes one of a finite number of values from a scale that measures changes in discrete units. These data are often whole numbers (integers) such as the number of mobile phones manufactured, number of occurrences of a particular word or phrase in employer associations' communications, or number of illustrations containing non-white people in each issue of a fashion magazine over the last ten years.

Chapter

08

Data coding

All data for quantitative analysis should, with few exceptions, be recorded using numerical codes. This enables you to enter the data quickly and with fewer errors using the numeric keypad on your keyboard. It also makes subsequent analyses, in particular those that require re-coding of data to create new variables, more straightforward.

If you are using a spreadsheet, you will need to keep a list of codes for each variable. Statistical analysis software can store these so that each code is automatically labelled.



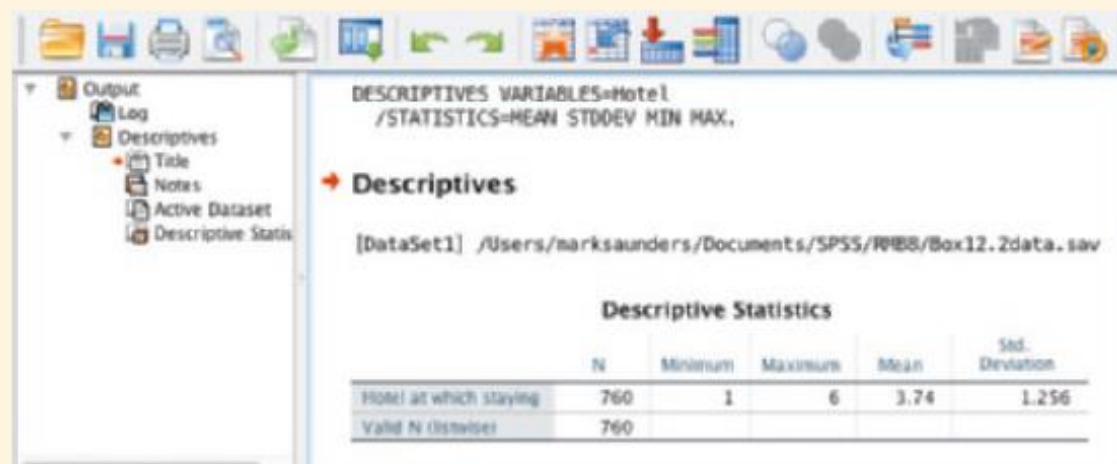
Box 12.2 Focus on student research

The implications of coding and data types for analysis

Pierre's research was concerned with customers' satisfaction for a small hotel group of six hotels. In collecting the data he had asked 760 customers to indicate the hotel at which they were staying when they completed their Internet questionnaires. When he downloaded his data, the survey design software had

automatically allocated a numerical code to represent the hotel, named the variable and labelled each of the codes. The code labels for the six hotels were:

Hotel at which staying	Code
Amsterdam	1
Antwerp	2
Eindhoven	3
Nijmegen	4
Rotterdam	5
Tilburg	6



In his initial analysis, Pierre used the analysis software to calculate descriptive statistics for every data variable, including the variable 'Hotel'. These included the minimum value (the code for Amsterdam), the maximum value (the code for Tilburg), the mean and the standard

deviation. Looking at his computer screen, Pierre noted that the mean (average) was 3.74 and the standard deviation was 1.256. He had forgotten that the data for this variable were categorical and, consequently, the descriptive statistics he had chosen were inappropriate.

Coding categorical data

For many secondary data sources (such as government surveys), a suitable coding scheme will have already been devised when the data were first collected. However, for other secondary sources such as documents (text, voice and visual) and all primary data you will need to decide on a coding scheme. Prior to this, you need to establish the highest level of precision required by your analyses .

Existing coding schemes can be used for many variables. These include industrial classification, occupation, social class and socioeconomic classification and ethnic group, social attitude variables as well as coding schemes devised and used by other researchers.

Coding categorical data

Wherever possible, we recommend you use these as they:

- save time;
- are normally well tested;
- allow comparisons of your findings with other research findings



Box 12.3

Focus on student research

Developing a codebook for open questions with multiple responses

As part of his research project, Amil used a questionnaire to collect data from the customers of a local themed restaurant. The questionnaire included an open list question, which asked 'List up to three things you like about this restaurant'. Respondents

could therefore provide more than one answer to the question, in other words multiple responses. Their answers included over 50 different 'things' that the 186 customers responding liked about the restaurant, the maximum number mentioned by any one customer being constrained to three by the phrasing of the question.

Once data had been collected, Amil devised a hierarchical coding scheme based on what the customers liked about the restaurant.

Extract from coding scheme used to classify responses:

Categories	Sub-categories	Response	Code
Physical surroundings		Decoration	1
		Use of colour	2
		Comfort of seating	3
Dining experience	Menu		10–49
			10–19
		Choice	11
	Food	Regularly changed	12
			20–29
		Freshly prepared	21
		Organic	22
		Served at correct temperature	23
	Staff attitude		30–39
		Knowledgeable	31
		Greet by name	32
		Know what diners prefer	33
		Discreet	34
		Do not hassle	35
		Good service	36
		Friendly	37
	Drinks	Have a sense of humour	38
			40–49
		Value for money	41
		Good selection of wines	42
		Good selection of beers	43
		Served at correct temperature	44

Content analysis

Content Analysis is a specific analytical technique of categorizing and coding text, voice and visual data using a systematic coding scheme to enable quantitative analysis. Although there are numerous definitions of Content Analysis, most draw on an early definition by Berelson (1952:18) as a “technique for the objective, systematic and quantitative description of the manifest content of communication.”

The technique is used to categorizes and code (and subsequently analyze) both the manifest and latent content of the data. **Manifest** content refers to those components that are clearly visible in the data and can be counted. **Latent** content refers to the meaning that may lie behind the manifest content.

Content analysis

‘Systematic’ in Berelson’s definition emphasizes that Content Analysis should be conducted in a consistent, transparent and replicable way with clear rules for defining and applying codes being detailed in a code book or coding manual.

This coding scheme can draw on existing schemes developed by other researchers or developed inductively from the data using similar techniques.



Box 12.4 Focus on management research

Developing a codebook for visual (and text) data

Lian and Yu (2017) conducted research into the online image of tourist destinations using photographic and textual information from official online media and online tourism marketers for the city of Huangshan in Japan. They published this in the *Asia Pacific Journal of Tourism Research*. In collecting data, they first selected a 10 per cent random sample of the 7,131 travel related pictures and 7,635 texts posted between January and June 2015. Building on previous research, this sample was used to establish categories of online images of tourist destinations and develop a codebook. Their final codebook comprised three categories, namely tourism facilities, tourism services and tourism resources. These were subdivided into 25 categories of photographic information and 31 sub categories of textual information. Those for the online photographic images are listed below:

Categories	Subcategories
Tourism facilities	<i>Cable car</i>
	<i>Accommodation</i>
	<i>Camping</i>
	<i>Vantage point</i>

Categories	Subcategories
Tourism services	<i>Safety fences</i>
	<i>Climbing stone steps</i>
	<i>Tour guide sign</i>
	<i>Catering</i>
Tourism resources	<i>Order maintaining</i>
	<i>Guest-greeting Pine</i>
	<i>Pinus taiwanensis</i>
	<i>Canyon</i>
	<i>Thermal springs</i>
	<i>Sea of clouds</i>
	<i>Absurd stones</i>
	<i>Waterfall</i>
	<i>Winter snow</i>
	<i>"The Light of Buddha"</i>
	<i>Rime</i>
	<i>Night sky</i>
	<i>Ravine stream</i>
	<i>"Concentric lock"</i>
	<i>Pediment</i>
	<i>Carved stone</i>

Subsequently all travel related pictures and texts were coded by both researchers and a content analysis undertaken. This revealed that the online image of Huangshan as a tourist destination comprised three elements: tourism resources, tourism facilities and tourism services, there being consistency across different media forms.

whether your system of categories is transparent and capable of being applied consistently by others is for you and a friend to code a sample of the same data separately using this system of categories and then to compare your results. This is known as inter-rater reliability and can be assessed by measuring the extent to which two coders agree. One way of doing this is to calculate the percentage agreement using the following formula:

$$PA = \frac{A}{n} \times 100$$

where:

PA = percentage agreement

A = number of agreements between the two coders

n = number of segments coded

Data entry and checking

When entering your data into analysis software you need to ensure the:

- data layout and format meet that required by the analysis software;
- data, once entered, have been saved and a back-up copy made;
- data have been checked for errors and any found corrected;
- need to weight cases has been considered

Chapter

08

Exploring and presenting data

Once your data have been entered and checked for errors, you are ready to start your analysis. Exploratory Data Analysis (EDA) approach useful in these initial stages.

This approach emphasizes the use of graphs to explore and understand your data. Although within data analysis the term graph has a specific meaning: ‘... a visual display that illustrates one or more relationships among numbers’ (Kosslyn 2006: 4), it is often used interchangeably with the term ‘chart’.

Consequently, while some authors (and data analysis software) use the term bar graphs, others use the term bar charts.



Box 12.8 Checklist

Designing your graphs and tables

For both graphs and tables

- ✓ Does it have a brief but clear and descriptive title?
- ✓ Are the units of measurement used stated clearly?
- ✓ Are the sources of data used stated clearly?
- ✓ Are there notes to explain abbreviations and unusual terminology?
- ✓ Does it state the size of the sample on which the values in the graph/table are based?

For graphs

- ✓ Does it have clear axis labels?
- ✓ Are bars and their components in the same logical sequence?
- ✓ Is more dense shading used for smaller areas?
- ✓ Have you avoided misrepresenting or distorting the data?
- ✓ Is a key or legend included (where necessary)?

For tables

- ✓ Does it have clear column and row headings?
- ✓ Are columns and rows in a logical sequence?
- ✓ Are numbers in columns right justified?

Exploring and presenting individual variables

The simplest way of summarizing data for individual variables so that specific amounts can be read is to use a table (frequency distribution).

For categorical data, the table summarizes the number of cases (frequency) in each category. For variables where there are likely to be a large number of categories (or values for numerical data), you will need to group the data into categories that reflect your research question(s) and objectives.

Chapter

08

Exploring and presenting individual variables

To show highest and lowest values

Tables attach no visual significance to highest or lowest data values unless emphasized by alternative fonts. Graphs can provide visual clues, although both categorical and numerical data may need grouping. For categorical and discrete data, **bar graphs** and **pictograms** are both suitable.

Chapter

08

Exploring and presenting individual variables

To show relative amounts

To emphasize the relative values represented by each of the bars in a bar graph, the bars may be reordered in either descending or ascending order of the frequency of occurrence represented by each bar.

For text data the relative proportions of key words and phrases can be shown using a word cloud, there being numerous free word cloud generators such as Wordle™ available online. In a word cloud the frequency of occurrence of a particular word or phrase is represented by the font size of the word or occasionally the color.

Exploring and presenting individual variables

Most researchers use a **histogram** to show highest and lowest values for continuous data. Prior to being drawn, data will often need to be grouped into class intervals. In a histogram, the area of each bar represents the frequency of occurrence and the continuous nature of the data is emphasized by the absence of gaps between the bars. For equal width class intervals, the height of your bar still represents the frequency of occurrences and so the highest and lowest values are easy to distinguish.

Exploring and presenting individual variables

To show a trend

Trends can only be presented for variables containing numerical (and occasionally ranked) longitudinal data. The most suitable diagram for exploring the trend is a **line graph** in which the data values for each time period are joined with a line to represent the trend.

Chapter

08

Exploring and presenting individual variables

To show proportions or percentages

Research has shown that the most frequently used diagram to emphasize the proportion or share of occurrences is the **pie chart**, although bar charts have been shown to give equally good results. A pie chart is divided into proportional segments according to the share each has of the total value and the total value represented by the pie is noted.

For numerical and some categorical data you will need to group data prior to drawing the pie chart, as it is difficult to interpret pie charts with more than six segments.

Chapter

08

Exploring and presenting individual variables

To show the distribution of values

Prior to using many statistical tests it is necessary to establish the distribution of values for variables containing numerical data. For continuous data, this can be visualized by plotting a histogram or frequency polygon. For discrete data a bar graph or frequency polygon can be plotted.

A frequency polygon is a line graph connecting the mid points of the bars of a histogram or bar graph. If your graph shows a bunching to the left and a long tail to the right, the data are **positively skewed**. If the converse is true, the data are **negatively skewed**. If your data are equally distributed either side of the highest frequency then they are **symmetrically distributed**. A special form of the symmetric distribution, in which the data can be plotted as a bell-shaped curve, is known as **normal distribution**.

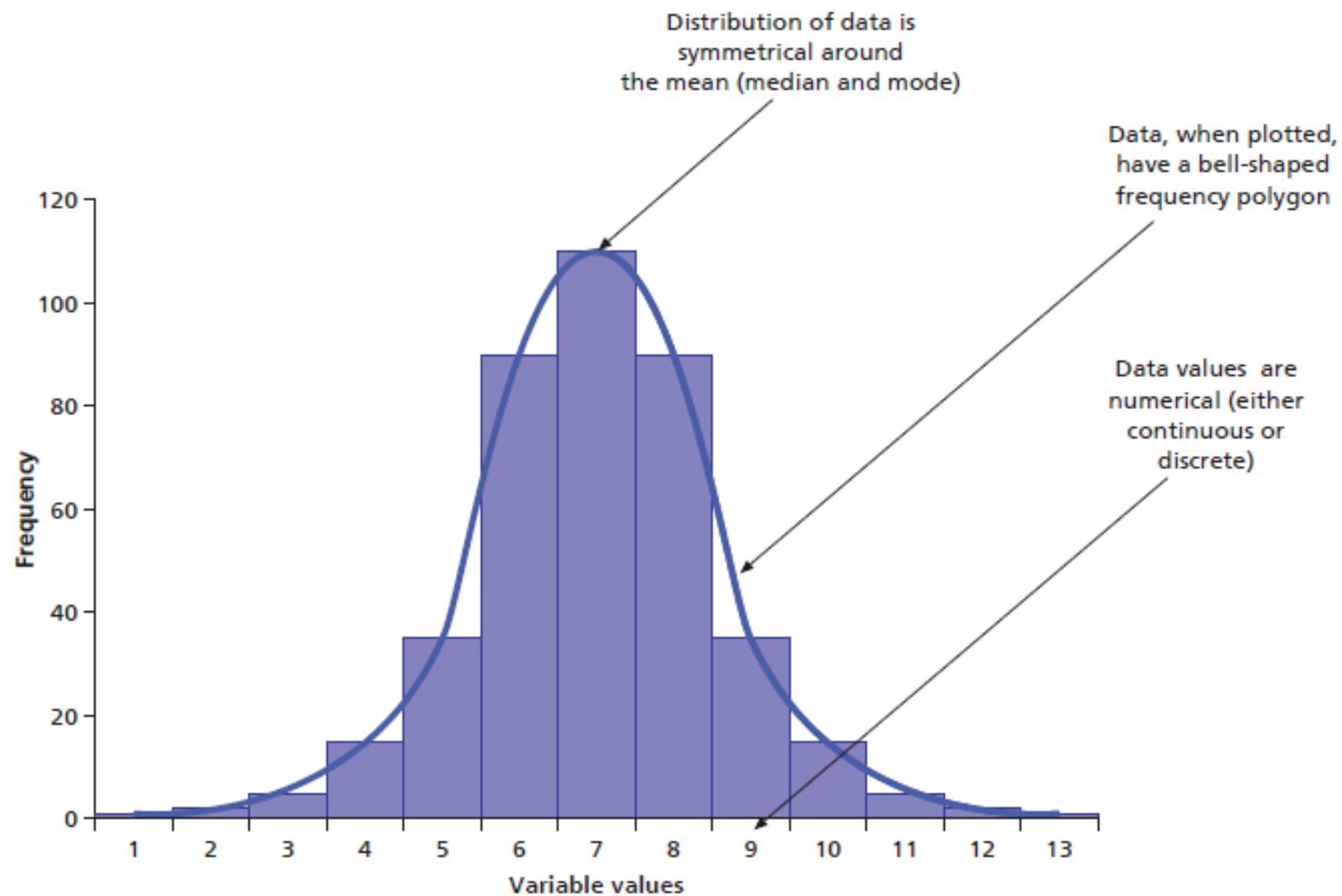


Figure 12.13 Annotated frequency polygon showing a normal distribution

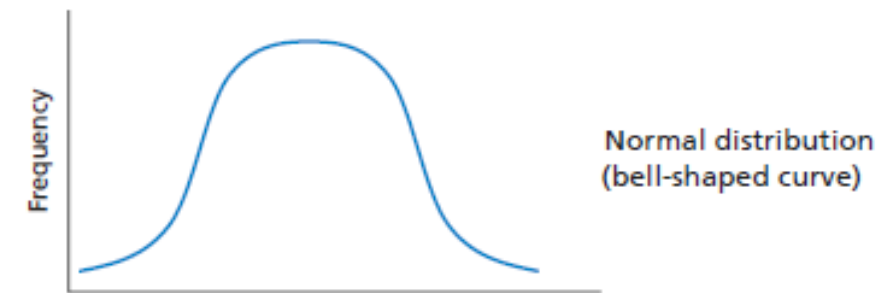
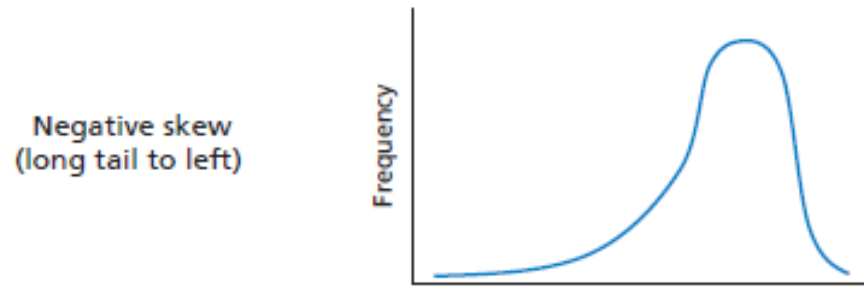
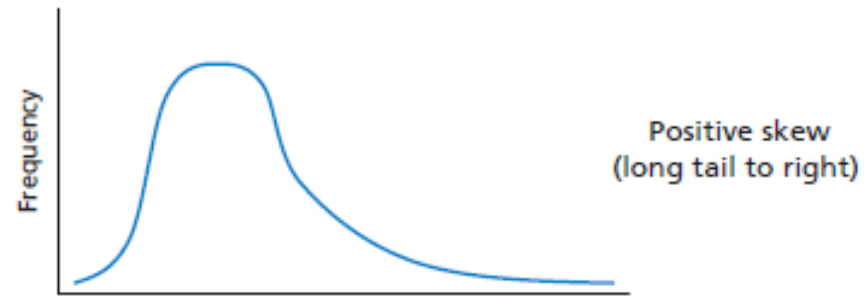


Figure 12.7 Frequency polygons showing distributions of values

Chapter

08

Comparing variables

To show interdependence and specific amounts

As with individual variables, the best method of showing interdependence between variables so that any specific amount can be discerned easily is a table. This is known as a **contingency table** or as a **cross-tabulation**.

For variables where there are likely to be a large number of categories (or values for numerical data), you may need to group the data to prevent the table from becoming too large.

Chapter

08

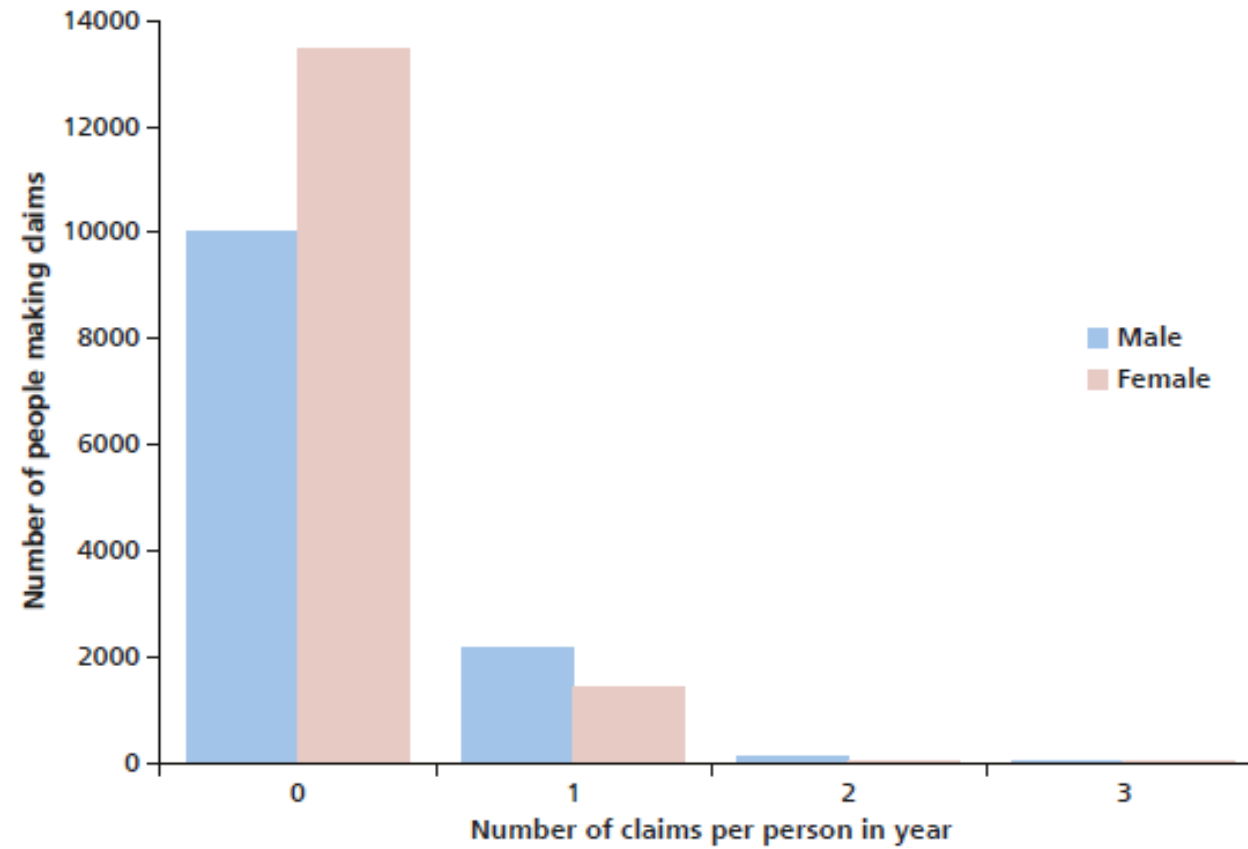
Comparing variables

To compare the highest and lowest values

Comparisons of variables that emphasize the highest and lowest rather than precise values are best explored using a **multiple bar graph**, also known as a **multiple bar chart**, alternatively known as a compound bar graph or compound bar chart.

Within any multiple bar graph you are likely to find it easiest to compare between adjacent bars. The multiple bar graph has therefore been drawn to emphasize comparisons between males and females rather than between numbers of claims.

Number of insurance claims by gender, 2018
Source: PJ Insurance Services



Chapter

08

Comparing variables

To compare proportions or percentages

Comparison of proportions between variables uses a percentage component bar graph (percentage component bar chart also known as a divided bar chart). This type of diagram can be used for all data types, provided that continuous data, and data where there are more than six values or categories, are grouped.

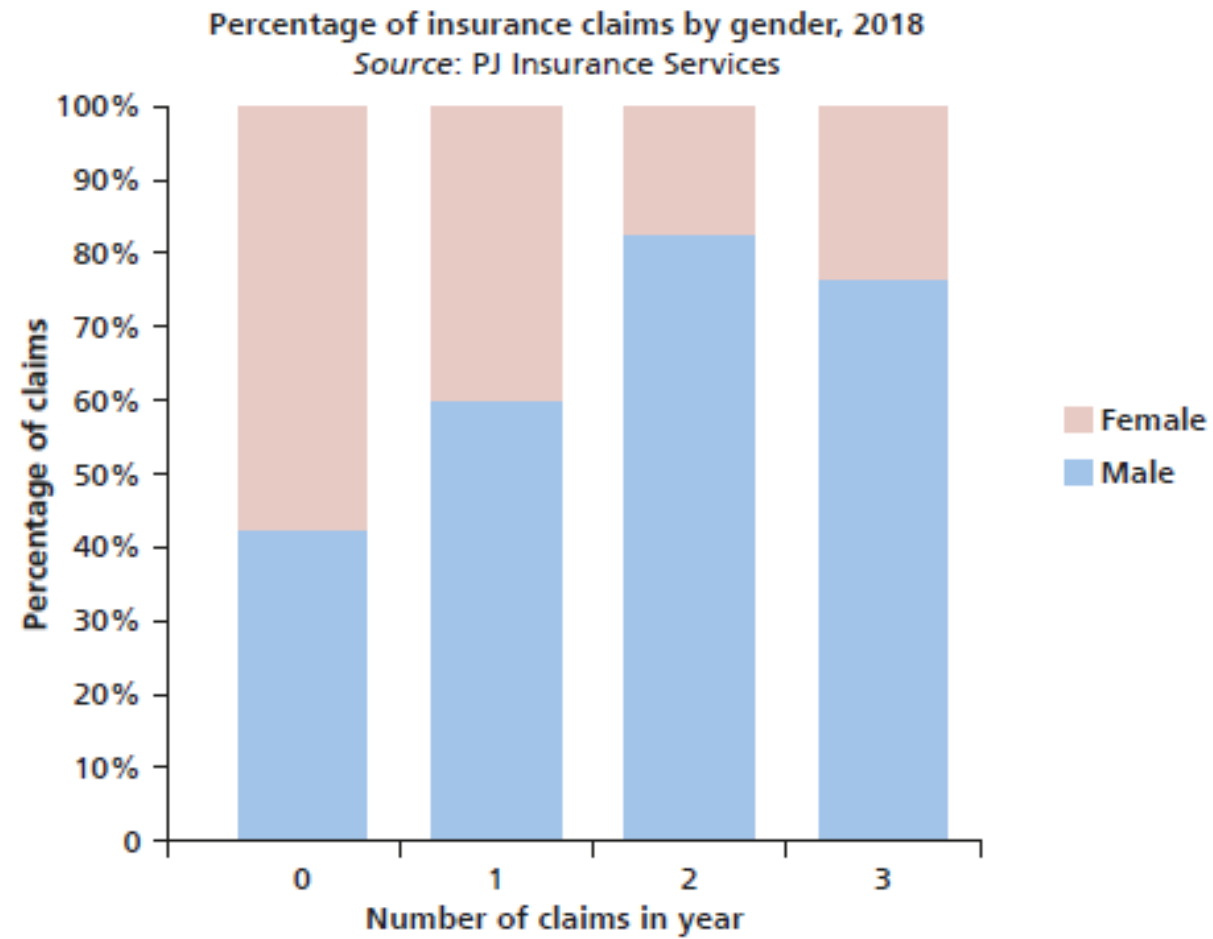


Figure 12.10 Percentage component bar graph

Chapter

08

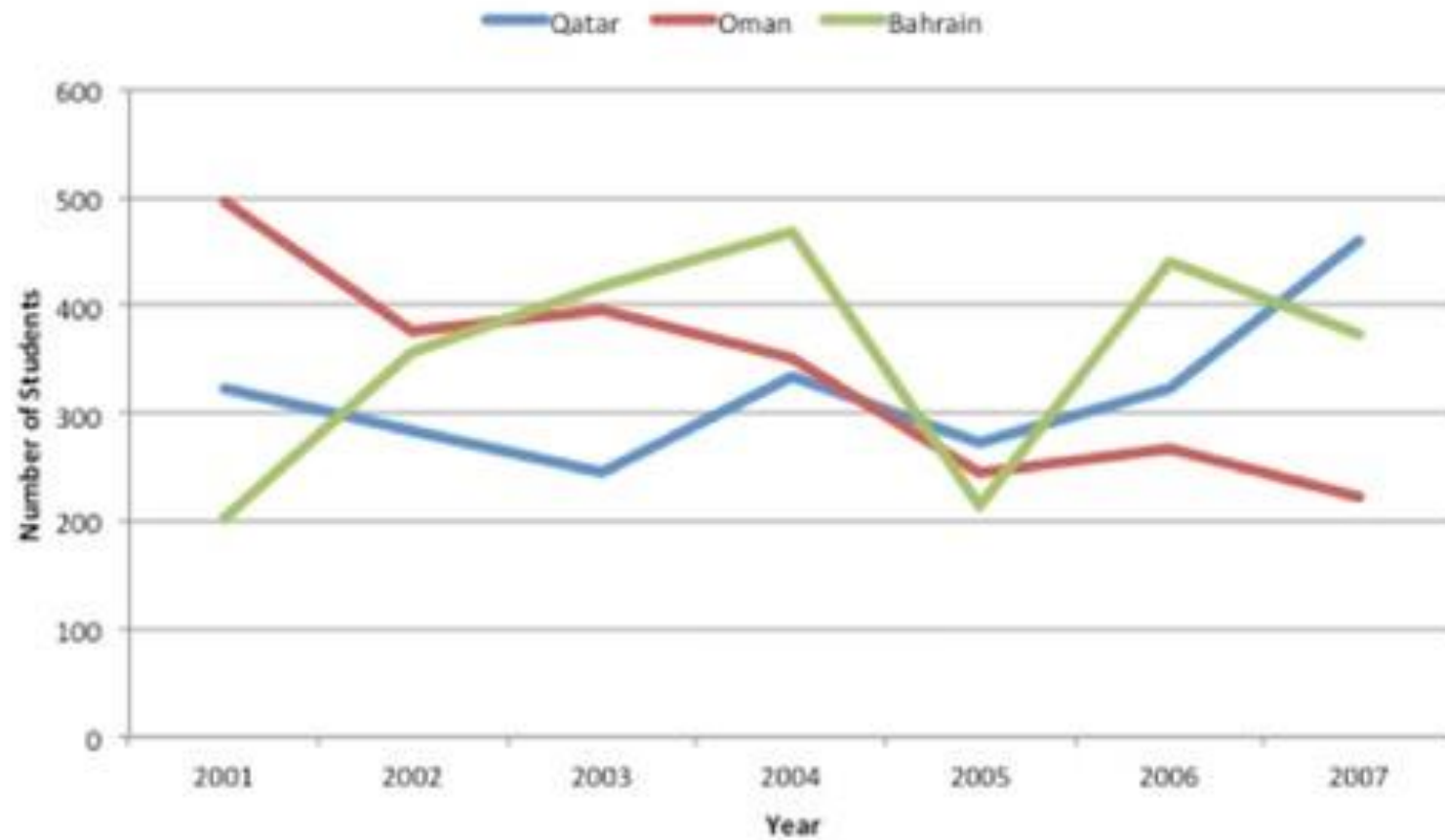
Comparing variables

To compare trends so the intersections are clear

The most suitable diagram to compare trends for two or more numerical (or occasionally ranked) variables is a **multiple line graph** where one line represents each variable. You can also use multiple bar graphs in which bars for the same time period are placed adjacent to each other.

Conjunctions in trends – that is, where values for two or more variables intersect – are shown by the place where the lines on a multiple line graph cross.

Number of Students Whose Major is Business



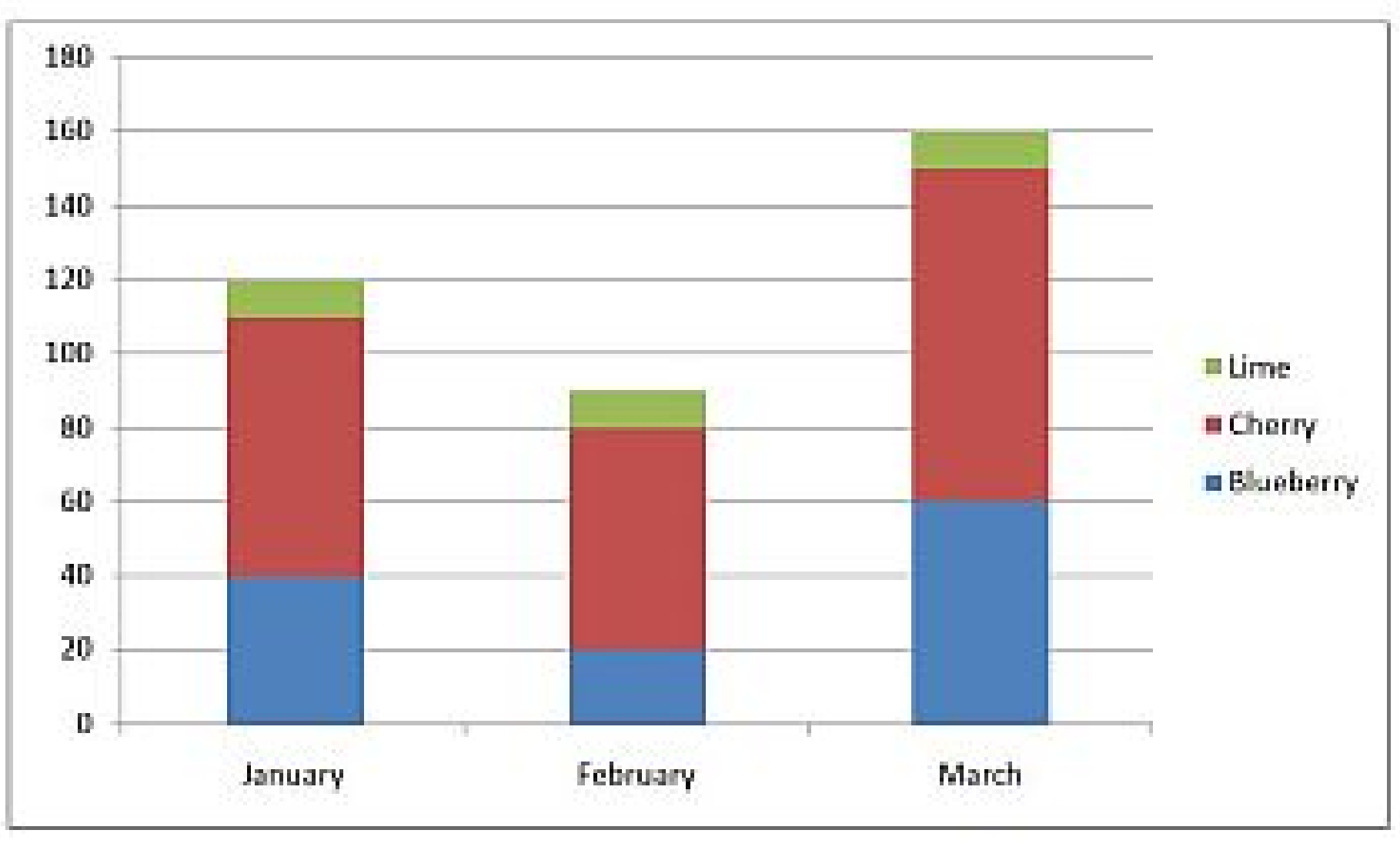
Chapter

08

Comparing variables

To compare the cumulative totals

Comparison of cumulative totals between variables uses a variation of the bar chart. A stacked bar graph, also known as a stacked bar chart, can be used for all data types provided that continuous data and data where there are more than six possible values or categories are grouped. As with percentage component bar graphs, the design of the stacked bar graph is dictated by the totals you want to compare.



Chapter

08

Comparing variables

To show the interrelationships between cases for variables

You can explore possible interrelationships between ranked and numerical data variables by plotting one variable against another. This is called a scatter graph (also known as a scatter plot), and each cross (point) represents the values for one case.

Convention dictates that you plot the dependent variable – that is, the variable that changes in response to changes in the other (independent) variable – against the vertical axis.

Chapter

08

Comparing variables

The strength of the interdependence or relationship is indicated by the closeness of the points to an imaginary straight line. If as the values for one variable increase so do those for the other then you have a positive relationship.

If as the values for one variable decrease those for the other variable increase then you have a negative relationship. Thus, in the figure below there is a negative relationship between the two variables. The strength of this relationship can be assessed statistically using techniques such as correlation or regression.



Figure 12.12 Scatter graph

Describing data using statistics

Descriptive statistics enable you to describe (and compare) a variable's data values numerically. Your research question(s) and objectives, although limited by the type of data (the table below), should guide your choice of statistics.

Statistics to describe a variable focus on two aspects of the data values' distribution:

- the central tendency;
- the dispersion.

Table 12.4 Descriptive statistics by data type: a summary

To calculate a measure of:		Categorical		Numerical	
		Nominal (Descriptive)	Ordinal (Ranked)	Continuous	Discrete
Central tendency that . . .	. . . represents the value that occurs most frequently	Mode			
	. . . represents the middle value			Median	
	. . . includes all data values (average)			Mean	
	. . . includes all data values other than those at the extremes of the distribution			Trimmed mean	

Dispersion that . . .	. . . states the difference between the highest and lowest values		Range (data need not be normally distributed but must be placed in rank order)
	. . . states the difference within the middle 50% of values		Inter-quartile range (data need not be normally distributed but must be placed in rank order)
	. . . states the difference within another fraction of the values		Deciles or percentiles (data need not be normally distributed but must be placed in rank order)
	. . . describes the extent to which data values differ from the mean		Variance, or more usually, the standard deviation (data should be normally distributed)
	. . . compares the extent to which data values differ from the mean between variables		Coefficient of variation (data should be normally distributed)
	. . . allows the relative extent that data values differ to be compared		Index numbers

Chapter

08

Describing data using statistics

Describing the central tendency

When describing data for both samples and populations quantitatively it is usual to provide some general impression of values that could be seen as common, middling or average. These are termed measures of central tendency and are discussed in virtually all statistics textbooks. The three main ways of measuring the central tendency most used in business research are the:

- value that occurs most frequently (mode);
- middle value or mid-point after the data have been ranked (median);
- value, often known as the average, that includes all data values in its calculation (mean).

Chapter

08

Describing data using statistics

To represent the value that occurs most frequently

The mode is the value that occurs most frequently. For descriptive data, the mode is the only measure of central tendency that can be interpreted sensibly.

Chapter

08

Describing data using statistics

To represent the middle value

If you have quantitative data it is also possible to calculate the middle or **median** value by ranking all the values in ascending order and finding the mid-point (or 50th percentile) in the distribution.

For variables that have an even number of data values, the median will occur halfway between the two middle data values. The median has the advantage that it is not affected by extreme values in the distribution.

Median

(Number of values is odd)

8 11 15 18 24 30 31



8 11 15 18 24 30 31

Median

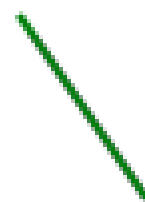
(Number of values is even)

11 15 18 24 30 31



11 15 18 21 24 30 31

18 19 20 21 22 23 24



Mean

7, 3, 4, 1, 7, 6

Sum of numbers divided
by the total numbers

$$\text{Mean} = (7+3+4+1+7+6)/6 \\ = 28/6 = 4.66$$

Median

7, 3, 4, 1, 7, 6

Arrange in order and
pick the middle value

1, 3, 4, 6, 7, 7

$$\text{Median} = (4+6)/2 = 5$$

Mode

7, 3, 4, 1, 7, 6

Most common number

7, 3, 4, 1, 7, 6

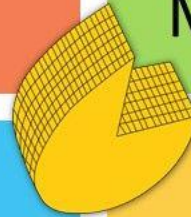
$$\text{Mode} = 7$$

Range

7, 3, 4, 1, 7, 6

Difference between
highest and lowest

$$\text{Range} = 7 - 1 = 6$$



Chapter

08

Describing data using statistics

To include all data values

The most frequently used measure of central tendency is the **mean** (average in everyday language), which includes all data values in its calculation. However, it is usually only possible to calculate a meaningful mean using numerical data.

The value of your mean is unduly influenced by extreme data values in skewed distributions. In such distributions the mean tends to get drawn towards the long tail of extreme data values and may be less representative of the central tendency.

Describing data using statistics

For this and other reasons, it is suggested that the median may be a more useful descriptive statistic. Alternatively, where the mean is affected by extreme data values (outliers) these may be excluded and a trimmed mean calculated.

This excludes a certain proportion (for example five per cent) of the data from both ends of the distribution, where the outliers are located. Because the mean is the building block for many of the statistical tests used to explore relationships, it is usual to include it as at least one of the measures of central tendency for numerical data in your report. This is, of course, provided that it makes sense!

Chapter

08

Describing data using statistics

Describing the dispersion

It is important to describe how the data values are dispersed around the central tendency. This is only possible for numerical data. Two of the most frequently used ways of describing the dispersion are the:

- difference within the middle 50 per cent of values (inter-quartile range);
- extent to which values differ from the mean (standard deviation).

Although these dispersion measures are suitable only for numerical data, most statistical analysis software will also calculate them for categorical data if you have used numerical codes.

Chapter

08

Describing data using statistics

To state the difference between values

In order to get a quick impression of the distribution of data values for a variable you could simply calculate the difference between the lowest and the highest values once they have been ranked in ascending order – that is, the range.

However, this statistic is rarely used in research reports as it represents only the extreme values. A more frequently used statistic is the **inter-quartile range**.

Chapter

08

Describing data using statistics

The median divides the range into two. The range can be further divided into four equal sections called quartiles. The lower quartile is the value below which a quarter of your data values will fall; the upper quartile is the value above which a quarter of your data values will fall.

As you would expect, the remaining half of your data values will fall between the lower and upper quartiles. The difference between the upper and lower quartiles is the inter-quartile range. As a consequence, it is concerned only with the middle 50 per cent of data values and ignores extreme values.

To describe and compare the extent by which values differ from the mean

Conceptually and statistically in research it is important to look at the extent to which the data values for a variable are spread around their mean, as this is what you need to know to assess its usefulness as a typical value for the distribution. If your data values are all close to the mean, then the mean is more typical than if they vary widely.

To describe the extent of spread of numerical data you use the standard deviation. If your data are a sample, this is calculated using a slightly different formula than if your data are a population, although if your sample is larger than about 30 cases there is little difference in the two statistics.

To describe and compare the extent by which values differ from the mean

You may need to compare the relative spread of data between distributions of different magnitudes (e.g. one may be measured in hundreds of tonnes, the other in billions of tonnes). To make a meaningful comparison you will need to take account of these different magnitudes. A common way of doing this is:

- 1 to divide the standard deviation by the mean;
- 2 then to multiply your answer by 100.

To describe and compare the extent by which values

This results in a statistic called the **coefficient of variation**. The values of this statistic can then be compared. The distribution with the largest coefficient of variation has the largest relative spread of data.

Alternatively, you may wish to compare the relative extent to which data values differ. One way of doing this is to use **index numbers** and consider the relative differences rather than actual data values. Such indices compare each data value against a base data value that is normally given the value of 100, differences being calculated relative to this value.

An index number greater than 100 represents a larger or higher data value relative to the base value and an index less than 100, a smaller or lower data value.

base value and an index less than 100, a marker of lower data value

To calculate an index number for each case for a data variable you use the following formula:

$$\text{Index number for case} = \frac{\text{data value for case}}{\text{base data value}} \times 100$$

Examining relationships and differences using statistics

When analyzing data quantitatively you are likely to ask questions such as: ‘Are these variables related?’, or ‘Do these groups differ?’ In statistical analysis you answer these questions by establishing the probability of the test statistic summarizing the relationship or difference in your data, or one more extreme, occurring.

This process of assessing the statistical significance of findings from a sample is known as **significance testing**, the classical approach to significance testing being hypothesis testing. Significance testing can therefore be thought of as assessing the possibility that your result could be due to random variation in your sample.

Examining relationships and differences using statistics

There are two main groups of statistical tests: **non-parametric** and **parametric**.

Nonparametric statistics are designed primarily for use with categorical (dichotomous, nominal, ordinal) data where there is no distributional model and so we cannot use statistics to estimate parameters. In contrast, **parametric statistics** are used with numerical (interval and ratio) data.

Although parametric statistics are considered more powerful because they use numerical data, a number of assumptions about the actual data being used need to be satisfied if they are not to produce spurious results.

Examining relationships and differences using statistics

These include:

- the data cases selected for the sample should be independent – in other words the selection of any one case for your sample should not affect the probability of any other case being included in the same sample;
- the data cases should be drawn from normally distributed populations;
- the populations from which the data cases are drawn should have equal variances (don't worry, the term variance will be explained later;
- the data used should be numerical.

Examining relationships and differences using statistics

In addition, you need to ensure that your sample size is sufficiently large to meet the requirements of the statistic you are using.

If the assumptions are not satisfied, it is often still possible to use non-parametric statistics. The way in which statistical significance is assessed using both non-parametric and parametric statistics can be thought of as answering one from a series of questions, dependent on the data type:

Examining relationships and differences using statistics

- Is the independence or association statistically significant?
- Are the differences statistically significant?
- What is the strength of the relationship and is it statistically significant?
- Are the predicted values statistically significant?

When assessing significance each question will usually be phrased as a hypothesis; that is a tentative, usually testable, explanation that there is an association, difference or relationship between two or more variables. The questions and associated statistics are summarized in table below along with statistics used to help examine trends.

	Categorical		Numerical	
	Nominal (Descriptive)	Ordinal (Ranked)	Continuous	Discrete
To test <i>normality</i> of a distribution			Kolmogorov–Smirnov test, Shapiro–Wilk test	
To test whether two variables are <i>independent</i>	Chi square (data may need grouping)		Chi square if variable grouped into discrete classes	
To test whether two variables are <i>associated</i>	Cramer’s V and Phi (both variables must be dichotomous)			
To test whether two groups (categories) are <i>different</i>		Kolmogorov–Smirnov (data may need grouping) or Mann–Whitney <i>U</i> test	Independent <i>t</i> -test or paired <i>t</i> -test (often used to test for changes over time) or Mann–Whitney <i>U</i> test (where data skewed or a small sample)	
To test whether three or more groups (categories) are <i>different</i>			Analysis of variance (ANOVA)	

	Categorical		Numerical	
	Nominal (Descriptive)	Ordinal (Ranked)	Continuous	Discrete
To assess the <i>strength of relationship</i> between two variables		Spearman's rank correlation coefficient (Spearman's rho) or Kendall's rank order correlation coefficient (Kendall's tau)	Pearson's product moment correlation coefficient (PMCC)	
To assess the <i>strength of a relationship</i> between one dependent and one independent variable			Coefficient of determination	
To assess the <i>strength of a relationship</i> between one dependent and two or more independent variables			Coefficient of multiple determination	
To <i>predict</i> the value of a dependent variable from one or more independent variables			Regression equation	

Examining relationships and differences using statistics

Testing for normality

As we have already noted, parametric tests assume that the numerical data cases in your sample are drawn from normally distributed populations. This means that the data values for each quantitative variable should also be normally distributed, being clustered around the variable's mean in a symmetrical pattern forming a bell-shaped frequency distribution.

Fortunately, it is relatively easy to check if data values for a particular variable are distributed normally, both using graphs and statistically.

Examining relationships and differences using statistics

A number of different types of graphs including histograms, box plots and frequency polygons. All of these can be used to assess visually whether the data values for a particular numerical variable are clustered around the mean in a symmetrical pattern, and so normally distributed.

For normally distributed data, the value of the mean, median and mode are also likely to be the same.

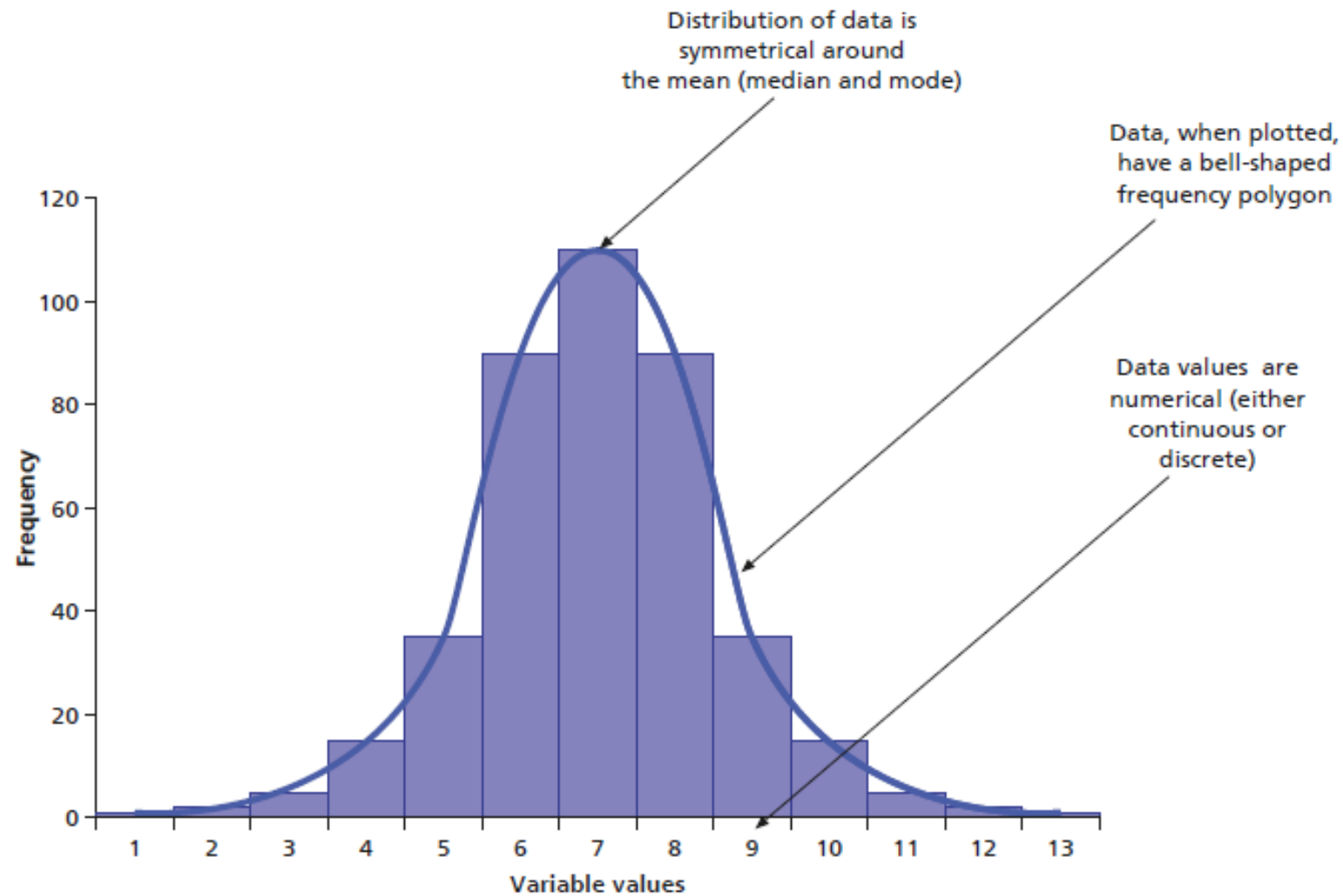


Figure 12.13 Annotated frequency polygon showing a normal distribution

Examining relationships and differences using statistics

Another way of testing for normality is to use statistics to establish whether the distribution as a whole for a variable differs significantly from a comparable normal distribution.

Fortunately, this is relatively easy to do in statistical software such as IBM SPSS Statistics using the **Kolmogorov–Smirnov test** and the **Shapiro–Wilk test**.

Chapter

08

Assessing the statistical significance of relationships and differences

Assessing the statistical significance of relationships and differences between variables usually involves hypothesis testing. As part of your research project, you might have collected sample data to examine the association between two variables.

You would have phrased this as a testable explanation that put forward the absence of that relationship (termed a null hypothesis) such as: ‘there is no association between. . . ’ Once you have entered data into the analysis software, chosen the statistic and clicked on the appropriate icon, an answer will appear as if by magic!

Assessing the statistical significance of relationships and differences

If the probability of your test statistic value or one more extreme having occurred is less than a prescribed significance value (usually $p < 0.05$ or lower), this is usually interpreted as casting doubt on or providing evidence against your null hypothesis and the associated underlying assumptions.

This means your data are more likely to support the explanation expressed in your hypothesis; in this example a testable statement such as: 'There is an association between. . . ' Statisticians refer to this as rejecting the null hypothesis and accepting the hypothesis, often abbreviating the terms null hypothesis to H_0 and hypothesis to H_1 .

Assessing the statistical significance of relationships and differences

Consequently, rejecting a null hypothesis could mean casting doubt on an explanation such as ‘there is no difference between . . . ’ or ‘there is no relationship between. . . ’ and accepting an explanation such as ‘there is a difference between . . . ’ or ‘there is a relationship between. . . ’

However, conclusions and policy decisions should not be based just on whether the p-value passes a specific threshold. Contextual factors such as the research design, quality of data, and other external evidence are also important in interpreting the findings.

Chapter

08

Assessing the statistical significance of relationships and differences

If the probability of obtaining the test statistic or one more extreme by chance alone is greater than or equal to a prescribed value (usually $p = 0.05$), this is normally interpreted as your data being compatible with the explanation expressed by your null hypothesis and its associated underlying assumptions. This indicates the null hypothesis can be accepted and is referred to by statisticians as failing to reject the null hypothesis. There may still be a relationship between the variables under such circumstances, but you cannot make the conclusion with any certainty.

Remember, when interpreting probabilities from software packages, beware: owing to statistical rounding of numbers a probability of 0.000 does not mean zero, but that it is less than 0.001.

Assessing the statistical significance of relationships and differences

The hypothesis and null hypothesis we have just stated are often termed **non-directional**. This is because they refer to a difference rather than also including the nature of the difference. A **directional hypothesis** includes within the testable statement the direction of the difference, for example 'larger'.

This is important when interpreting the probability of obtaining the test result, or one more extreme, by chance. Statistical software often states whether this probability is **one-tailed** or **two-tailed**. Where you have a directional hypothesis such as when the direction of the difference is larger, you should use the one-tailed probability. Where you have a non-directional hypothesis and are only interested in the difference, you should use the two-tailed probability.

Type I and Type II errors

Inevitably, errors can occur when making inferences from samples. Statisticians refer to these as Type I and Type II errors. The analogy of legal decisions is used to explain Type I and Type II errors. In this analogy Type I error is equated to a person who is innocent being unjustly convicted and a Type II error to a person who is guilty of a crime being unjustly acquitted.

In entrepreneurship research we would say that an error made by wrongly rejecting a null hypothesis and therefore accepting the hypothesis is a Type I error. Type I errors might involve you concluding that two variables are related when they are not, or incorrectly concluding that a sample statistic exceeds the value that would be expected by chance alone.

		Likelihood of making a	
		Type I error	Type II error
Significance level at	0.05	Increased	Decreased
	0.01	Decreased	Increased

Type I and Type II errors

Type I and Type II errors

This means you are rejecting your null hypothesis when you should not. The term ‘statistical significance’ discussed earlier therefore refers to the probability of making a Type I error.

A Type II error involves the opposite occurring. In other words, you fail to reject your null hypothesis when it should be rejected. This means that Type II errors might involve you in concluding that two variables are not related when they are, or that a sample statistic does not exceed the value that would be expected by chance alone.

Type I and Type II errors

Given that a Type II error is the inverse of a Type I error, it follows that if we reduce our likelihood of making a Type I error by setting the significance level to 0.01 rather than 0.05, we increase our likelihood of making a Type II error by a corresponding amount.

This is not an insurmountable problem, as researchers usually consider Type I errors more serious and prefer to take a small likelihood of saying something is true when it is not. It is therefore generally more important to minimize Type I than Type II errors.

To test whether two variables are independent or associated

Often descriptive or numerical data will be summarized as categorical data using a two way contingency table. The **chi square test (x2)** enables you to find out how likely it is that the two variables are independent. It is based on a comparison of the observed values in the table with what might be expected if the two distributions were entirely independent.

Therefore you are assessing the likelihood of the data in your table, or data more extreme, occurring by chance alone by comparing it with what you would expect if the two variables were independent of each other. This could be phrased as the null hypothesis: 'there is no dependence ... '.

To test whether two variables are independent or associated

Some software packages, such as IBM SPSS Statistics, calculate the statistic Cramer's V alongside the chi square statistic. If you include the value of **Cramer's V** in your research report, it is usual to do so in addition to the chi square statistic. Whereas the chi square statistic gives the probability of data in a table, or data more extreme, occurring by chance alone, Cramer's V measures the association between the two variables within the table on a scale where 0 represents no association and 1 represents perfect association.

Because the value of Cramer's V is always between 0 and 1, the relative strengths of associations between different pairs of variables that are considered statistically significant can be compared..

Chapter

08

To test whether two groups are different

Numerical data

If a numerical variable can be divided into two distinct groups using a descriptive variable, you can assess the likelihood of these groups being different using an **independent groups t-test**. This compares the difference in the means of the two groups using a measure of the spread of the scores.

If the likelihood of an observed difference or one greater between these two groups occurring by chance alone is low, this is represented by a large t statistic with a low probability (p -value).

Chapter

08

To test whether two groups are different

A p-value of 0.05 or less is usually termed statistically significant. Alternatively, you might have numerical data for two variables that measure the same feature but under different conditions. Your research could focus on the effects of an intervention such as employee counselling. As a consequence, you would have pairs of data that measure work performance before and after counselling for each case.

To assess the likelihood of any difference or one greater between your two variables (each half of the pair) occurring by chance alone, you would use a paired t-test. Although the calculation of this is slightly different, your interpretation would be the same as for the independent groups t-test.

To test whether two groups are different

The t-test assumes that the data are normally distributed and this can be ignored without too many problems for sufficiently large samples, this often being defined as less than 100 and by some as less than 30.

The assumption that the data for the two groups have the same variance (standard deviation squared) can also be ignored provided that the two samples are of similar size.

To test whether two groups are different

If the data are skewed or the sample size is small, the most appropriate statistical test is the **Mann–Whitney U Test**. This test is the non-parametric equivalent of the independent groups t-test.

Consequently, if the likelihood of a difference or one greater between these two groups occurring by chance alone is low, this will be represented by a large U statistic with a probability less than 0.05. This is termed statistically significant..

To test whether three or more groups are different

If a numerical variable is divided into three or more distinct groups using a descriptive variable, you can assess the likelihood of these groups being different occurring by chance alone by using **one-way analysis of variance or one-way ANOVA**.

As you can gather from its name, ANOVA analyses the variance, that is, the spread of data values, within and between groups of data by comparing means. The **F ratio** or **F statistic** represents these differences. If the likelihood of the observed difference or one greater between groups occurring by chance alone is low, this will be represented by a large F ratio with a probability of less than 0.05.

To test whether three or more groups are different

This is usually considered statistically significant. The following assumptions need to be met before using one-way ANOVA. More detailed discussion is available.

- Each data value is independent and does not relate to any of the other data values. This means that you should not use one-way ANOVA where data values are related in some way, such as the same case being tested repeatedly.
- The data for each group are normally distributed. This assumption is not particularly important provided that the number of cases in each group is large (30 or more).
- The data for each group have the same variance (standard deviation squared).

However, provided that the number of cases in the largest group is not more than 1.5 times that of the smallest group, this appears to have very little effect on the test results.

Assessing the strength of relationship

If your data set contains ranked or numerical data, it is likely that, as part of your Exploratory Data Analysis, you will already have plotted the relationship between cases for these ranked or numerical variables using a scatter graph. Such relationships might include those between weekly sales of a new product and those of a similar established product, or age of employees and their length of service with the company.

These examples emphasize the fact that your data can contain two sorts of relationship:

- those where a change in one variable is accompanied by a change in another variable but it is not clear which variable caused the other to change, a correlation;
- those where a change in one or more (independent) variables causes a change in another (dependent) variable, a cause-and-effect relationship.

To assess the strength of relationship between pairs of variables

A **correlation coefficient** enables you to quantify the strength of the linear relationship between two ranked or numerical variables. This coefficient (usually represented by the letter r) can take on any value between $+1$ and -1 . A value of $+1$ represents a perfect positive correlation.

This means that the two variables are precisely related and that as values of one variable increase, values of the other variable will increase. By contrast, a value of -1 represents a perfect negative correlation. Again, this means that the two variables are precisely related; however, as the values of one variable increase those of the other decrease. Correlation coefficients between $+1$ and -1 represent weaker positive and negative correlations, a value of 0 meaning the variables are perfectly independent. Within business research it is extremely unusual to obtain perfect correlations.

To assess the strength of relationship between pairs of variables

For data collected from a sample you will need to know the probability of your correlation coefficient or one more extreme (larger) having occurred by chance alone. Most analysis software calculates this probability automatically.

As outlined earlier, if this probability is very low (usually less than 0.05) then the relationship is usually considered statistically significant. In effect you are rejecting the null hypothesis, that is a statement such as: “there is no correlation between. . .” and accepting a hypothesis such as: “there is a correlation between. . .” If the probability is greater than 0.05 then your relationship is usually considered not statistically significant.

To assess the strength of relationship between pairs of variables

If both your variables contain numerical data you should use **Pearson's product moment correlation coefficient (PMCC)** to assess the strength of relationship. Where these data are from a sample then the sample should have been selected at random and the data should be normally distributed.

However, if one or both of your variables contain ranked data you cannot use PMCC, but will need to use a correlation coefficient that is calculated using ranked data. Such rank correlation coefficients represent the degree of agreement between the two sets of rankings. Before calculating the rank correlation coefficient, you will need to ensure that the data for both variables are ranked. Where one of the variables is numerical this will necessitate converting these data to ranked data. Subsequently, you have a choice of rank correlation coefficients

To assess the strength of relationship between pairs of variables

The two used most widely in entrepreneurship research are **Spearman's rank correlation coefficient** (Spearman's r , the Greek letter rho) and **Kendall's rank correlation coefficient** (Kendall's t , the Greek letter tau). Where data are being used from a sample, both these rank correlation coefficients assume that the sample is selected at random and the data are ranked (ordinal).

Given this, it is not surprising that whenever you can use Spearman's rank correlation coefficient you can also use Kendall's rank correlation coefficient. However, if your data for a variable contain tied ranks, Kendall's rank correlation coefficient is generally considered to be the more appropriate of these coefficients to use. Although each of the correlation coefficients discussed uses a different formula in its calculation, the resulting coefficient is interpreted in the same way as PMCC.

Chapter

08

To assess the strength of a cause-and-effect relationship between dependent and independent variables

In contrast to the correlation coefficient, the **coefficient of determination** enables you to assess the strength of relationship between a numerical dependent variable and one numerical independent variable and the **coefficient of multiple determination** enables you to assess the strength of relationship between a numerical dependent variable and two or more independent variables.

Once again, where these data have been selected from a sample, the sample must have been selected at random. For a dependent variable and one (or perhaps two) independent variables you will have probably already plotted this relationship on a scatter graph. If you have more than two independent variables this is unlikely as it is very difficult to represent four or more scatter graph axes visually!

Chapter

08

To assess the strength of a cause-and-effect relationship between dependent and independent variables

The coefficient of determination (represented by r^2) and the coefficient of multiple determination (represented by R^2) can both take on any value between 0 and +1. They measure the proportion of the variation in a dependent variable (amount of sales) that can be explained statistically by the independent variable (marketing expenditure) or variables (marketing expenditure, number of sales staff, etc.).

This means that if all the variation in amount of sales can be explained by the marketing expenditure and the number of sales staff, the coefficient of multiple determination will be 1. If 50 per cent of the variation can be explained, the coefficient of multiple determination will be 0.5, and if none of the variation can be explained, the coefficient will be 0.

Chapter

08

To assess the strength of a cause-and-effect relationship between dependent and independent variables

The coefficient of determination (represented by r^2) and the coefficient of multiple determination (represented by R^2) can both take on any value between 0 and +1. They measure the proportion of the variation in a dependent variable (amount of sales) that can be explained statistically by the independent variable (marketing expenditure) or variables (marketing expenditure, number of sales staff, etc.).

This means that if all the variation in amount of sales can be explained by the marketing expenditure and the number of sales staff, the coefficient of multiple determination will be 1. If 50 per cent of the variation can be explained, the coefficient of multiple determination will be 0.5, and if none of the variation can be explained, the coefficient will be 0.

Chapter

08

To predict the value of a variable from one or more other variables

Regression analysis can also be used to predict the values of a dependent variable given the values of one or more independent variables by calculating a regression equation. You may wish to predict the amount of sales for a specified marketing expenditure and number of sales staff. You would represent this as a regression equation:

This equation can be translated as stating:

$$\text{Amount of sales}_i = \text{value} + (\beta_1 * \text{Marketing expenditure}_i) + (\beta_2 * \text{Number of sales staff}_i)$$

To predict the value of a variable from one or more other variables

Using regression analysis you would calculate the values of the constant coefficient a and the slope coefficients b_1 and b_2 from data you had already collected on amount of sales, marketing expenditure and number of sales staff.

A specified marketing expenditure and number of sales staff could then be substituted into the regression equation to predict the amount of sales that would be generated. When calculating a regression equation you need to ensure the following assumptions are met:

Chapter

08

To predict the value of a variable from one or more other variables

The relationship between dependent and independent variables is linear. **Linearity** refers to the degree to which the change in the dependent variable is related to the change in the independent variables.

Linearity can easily be examined through residual plots (these are usually drawn by the analysis software). Two things may influence the linearity. First, individual cases with extreme values on one or more variables (outliers) may violate the assumption of linearity.

It is, therefore, important to identify these outliers and, if appropriate, exclude them from the analysis. Second, the values for one or more variables may violate the assumption of linearity. For these variables the data values may need to be transformed.

Chapter

08

To predict the value of a variable from one or more other variables

The extent to which the data values for the dependent and independent variables have equal variances, also known as **homoscedasticity**. Again, analysis software usually contains statistical tests for equal variance. For example, the **Levene test** for homogeneity of variance measures.

The equality of variances for a single pair of variables. If **heteroscedasticity** (that is, unequal variances) exists, it may still be possible to carry out your analysis. Further details of this can again be found in more specialized books on multivariate analysis, such as Hair et al. (2014).

To predict the value of a variable from one or more other variables

Absence of correlation between two or more independent variables (**collinearity** or **multicollinearity**), as this makes it difficult to determine the separate effects of individual variables. The simplest diagnostic is to use the correlation coefficients, extreme collinearity being represented by a correlation coefficient of 1.

The rule of thumb is that the presence of high correlations (generally 0.90 and above) indicates substantial collinearity. Other common measures include the tolerance value and its inverse – the **variance inflation factor (VIF)**. Hair et al. (2014) recommend that a very small tolerance value (0.10 or below) or a large VIF value (10 or above) indicates high collinearity.



مجموعة جامعة
الجميع الذكية
للتعليم القابضة

MIDOCEAN
UNIVERSITY
BY EVERYONE'S SMART UNIVERSITY



Thank You

