TOPIC 1

Business Analytics in Organisations

Contents

Introduction

The economic environment is becoming increasingly competitive due to global marketing, the electronic age and the explosion in the amount of information available to us.

Increased competition means that modern managers need, more than ever before, to run highly efficient organisations and to focus on the quality of products and services that they offer their clients. Failure in either of these two areas—the internal operations and the end-user market—can mean failure of the organisation altogether.

To achieve these goals, a basic requirement is to understand the organisation and the market. And the key to this understanding is data.

Data (the plural word for datum) can be described as facts and figures which have meaning. The numbers '315,700', '21,017,200', '1.5', etc. are not data as such, but they are data if we say the following:

> In the 12 months to 30 June 2007, the Australian population increased by 315,700 people, reaching 21,017,200. The annual growth rate for the year ended 30 June 2007 (1.5%) was slightly higher than that recorded for the year ended 30 June 2006.
>
> (Australian Bureau of Statistics (ABS) 2007, (3201.0) 'Population by Age and Sex, Australian States and Territories', released 12 December 2007 <http://www.abs.gov.au> retrieved 5 August 2010.

Similarly, 'female', '27', 'yes' and '8' don't tell us anything by themselves. They do tell us something, however, if we know that (from a survey of our staff) one female respondent aged 27 said she was happy with her current boss and she gave a score of 8 for her own personal job satisfaction (on a scale of 0 = lowest to 10 = highest).

Data are the basic requirements for decision making, and managers must have the ability to collect, organise and interpret data in order to arrive at rational decisions regarding the efficiency and effectiveness of an organisation.

Increasingly organisations are collecting or have access to larger and larger volumes of data through point of sale systems, website traffic and social media, cheap instrumentation such as RFID tags and so on. This is now commonly referred to as 'Big Data' and is proving challenging for many organisations to capture and analyse effectively and in a timely fashion.

Decision making—business or otherwise—comes in different guises. It could be helping to make a *choice* (choosing between three different locations for a new warehouse), solving a *problem* (why productivity in one plant in our organisation is noticeably lower than in another) or investigating an *opportunity* (seeking out new markets or products). Underpinning good decision making is day-to-day monitoring of an organisation and its various processes (tables, graphs and summary measures on sales, costs, market share, competitors' pricing, etc.).

Thus, data and data analysis are keys to successful decision making. The importance extends to business analytics which used data, information technology, statistical analysis, and mathematical or computer-based models to gain improved insights. Common applications include areas like customer relationship management, supply chain optimisation, and pricing decisions.

Central to business analytics is Statistics—as a source of information and as a field of study—which is also used by all of us every day.

In the sporting arena, the Monday newspapers are full of statistics about all the details of participants in the weekend's cricket, football, racing, etc. Newspapers and TV news programs on any day of the week carry numerous articles involving 'surveys', 'research', 'latest findings', 'recent trends', etc.

As a student in this course, you are—probably unconsciously—using statistical techniques by choosing this course since 'in all probability' it will 'increase your promotion and future prospects'.

The route and means you choose to go to work or campus each day is a subconscious use of statistics in terms of evaluating the quickest, safest and most enjoyable option.

Your university is constantly updating and monitoring student numbers, student characteristics, their geographical area and student results (the percentages of passes and credits, etc.)

At work you would receive and use statistical data daily, yet might not realise it. Examples might include:

- economic data (e.g. inflation, GDP, interest rate forecasts)
- finance data (e.g. share prices, credit risk, costing analysis)
- marketing data (e.g. market-share analysis)
- human resource data (e.g. productivity of various sections, labour union trends)
- corporate environment data (e.g. ecological impact studies, operations data such as defect rates and productivity).

Thus, statistics (or data analysis) provides information necessary to make effective decisions. There is no point in collecting, processing and analysing data if there is no real purpose to the exercise.

In the past, *statistics* as a subject to study has been treated with fear and concern. However, modern courses need to, and should not, be treated with any fear. You will find this course is:

- non-mathematical (with little requirement to derive, remember, or even understand formulae)
- computer based (with the computer doing most of the work for you)
- data focused (using data sets to explain and understand statistical concepts).

Thus, some advice from the beginning of the course: concentrate on the 'big picture' not the 'detail':

- Don't become anxious about the mathematics and detailed workings.
- For each statistical technique you come across, determine how it can be of use as a decision-making tool (the 'big picture').

Keeping this in mind, be assured this unit is *not* written for professional statisticians. It is tailored to provide you with a detailed understanding of data analysis techniques, in order that it may assist you as a manager in your decision-making capacity.

Objectives

At the completion of this topic you should be able to:

- explain key terms such as business analytics, big data and statistics
- identify the main sources of data available to assist in statistical decision making
- distinguish between different data types and formats

Introduction to business analytics, big data and statistics

Business Analytics is a relatively new discipline that involves the use of a variety of tools and techniques to achieve an understanding of the large amounts and variety of data that many businesses collect or have access to. The goal of business analytics is to ultimately aid in making better (data driven) business decisions and realising the value of an organisation's data.

Business Analytics begins with the collection, organisation and manipulation of data and is supported by three major components:

Descriptive Analytics—uses data to understand past and present performance and make informed decisions. It is arguably the most commonly used and well understood type of analytics. The techniques involved use fundamental tools and methods of data analysis and statistics, focusing on:

- Descriptive statistical measures and Data visualisation
- Probability distributions / Sampling and estimation
- Statistical inference

Predictive Analytics—analyses past performance in an effort to predict the future by examining historical data, detecting patterns or relationships in these data, and then extrapolating these relationships forward in time. Techniques include:

- Regression

- Forecasting

Prescriptive Analytics—also referred to as 'Decision Analytics'. Uses optimization to identify the best alternative to minimize or maximize some objective. Techniques include linear programming and other advanced modelling techniques.

Although the tools used in descriptive, predictive and prescriptive analytics are different, many applications involve all three.

In MSQ791, the focus is on descriptive analytics and the statistical techniques that underpin it.

READING

Read Evens J R Business Analytics 2013, pp. 2–7.

Statistics is a large discipline that comprises three broad tasks. These are:

1 collection of data

2 processing and presentation of data

3 analysis and interpretation of data.

The tasks, in fact, are interdependent and largely overlap. For example, sometimes we are able to process data at the same time as we are collecting it, and often we analyse and interpret the results at the same time as processing data (for example, while generating tables and graphs, or while performing a test).

The overall objective is to make valid conclusions about the characteristics of the sources from which the data were obtained. For managers, statistics is a tool to be used in conjunction with other fields to aid in decision making. The challenge of this *Data Analysis for Managers* unit is to demonstrate why statistics is an essential tool in the decision-making process, and to understand how it can be utilised.

Managers need an understanding of statistics for the following four key reasons:

1 to properly present and describe business data and information

2 to draw conclusions about large populations based solely on information collected from samples

3 to make reliable forecasts about business trends

4 to improve business processes.

Descriptive and inferential statistics

The reason for the existence of statistics as a discipline is the concept of 'variation'. Variation exists in countless instances: why don't all cars sell for the same price? Why don't all employees work to the same level of productivity? Why don't all students score the same mark in exams? Why don't all seedlings survive, and why don't those that do survive grow to exactly the same height and shape?

Statistics, then, could be called the 'study of variation'.

As such, as a field of study, statistics can be split into two main groups:

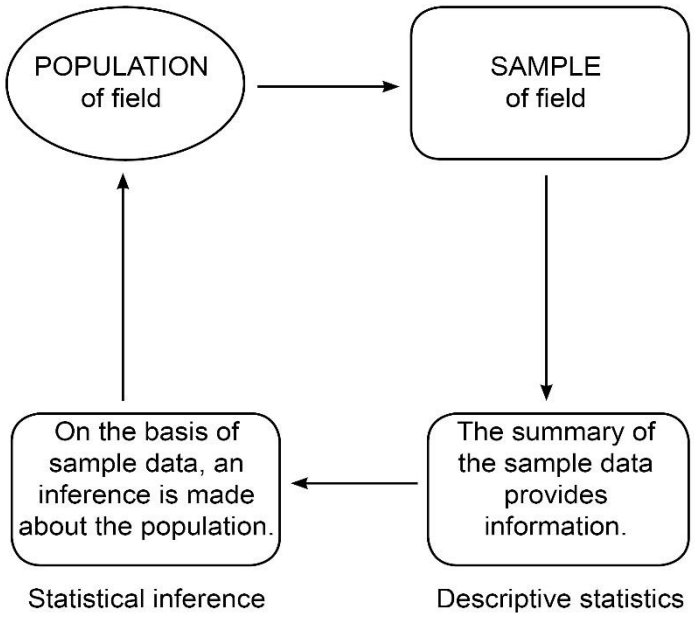
1 *Descriptive statistics*, which relates to a set of techniques based around certain tables, graphs and calculated summary measures used for describing the important features of a given set of data.

2 *Inferential statistics*, which relates to the use of sample data to draw inferences and conclusions about the whole population of individuals or items from which the sample was drawn.

We will use descriptive and/or inferential techniques to analyse variation throughout much of this unit.

The process of statistical inference is illustrated in Exhibit 1.1.

Exhibit 1.1: The process of statistical inference



Descriptive techniques also play a key part in inferential statistics. Thus, when a sample is taken, the data are first thoroughly investigated using descriptive

measures (tables, graphs and summary measures). From this summarised form we apply statistical inference techniques to allow for potential error from the sample. The end result is inferences about the population, which in turn will assist us in decision making.

To understand inferential statistics there are several key terms that must be defined:

- A *population* is the whole group of items or individuals about which we wish to draw conclusions (for example, the coffee manufacturer wishes to draw conclusions about *all* Australian households).
- If we can investigate every member of a population we often say we conduct a *census*.
- Frequently, however, we can take only a *sample*. A sample means that only a fraction or proportion of the population is included in the investigation (for example, a random sample of 1000 households).

EXERCISE 1.1

Consider the following situations:

1 A suburban printery has 15 employees. The owner of the business would like to obtain the views of the employees relating to changes in a number of workplace practices.

2 The circulation manager of a large metropolitan daily newspaper wishes to ascertain readers' views on a number of planned changes to the newspaper.

In each case:

(a) explain whether a sample or a census would be appropriate

(b) explain whether descriptive statistics or inferential statistics would be appropriate

(c) explain why such a survey may be needed.

Data

Almost everyone deals with data. These include chief executives, accountants, economists, marketing representatives, social scientists, chemists, occupational hygienists, consumers and managers. Data could be in a multitude of forms, including quarterly sales figures, expenditures for goods and services, efficiency rates on a production line, customer interactions on social media, web based traffic, census figures, contamination levels associated with toxic material, and so on.

Big Data

Organisations have always collected and used data to undertake various business functions. What has changed over the years is the sheer magnitude

of data, the variety of sources of data as well as the format and type of data that is now available. Improvements in technology and instrumentation have made data more readily available and very cheap. Use of the world wide web and social media by organisations have also resulted in an explosion of unstructured data such as comments made about a product on Facebook.

READING

Read Zikopoulos et al. Harness the Power of Big Data 2013, pp. 9–15.

Types of data

In this section we study different ways of classifying data. Let us look at one question we may have asked while collecting data:

- How many hours do you work?
- Are you currently working full time?

If we examine these questions, we see that the answers will supply us with two different types of data:

- numerical
- categorical.

The first question will provide us with a number, e.g. 37.25 hours (i.e. *numerical data*) For the second question, the answer is either 'yes' or 'no' (i.e. *categorical data*). **The analytical techniques used will depend upon the type of data.**

Understanding which data type you are working with in statistics is very important in a whole range of ways.

Numerical data v categorical data

The following data variables can be classified as being 'numerical' or 'categorical':

- Numerical, or quantitative, variables are already in numerical form. A car's 'price', 'number of seats' and 'kilometres travelled' are examples.
- Categorical, or qualitative, variables are not in numerical form. A car's 'colour', 'style' and 'make' are examples.

Numerical variables: Discrete v continuous

Numerical (quantitative) variables, themselves, can be split into two broad groups:

- *Discrete variables*—ones that can take on only limited values in a given range (generally values limited to whole numbers). For example, the number of brothers and sisters you have can take on only the values 0, 1, 2, 3, etc.

- *Continuous variables*—ones that can take on any value in a given range (values capable of including decimal places). For example, in theory, when measuring people's heights, any value is possible in the range 160 cm to 170 cm, say.

The distinction between discrete and continuous variables becomes important in a number of areas; in particular, when constructing the classes for frequency distributions and for graphical purposes.

We often use variables which are strictly discrete but which are best treated as being continuous, that is, they are *effectively continuous*. For example, people's weekly incomes are strictly a discrete variable since incomes can only be measured to the nearest cent. However, for most practical purposes this is more than the degree of accuracy required and thus we treat weekly incomes as being effectively continuous.

Levels of measurement and scaling

From topic 1, recall levels of measurement and types of measurement scales. We noted the following different scales:

Categorical data

- nominal—data are labels or names used to identify an attribute of the entity.
- ordinal—data have the properties of nominal data and the order or rank of the data is meaningful.

Numerical data

- interval—data have the properties of ordinal data, and the interval between observations is expressed in terms of a fixed unit of measure.
- ratio—data have all the properties of interval data and the ratio of two values is meaningful.

Cross-sectional data v time-series data

Another way of classifying data is as *cross-sectional* data or *time-ordered* (or *time-series*) data.

Cross-sectional data relates to a group of items or individuals at a given point of time. For example, the Australian Bureau of Statistics may take a random sample of households on 31 May 2012 to gather data on household expenditure. The principal reason for analysing cross-sectional data is to investigate patterns and relationships to help us better understand the whole group (the population) from which the data were taken.

Time-ordered (or time-series) data relate to a particular entity or situation at different points of time. For example, we may have data on the sales of a firm for the last 10 years. The primary reason for analysing time-series data is for forecasting, or predicting what might happen. By analysing the history of a

time-series variable, we can identify past trends and patterns which may indicate what will happen in the future.

In this subject we mainly examine cross-sectional data. Generally, these two broad types of numerical data need to be analysed in quite different ways. The main reason is that in a time series there may be trends—seasonal and other influences—which render cross-sectional analysis meaningless, and even dangerous. Thus, time-series data require special statistical tools of their own.

Univariate and multivariate data

Another helpful classification of cross-sectional data is as 'univariate' and 'multivariate' data sets. We use these terms to describe the number of variables in a data set.

Univariate means that there is only one variable of interest. A *multivariate* data set means two or more variables are given (sometimes the term bivariate is used for a two variable data set).

Even if a data set is multivariate, we often analyse just a single variable at a time. That is, we analyse it independently of the other variables in the data set. That way we can explore the patterns for that particular variable (we will investigate these patterns—such as central tendency, dispersion and shape—later in this topic and the next). Part of your computer exercises with the Conrobar data set in this and other topics will be to thoroughly investigate certain variables on a single variable basis—in particular, productivity, days absent and job satisfaction.

If something of interest is discovered with a single variable, we may need to find out more. For example, we may find that the productivity level of workers is below the company's objective of 100%. Then we may begin to introduce other variables in order to help explain why this has occurred: does productivity vary by gender, by age, etc.?

Generally, in a multivariate set we have a key variable in which we are interested. For example, we may be interested in the 'price' of second-hand cars: Why don't second-hand cars all sell for the same price? What are some of the factors influencing selling price? Or, we may be interested in the productivity performance of workers in a firm: Why don't they all meet the goal of 100%? Does overtime or salary have any predictive powers over productivity?

This key variable is often called the 'dependent' variable. Normally, we will start by looking at this variable just by itself: understanding the degree of variation, investigating central (or typical) values, and looking for any shapes that help us understand the variable.

Once we have a good appreciation of our key variable, we can then introduce other variables which may help explain variation in that key variable.

Summary

The topic began with a broad overview of the subject of business analytics and statistics and its applications.

The collection, presentation and characterisation of information were introduced as modern statistics' key features, which assist in the decision-making process.

The distinction was made between descriptive statistics and inferential statistics, the latter being the use of sample statistics from random samples to draw conclusions about unknown population parameters. It is important to appreciate the distinction between descriptive statistics and inferential statistics. However, it is also important to appreciate that the two supplement each other.

In data collection and presentation the emphasis was on the importance of obtaining good data. Classifying the data is important in both understanding the data but also in deciding what analytical techniques are appropriate.

Further resources

Berenson, ML, Levine, DM, Krehbiel, TC, Watson J, Jayne, N, Turner, L & O'Brien, M 2010, *Basic business statistics: concepts and applications* (Australasian and Pacific edition) Pearson Education Australia, Frenchs Forest.

Black, K 2008, *Business statistics for contemporary decision making*, 5th edn, Wiley, NJ.

Anderson, DR, Sweeney, DJ & Williams, TA 2008, *Statistics for business and economics*, 10th edn, South-Western Thomson Learning, Cincinnati, Ohio.

Selvanathan, A, Selvanathan, S, Keller, G & Warrack, B 2006, *Australian business statistics*, 4th edn, Nelson Thomson Learning, Melbourne.

TOPIC 2

Summarising business data: Summary measures

Contents

Introduction

The process of statistical decision making involves three broad but highly interdependent stages, which are:

1 collecting data for a particular purpose

2 processing the data to produce suitable summary measures, tables and graphs

3 analysing and interpreting the results in regard to the final decision that has to be made.

In topics 2 – 4 we investigate ways in which data, once collected, can be processed and used for analysis and interpretation, for the eventual goal of decision making. The methods covered provide ways of presenting data for you and others to use. We call these *descriptive* or *exploratory* tools. They fall under three main types:

1 *Tables*—these include the array, the frequency distribution and the cross-tab
(topics 3 and 4)

2 *Graphs*—these include histograms, bar charts, dot plots, and scatter diagrams
(topics 3 and 4)

3 *Summary measures*—these are calculations such as the *mean*, *median*, *mode*, *range*, etc that, with single values, describe characteristics of a data set (this topic)

Each type of table, graph and summary measure has its advantages and disadvantages. We will find that what might be suitable for one purpose may not be suitable for another. We will also find that we often need to use a combination of descriptive or exploratory tools to help us fully understand the data.

In summary, over the next three topics, we will be investigating ways of describing and exploring data to help us better understand a given variable and to help explain variation in that variable. Because of the power of computers, which enable us to quickly generate a large range of tables, graphs and calculations, exploring data is one of the most interesting and challenging aspects of modern statistics.

Objectives

At the completion of this topic you should be able to:

- distinguish between, and create, a wide range of statistical measures which can be used for summarising and exploring data

- identify when each descriptive tool should be used
- perform some *basic* manual calculations of some summary measures
- use computer software for calculating summary measures (after completing the corresponding tutorial)
- describe comprehensively the main features of a data set

Introduction to summary measures

The aim of a summary measure is to summarise—in a single figure—a particular feature of *numerical* data. Note that categorical data is not considered in this topic, but will instead be covered by techniques in the next two topics.

While summary measures attempt to give precision to our findings, two points should be made:

- It is generally not wise to use a single summary measure by itself for a particular purpose. Thus, always use a range of summary measures in conjunction with each other before drawing important conclusions.
- It is generally not wise to use summary measures without also resorting to tabular and graphical methods (see topic 3). Thus, always use a range of summary measures, tables and graphs when attempting to explore and describe a particular set of data.

Sample data v census data

Before generating any summary measures, and indeed before carrying out any other analysis, we should always note whether our data are from a *census* (of the entire population) or from a *sample* (a fraction of the population).

The source of data—from a census/whole population or from a sample—have important consequences for our calculations, our interpretation and our use of the data. You will see this especially from topic 5 onwards.

To ensure the distinction is made, statisticians use different terms and notation as follows:

- A summary measure computed from a sample is called a *statistic*.

  A summary measure computed from an entire population (census) is called a *parameter*.
- Greek letters are often used to denote population parameters. English (Roman) letters are generally used to denote sample statistics. The different notation means we instantly know the source of the data.

For example, you will see that the symbol we use for the sample mean is $\bar{X}$, while the symbol we use for the population mean is μ (mu). The symbol we use for the sample standard deviation is *s*, while for the population standard deviation we use σ (sigma). Therefore, if we see statistical results involving $\bar{X}$

and *s*, we know the data are sample data, and therefore generalisations may need to be made about the unknown population parameters, μ and σ, respectively.

Features of data

There are five broad patterns or features we look for in numerical data. These are:

Central tendency

Central tendency means the tendency for data to cluster about certain 'central' values. It is usually possible to detect some central value (or typical or average value) of the data around which all the data cluster. We use '*averages*' as indications of central tendency. The three averages you will come across are the *mean* (or *arithmetic mean* or *common average*), the *median* and the *mode*.

Location

We are also interested in *location* of the data. For example, below which values lie the bottom 50%, 25%, 95%, etc? Or above which values lie the top 25% or top 5%? We use summary measures called the quartiles, deciles and percentiles to examine these features. The measures of central tendency are also considered to be measures of location: the median provides the exact location of the middle value in an array; the mean will generally fall towards the centre of the data as well; the mode is less predictable but, generally, it too falls towards the middle of the data.

Variation

Variation is often referred to as dispersion or spread. Basically this means how spread out are the data around the central value(s). You will see we use summary measures such as the range, interquartile range, standard deviation and coefficient of variation to examine this feature.

Shape

The third important pattern we look for in a set of data is *shape*. Data can present itself in a number of different shapes or patterns. We best identify shape when data are graphed (as a dot plot, box plot or histogram – see topic 3), but we can also use summary measures to help examine shape.

The above four concepts relate to statistical features of the data. However, before using a particular data set, we need to investigate potential problems. Outliers are thus the fifth feature we investigate.

Outliers

An outlier can be thought of as being a value that is quite apart from the rest of the data or one that is separated from the rest of the data. It is important to detect them since they may involve an error in the data (someone made a mistake keying in the data or in responding to the data) or may indicate an extraordinary, but legitimate, occurrence. In this topic we introduce mathematical and graphical methods of detecting potential outliers.

READING

Read Black (2010), Business statistics for contemporary decision making

Note that you are not expected to perform many of the manual calculations shown in the reading (apart from some of the more straight-forward calculations). However, having a basic understanding of the underlying formulas does give you a better understanding of the measures and how they are interpreted.

Measures of central tendency

READING

Read Black et al (2010), Australasian Business Statistics, 2nd Edition, pp. 52–58.

Note that you are not expected to perform many of the manual calculations shown in the reading (apart from some of the more straight-forward calculations). However, having a basic understanding of the underlying formulas does give you a better understanding of the measures and how they are interpreted.

There are three key measures of central tendency for numerical data:

- Mean (arithmetic mean)
- Median
- Mode.

These are also known as *averages* since in some way or other they typify the data in a single figure. The *mean* is often referred to as '*the average*'. Thus, in this unit, when we talk about '*the average*' we are generally referring to the *mean*. However, you will see later in this topic that the mean is not always a good figure to use as a typical value, and hence we may prefer to use the median or mode as a better indication of an *average* value.

Measures of location

READING

Read Black et al (2010), Australasian Business Statistics, 2nd Edition, pp. 59–62.

The averages are by default measures of location, since they give an indication of the location of the *middle* or *centre* of the data.

However, there are more specific measures of location, all of which can be found in a dataset:

- Minimum and maximum values
- First and third quartiles
- Deciles
- Percentiles.

The median is a specific measure of location since it gives the location of the bottom and top halves of the data: it is equivalent to the *second quartile*. The *first and third quartiles* are also highly useful measures as they break the data into quarters. D*eciles* (which break the data into tenths) and *percentiles* (which break the data into hundredths). Clearly these last two terms are very useful: we often talk about a baby's weight or someone's income or height being at a particular decile or percentile.

Please note that there are varying formulae for calculating percentiles. The reading follows one approach. Below is an alternative way of calculating percentiles (you can use either technique. In most cases you will get the same or similar answers):

- Rank the numbers into ascending order.
- Find i = (p/100)×(n+1) where p is the required percentile, and n is the number of observations.
- If i is an integer, the pth percentile is the value in the i th position.
- If i is a fractional half, the pth percentile is the average of the values either side of the ith position.
- If i is neither an integer nor a factional half, round the result to the nearest integer, the pth percentile is the value in the ith position.

For example to determine Q_1 from a sample of 12 exam marks, we firstly arrange them in ascending order.

36 39 62 67 70 74 77 79 81 87 89 91

To compute Q_1, we find the 25^{th} percentile. Using the formula this equals

$$i=(p/100) \times(n+1)= (25/100)*13 = 3.25$$

Rounding i to the nearest integer gives 3. Q_1, is thus the value in the third position which equals 62.

EXERCISE 2.1

Using the above data set of 12 exam marks:

1 Calculate the mean, median and mode.

2 Calculate the third quartile.

3 What exam mark would you need to obtain to be in the top 10%?

Measures of variation

READING

Read Black et al (2010), Australasian Business Statistics, 2nd Edition, pp. 63–68, p.76.

Note that you are not expected to perform any of the manual calculations shown in the reading related to the standard deviation or variance. Skim read the material to obtain a basic understanding of the underlying formulas. This will give you a better understanding of the measures and how they are interpreted.

The measures of variation studied can be distinguished as follows:

- The *minimum* and *maximum* are two values that can be used to provide a good starting point to analysing overall variability.
- The *range* is a distance measure of variation. It measures the total distance between the minimum and maximum values.
- The *interquartile range* is also a distance measure of variation, and specifically measures the distance between the two quartiles. It can be called the *midspread* as it gives the location or spread of the middle 50% of values.
- The *standard deviation* is an average measure of variation. While the formulae may be difficult to understand and the technical definition ('the square root of the variance') may be less than enlightening, it does have a simple intuitive meaning. Think of the standard deviation as an '*estimate of the average distance that individual values are away from the mean*'. The standard deviation is the most important measure of variation you will work with in this unit. Hence, it is important to try and get a feel for what it attempts to measure.
- The *coefficient of variation* is a relative measure of variation. The range, interquartile range and standard deviation all measure variation in absolute terms (that is, they are measured in the same units as the original data). The coefficient of variation takes into account the magnitude of the data being analysed, and is the means of comparing variability between two or more data sets in quite different magnitudes or in different units.

EXERCISE 2.2

Using the above data set of 12 exam marks from Exercise 2.1:

1 Calculate the range and IQR

2 Interpret both values. Which value provides a better summary of the variation in the data set?

Shape

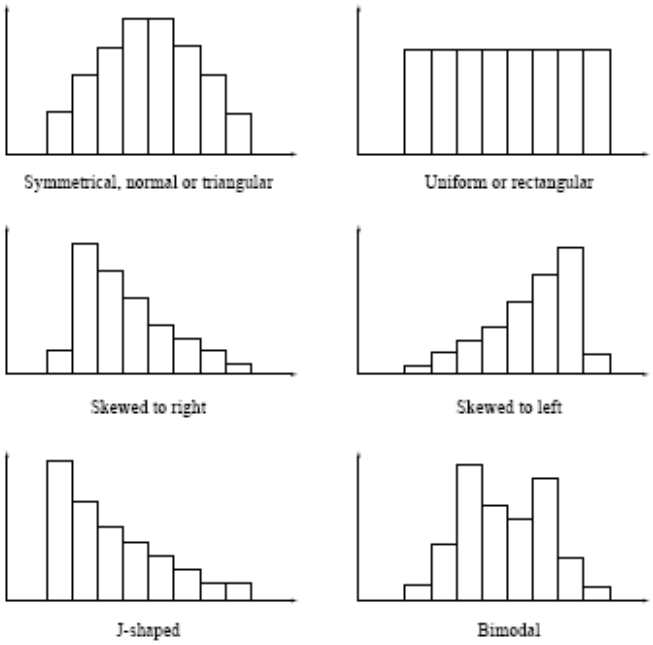
READING

Read Black et al (2010), Australasian Business Statistics, 2nd Edition, pp. 86–88.

The shape of a data set refers to how the data is grouped or organised when displayed as a graph. We will investigate various graphical techniques in topic 3 for indentifying shape (the figures shown below are examples of some of the graphs that will be covered in the next topic). For this topic, we briefly cover the terms used to describe shape:

- *Symmetrical*: Both sides of the distribution are identical.
- *Uniform (rectangular)*: All classes appear with equal frequency.
- Skewed: One tail is stretched out longer than the other. The direction of skewness is on the side of the longer tail.
- J-shaped: There is no tail on the side of the class with the highest frequency.
- Bimodal: The two most populous classes are separated by one or more classes. This situation implies that two populations may have been sampled.
- *Normal*: A symmetrical distribution that is mounded up in the middle and becomes sparse at the extremes.

These can be illustrated as follows:



A related formula is the Pearson's *skewness coefficient*. It is given by:

$$\text{Skewness coefficient, } s_k = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

It is a useful starting point to help detect skewness and it can be used for comparing the extent of skewness between different data sets. (There are other measures of skewness which could be used. For example, Microsoft Excel uses a different formula. See http://office.microsoft.com/en-au/excel-help/skew-HP005209261.aspx for details)

Note: When the mean is positive, then for positive skewness (where Mean > Median), the value of sk is positive. For negative skewness (where Mean < Median), the value of sk is negative.

As an example, if we found the mean = 8.3, median = 8.3 and standard deviation = 2.52, we find:

$$s_k = \frac{3(8.3 - 8.3)}{2.52} = 0$$

Confirm with a calculator that if the median had been 9.1 then the skewness coefficient would be -0.95.

The higher the s_k coefficient, the more likely the shape is skewed rather than 'symmetrical' or 'approximately symmetrical'. You should not rely solely on the s_k coefficient to determine skewness but always compliment it with a chart (e.g. histogram or polygon) in order to make your final judgment.

Outliers

An outlier can be thought of as a value that is separated from the rest of the data. It is important to detect them since they may involve an error in the data (someone made a mistake keying in the data, or in responding to the data) or may indicate an extraordinary, but legitimate, occurrence.

There might be more than one outlier in a set of data; there might not be any. Outliers should always be investigated to determine the cause.

In topic 3 you will see some graphical methods, especially the dot plot, which highlight potential outliers. The second method of detecting outliers is through simple calculation rules. The first method is based on what is known as the empirical rule.

Bell-shaped distributions

If our distribution of data is approximately bell-shaped, under the *empirical rule* we can use the following rules of thumb (where μ denotes the mean, and σ denotes the standard deviation):

Empirical rule ranges:

- about 68% of values lie in the range $\mu \pm 1\sigma$
- about 95% of values lie in the range $\mu \pm 2\sigma$

- about 99% of values lie in the range $\mu \pm 3\sigma$

A value in a bell-shaped distribution is considered to be a potential outlier if it lies outside the range $\mu \pm 3\sigma$.

Non-bell-shaped distributions

If our distribution of data is not approximately bell-shaped—meaning that we are referring to practically any other shaped distribution—we can use the following rule of thumb, sometimes referred to as *Tukey's rule*, where IQR stands for the inter-quartile range.

- 1.5 IQR rule (Tukey):
  - A value in a non-bell-shaped distribution is considered to be a potential outlier if it lies above the upper fence limit of 1.5xIQR above Q_3 or below the lower fence limit of 1.5xIQR below Q_1.

Note that the calculation rules are not the 'be all and end all' of outlier detection. They may highlight values which are not true outliers, or they might not highlight values that really are outliers. Thus, use the calculations methods in conjunction with graphs and plots of the data.

What do we do with outliers?

If outliers exist in a data set without correction or modification, they can severely distort some results. For example, the arithmetic mean (average) will be pulled up (or down) by an outlier at the top (bottom) end of the data. Thus, it is important to check for outliers.

If an outlier is found, a judgment must be made as to what to do with it. If an outlier is clearly an error (for example, the number '55' was entered instead of '5'), then the correction must be made. An outlier may be left in the data if it is a legitimate observation. Sometimes it will be removed, if the distortion is significant, even though it is a legitimate observation, and only the remaining values are analysed.

Note that extremes are not necessarily outliers. For example, the top income earner in a group of employees may be an extreme for that data set, but not necessarily an outlier, particularly if they earn only marginally more than the second highest income earner.

Which summary measures do we use?

When you use the statistics routines of spreadsheet or statistical software packages, you will find results are given for a large number of summary measures. Why are there so many summary measures?

The reason for this is associated with concepts such as: 'a little knowledge is a dangerous thing', the 'use and abuse of statistics', 'how to lie with statistics' and 'there are lies, damned lies and statistics'.

There is an inherent danger in using just a single summary measure to describe the features of a set of data. For example, the mean is often put

forward as being typical or representative of a given set of data. However, the more variability (as measured, say, by the standard deviation) there is in a data set, the less the mean (and perhaps any other average) is representative of the data. Thus, the use of averages should always be accompanied by the use of one or more suitable measures of variability.

Furthermore, sometimes one average (out of the mean, median or mode) may be more representative or meaningful than the others. This is of particular importance in skewed data where the value of the mean is affected by the extreme values at just one end of the data. In this case, the median (a positional measure) will be more representative as the average as it is less affected by extreme values. Thus, the use of the mean and other averages should also be accompanied by some investigation of the skewness in the data set (for example, by comparing the mean, median and mode, or by calculating an appropriate measure of skewness). The mode best represents the average when the data is categorical.

If you wish to explore and understand a set of data, it is advisable to always look at a range of summary measures, and to supplement those measures with graphical and tabular analysis.

VIDEO

View the Media Watch clip on online gambling

Summary

To assist you in the decision-making process you will require data of one form or another. The focus of this topic has been to examine the techniques for processing data, which you need in order to explore and summarise data and to present your findings to others.

You should always use a range of summary measures, as well as tables and graphs (see topic 3) when attempting to explore and describe a particular set of data.

Further resources

Anderson, DR, Sweeney, DJ & Williams, TA 2008, *Statistics for business and economics*, 10th edn, South-Western Thomson Learning, Cincinnati, Mass.

Selvanathan, A, Selvanathan, S, Keller, G & Warrack, B 2006, *Australian business statistics*, 4th edn, Nelson Thomson Learning, Melbourne.

Summarising business data: Graphs and tables

Contents

Introduction

In this topic we will look at ways of exploring data to help us summarise and highlight the important features of a given set of data for a single variable using either graphical or tabular methods.

Each type of table and graph has its advantages and disadvantages. We will find that what might be suitable for one purpose may not be suitable for another. We will also find that we often need to use a combination of descriptive or exploratory tools to help us fully understand the data.

Because of the power of modern computers and software, which enable us to quickly generate a large range of tables, graphs and calculations, exploring data is one of the most interesting and challenging aspects of analytics. Although this topic focuses on the fundamentals of good graphing practices, you will see examples of sophisticated charts to gain an appreciation of the power of graphical representation.

Objectives

At the completion of this topic you should be able to:

- distinguish between, and create, a wide range of tables, graphs and charts which can be used for presenting and exploring data
- identify when each descriptive tool should be used
- understand the principles of proper practice in tabular and graphical presentation
- use computer software for describing and exploring data (in the tutorial)

Introduction to data presentation

Selecting the right statistical tool

You will soon see that statistics seems to be an endless string of techniques, options, concepts and terms. The reason for such a large range is that statistical problems come under numerous guises, so statisticians have had to develop methods to meet each of these.

Because there is a wide range of statistical tools to choose from, you must choose 'the right tool for the right job'.

As an analogy, consider the job of cooking. The kitchen is the storage place for all your cooking tools. If you wish to bake a cake then you choose the right tools (utensils) for the job: a bowl, a mixer, a baking tin. Or consider the job of household repairs, and the garage or toolshed as the storage place for all of your tools. If you have to repair a broken chair, you choose the appropriate

tools: a chainsaw is clearly not appropriate, nor is a lawnmower, but a hammer and nails or screwdriver and screws are.

Good practice in presenting tables and graphs

It is extremely important to understand how to organise and most effectively present collected data in tables and charts.

There are some basic rules which should always be borne in mind for proper practice in presenting tables and graphs:

- First, every table, graph and chart should have a clear title, and all columns, rows, axes and sections must be clearly labelled.
- Second, all time periods (for example, 'year ended 30 June 2007') and units of measurements (such as $m) must be clearly stated.
- Third, the source of the data must be clear so that others know where the data came from and so that you and others can return to that source later.
- Finally, the data must not be distorted or misrepresented (for example, by truncating axes or by making one bar or section stand out).

For example, how useful is an exhibit (table or graph) with no title or no time period started?

From now on, take critical note of each table or graph you come across, or any reference to data to see whether or not it has been properly presented. Similarly, you need to be very conscientious about ensuring that exhibits you prepare meet these standards.

READING

Read UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE, Making Data Meaningful Part 2: A guide to presenting statistics (2009), pp. 7–10 (http://www.unece.org/fileadmin/DAM/stats/documents/writing/MDM_Part2_English.pdf).

Tables and charts for categorical data

In this section we will concentrate on variables that are strictly categorical variables. In particular, we will only look at a single variable at a time since our aim is to understand that variable as much as possible.

As a general rule, the most we can do with categorical data is to construct a *frequency count.* Thus, we count the number of occurrences in each category, and we generally calculate the percentage in each category. For example, we can count the number of people who would vote for each political party, such as the Australian Labor Party, the Liberal Party, and so on. We can also determine the proportion of people who would vote for these parties. The only summary measure that is really relevant to categorical data is the mode: the most common occurrence. For example, we could say that in an election 'the most common vote was for party X'.

The specific techniques you need to understand from this section are:

- Frequency count (or summary table)
- Bar chart (bars shown horizontally) / Column chart (bars shown vertically)
- Pie chart.

All are easily created, particularly using computer packages (see tutorial for details on how to create the above charts and table in Excel).

READING

Read Hardin, M. et al., *Which chart of graph is right for you?*, Tableau White Papers, pp. 2–6.

Tables and charts for numerical data

The specific techniques you need to understand from this section are:

- Array
- Ranking
- Frequency count
- Dot plot
- Bar chart* / Column chart*
- Pie chart*.

All are very useful tools for organising and previewing numerical data. They are most useful if the number of observations is not too large: for example, if there are 50 or fewer observations, but especially for 20 or fewer. (For a larger number of observations, other techniques are recommended and these are covered later.)

*Note that under our dictum of 'the right tool for the right job', we only use the *column*, *bar* and *pie* charts with caution for numerical data in raw form. An explanation of when we can use these is given later in this topic.

Frequency count

A *frequency count* is a table which provides information on the number of occurrences of individual values in a data set.

- It is a useful way of summarising raw data if individual values are repeated frequently.
- Exhibit 3.1 gives the frequency count for the Price of second hand cardata.
- It is also useful to provide the *per cent count* (also known as the *relative frequency,* that is, the percentage of the total items having a particular value), the *cumulative frequency/count* (that is, the number of observations up to and including a particular value) and the *cumulative*

per cent (that is, the proportion of observations up to and including a particular value).

- Say we are interested in $8000 or less. Thus, we observe that of the 20 cars, two, or 10%, were priced exactly at $8000, while 17 cars, or 85%, were priced at or below $8000.
- In this case we have produced the frequency count from the raw data. Often you will find data already presented to you as a frequency count table. The applications and interpretation are the same.
- You should be able to produce a frequency count 'by hand' for small data sets (say, 20 or fewer observations).

Exhibit 3.1: Frequency Count of Price for 20 second-hand cars

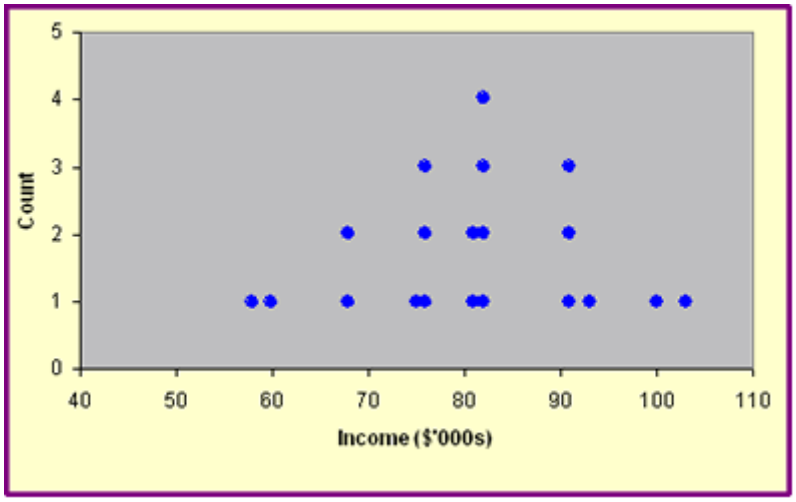
Price	Count	Percent	Cum. Freqy	Cum. Percent
3000	2	10.00%	2	10.00%
4000	3	15.00%	5	25.00%
5000	5	25.00%	10	50.00%
6000	3	15.00%	13	65.00%
7000	2	10.00%	15	75.00%
8000	2	10.00%	17	85.00%
9000	1	5.00%	18	90.00%
10000	2	10.00%	20	100.00%
Total	20	100.00%		

Dot plot

A *dot plot* is a graph (plot or chart) which plots a separate point for each occurrence of a value.

- It has a standard/arithmetic scale on both axes. On the X axis we plot the variable under consideration, while we plot the frequency count on the Y axis. In the body of the graph we place a dot above each successive occurrence of each value.
- It is a very simple but useful graphical technique for exploring the features of a given set of data where there are no more than about 50 observations, and those values tend to repeat reasonably often.
- Exhibit 3.2 shows a dot plot of salaries. It is a very useful graph as we can easily see the shape of the data, the range, and any clustering around central values, and any outliers. You need to be able to interpret this type of graph.

Exhibit 3.2: Dot plot for Income



- A dot plot will not necessarily be useful for highlighting central tendency and clustering if you have a very large number of observations, and/or the values tend not to repeat. However, it may still be worth drawing in order to view the variability in the data, and to inspect for potential outliers.
- You should be able to produce a dot plot 'by hand' for small data sets (say, 20 or fewer observations).

Column, bar and pie charts

Column, bar and parts are some of the most often seen graphs.

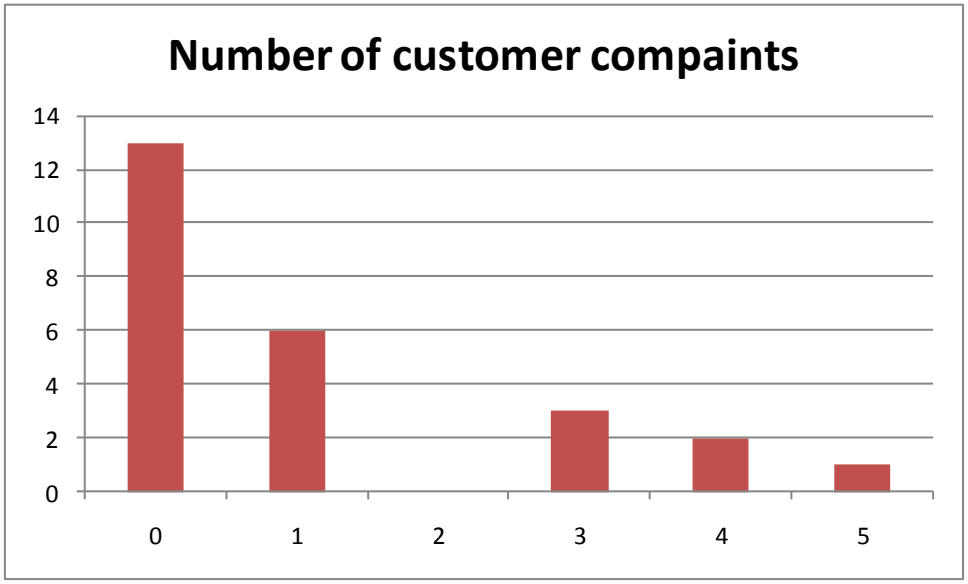
A *column chart* has a vertical bar drawn to the height of the frequency count for each value or category.

A *bar chart* has a horizontal bar drawn to the length of the frequency count for each value or category.

A *pie chart* breaks a circle into segments in proportion to the frequency count for each value or category.

- These charts can be created from the raw data via a frequency count as we did above for the dot plot. Or they can be created from a table already in frequency count form.
- If you are working with a numerical variable, it is advisable to use a numerical-based plot with an arithmetic scale on each axis (such as the dot plot, or the histogram which we discuss later). However, on occasions, the column, bar or pie chart may be more appropriate when that data is in a discrete format: just be cautious.
- A n example of a column chart is given in exhibit 3.3.

Exhibit 3.3: Daily customer complaints across a 25 day period



- One of the big advantages of the column, bar and pie charts is that they can be drawn under a number of different formats. Again, while these look effective, ensure they are conveying the information correctly, without the possibility of distortion or misinterpretation.

Practice exercises

Now complete the following computer exercises which are important in developing your computer skills. (Use methods you have learned in Excel from the tutorials).

In completing these exercises, keep the following in mind as good practice when studying a variable or set of data:

- Do I have a feel for the data?
- What do the exhibits tell me about central tendency/location, variability and shape?
- Are there any potential outliers?

EXERCISE 3.1

RESTAURANT MEALS

The data come from the file RESTRATE. When you open this data set, you will find it relates to 100 restaurants. From the 'Location' column you would find that 50 are NYC (New York CITY) restaurants, and the remaining 50 are LI (Long Island, SUBURBAN) restaurants.

(a) Open the file RESTRATE.

(b) Using the variable 'Price', create an ordered array for NYC restaurants. (While the array is useful for a small number of observations, because there are so many values here—50 of them—the array is not that useful in helping to understand the data. We proceed to show other ways of summarising the data.)

(c) Produce a frequency count for this variable.

(d) (Optional) produce a dot plot for this variable. You should see that the dot plot is informative in several ways. The range of the data and the clustering towards the middle are made quite clear. The fairly symmetrical shape of the data is made clear. We can also use this graph for checking for outliers: you should see that there appears to be one potential outlier at the negative end and one at the positive end.

Grouping numerical data using frequency distributions and histograms

The techniques described in the previous section will not be of use if there is a large number of values to analyse (say 50 or more) and/or the values tend not to repeat. You will have seen from exercise 3.3 that the frequency count (and subsequently the dot plot) did very little to help summarise the features of the data. In such instances, we need other methods for effectively summarising the data. Two such techniques are the frequency distribution and the histogram, which we study in this section. These techniques are appropriate when the data are continuous or effectively continuous (for example, where the data are strictly discrete but with many values occurring, such as the number of customers entering a large department store which might range from 1000 to 4000 per day).

We still concentrate on variables that are strictly numerical variables, and we look at just one variable at a time, showing how to group it into useful classes.

The specific techniques you need to understand from this section are:

- Frequency distribution
- Absolute, per cent (relative) and cumulative frequencies
- Histogram
- Frequency polygon and curve
- Ogive.

While in this subject we use software to do most of the calculation and graphic work for us, you will still need to know how to produce some tables, graphs and summary measures 'by hand', using a hand-held calculator if necessary. The notes indicate which techniques you need to be able to reproduce 'by hand'.

Frequency distribution

From your reading, you should have noted that grouping data in the form of a *frequency distribution* is useful for large data sets. Class intervals replace the individual values used in a frequency count. In each class we list the frequency with which values occur. Even by itself a frequency distribution can provide a convenient tabular 'picture' of the variable. To do this it is important to closely scrutinise the results in the table.

You should be able to produce a frequency distribution 'by hand' for small data sets (say 20 or fewer observations).

Note that often you will not have access to the actual raw data, but to the grouped data itself. That is, you will be given the data in frequency distribution form. The method of analysis from there on is the same as given below. Thus, you can still create per cent and cumulative frequencies, draw a histogram, etc.

Absolute, per cent (relative) and cumulative frequencies

The frequency distribution shows the absolute frequencies. However, like the frequency count, sometimes we need to work with per cent (relative) frequencies or with cumulative count/per cent frequencies. To summarise:

- The *absolute frequency* in a class is the actual number of occurrences in that class.
- The *relative frequency* in a class is the proportion of occurrences falling in that class.
- *Cumulative 'less than'* frequencies show the number or proportion less than particular values.
- You should be able to calculate relative and cumulative frequencies 'by hand' for small frequency distributions (say, 10 or fewer classes).

Histogram and related graphs

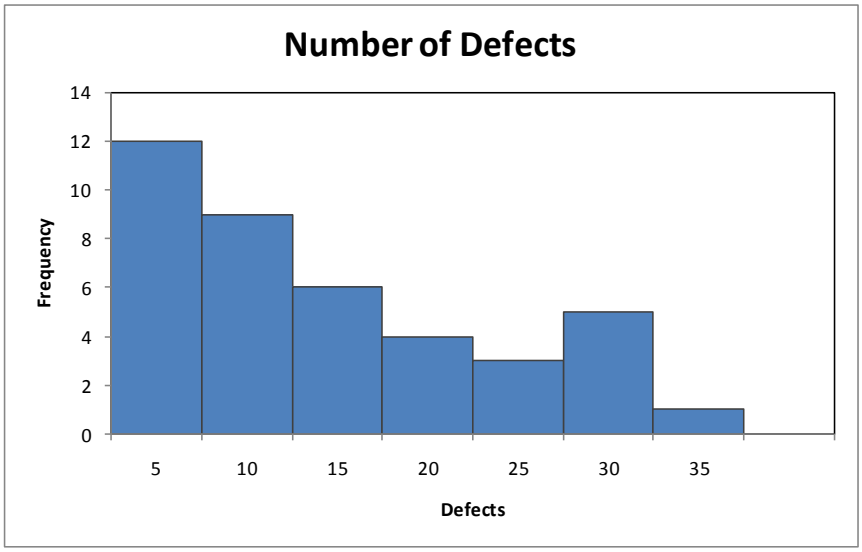
For a true picture of the data, we can graph a frequency distribution as a *histogram*. A histogram is like a column chart of the frequencies in a frequency distribution, except that the columns are joined together to represent the continuous nature of the data. The variable is plotted on the horizontal axis, with each bar width equal to the width of an interval. The frequencies are plotted on the vertical axis, with each bar height equal to the class frequency.

Instead of the absolute frequencies, relative frequencies can be plotted on the histogram. Clearly this type of graph is most useful for helping us get a feel for the data, for analysing important patterns and features (shape, central tendency, etc.), and for looking for potential outliers. Outliers may not always been seen in a histogram due to the interval nature of the data. You may find a dot plot is a better chart for identifying potential outliers.

You should be able to draw a histogram 'by hand' for small frequency distributions (say, 10 or fewer classes).

Exhibit 3.4 shows an example of a histogram for the number of defective products across 40 shifts.

Exhibit 3.4: Histogram of number of defective products across 40 shifts



Frequency polygon and frequency curve

A *frequency polygon* and the related concept, the *frequency curve*, are derivatives of the histogram in the sense that the top (middle) of each column is joined by a line, however the columns are not shown in the final chart. They attempt to provide a less cluttered, smoother representation of the data. The frequency curve is a 'free-hand' representation of the polygon, hence, it is not totally accurate.

You should be able to draw a frequency polygon or curve 'by hand' for small frequency distributions (say, 10 or fewer classes).

Ogive

An *ogive* is a graph of either cumulative absolute or cumulative relative frequencies. It is useful for assessing the number (per cent) of items 'no more than' or 'more than' a given value. As for a histogram, the variable under consideration is drawn along the horizontal axis; however, on the vertical axis we plot the cumulative frequencies.

You should be able to draw an ogive 'by hand' for small frequency distributions (say, 10 or fewer classes).

In summary, we use histograms or frequency polygons or curves for graphing actual or relative frequencies, and use ogives for graphing cumulative actual or relative frequencies.

EXERCISE 3.2

Complete the following exercises using what you have learnt in the tutorial.

(a) Open the file RESTRATE which you should have stored in your working folder.

(b) Using the variable 'Price', create the frequency distribution. Use the starting values for the first two classes will be '5.0 10.0'.

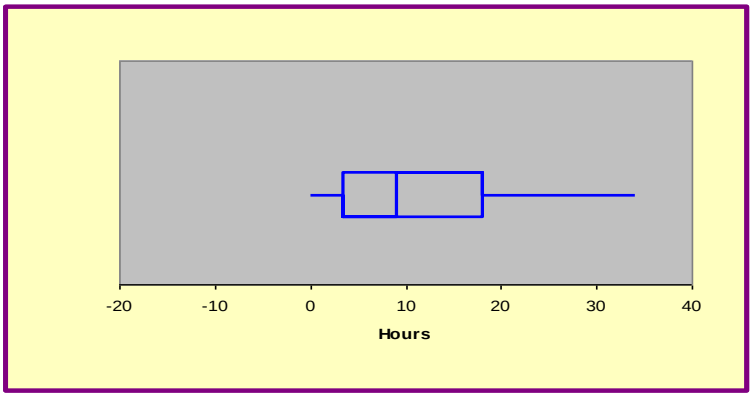
(c) Produce a histogram.

Outliers and box plots

Five number summary and box plots

One of the statistician's most useful graphs/plots is the box plot (or box-and-whisker plot). It is used in explanatory analysis and is based around five summary measures (the minimum, maximum, the first and third quartiles, and the median). To draw a box plot, start by drawing a box with its ends at Q1 and Q3. A vertical line is then drawn at the median. Finally, 'whiskers' are drawn to the minimum and maximum. Exhibit 3.5 shows an example of a box plot.

Exhibit 3.5: Box plot of number of defective products across 40 shifts



Box plots provide a very useful summary of data and allow us to compare for different shapes of data, and for highlighting central values, variability, location and outliers. Ensure you can interpret a box plot in regard to these features. For example, the diagram in Exhibit 3.5 highlights that the data is positively skewed.

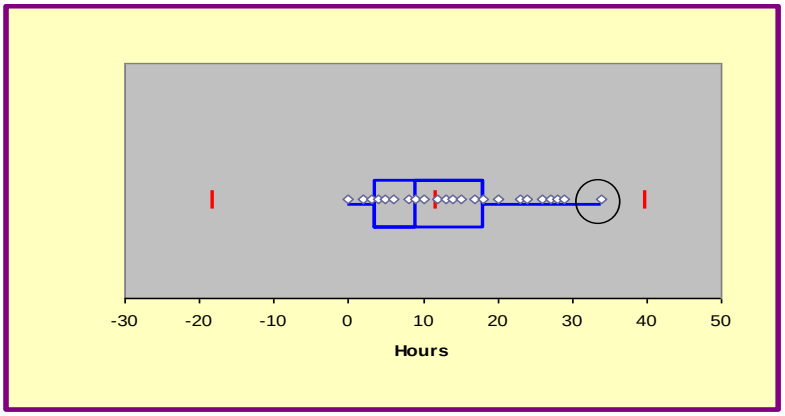
A particular advantage of box plots is in identifying outliers. The specific calculation ranges were discussed in topic 2. Any value outside these ranges are considered to be potential outliers. The standard range to use with box plots is the 1.5 IQR rule.

Exhibit 3.6 shows a box plot with the 'fences' shown from the 1.5 IQR rule. The graph initially suggests that there are no outliers as both the maximum and minimum are within the 'fences'. However, looking at the actual data (white dots in the diagram), we can see that one value (34 defects) does stand apart from the rest of the data which indicates that it may indeed be an outlier requiring further analysis.

You should be able to calculate the five number summary 'by hand' and draw a box plot 'by hand', including marking in the 1.5 IQR or 3 IQR limits.

Exhibit 3.6: Box plot of number of defective products across 40 shifts

with ‘fences’



Graphical excellence

In this topic we have explored some of the ways that data can be presented graphically. In order for these graphical displays to be useful in analysis and decision making, we must ensure that they are presented in a clear and accurate manner. The reading below provides some guidelines for proper graphing methods and some of the common errors that distort our visual interpretation. See also the lecture notes for good and bad examples of graphing.

READING

Skim read UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE, Making Data Meaningful Part 2: A guide to presenting statistics (2009), pp. 11–29 (http://www.unece.org/fileadmin/DAM/stats/documents/writing/MDM_Part2_English.pdf).

Summary

To assist you in the decision-making process you will require data of one form or another. The focus of this and the previous topic has been to examine the techniques for processing data, which you need in order to explore and summarise a *single* data item and to present your findings to others.

Data can initially be examined using graphical and tabular methods (this topic), and also via summary measures (topic 2). Graphical and tabular techniques represent powerful methods of easily and quickly conveying results to others whilst summary measures are used to summarise particular features (central tendency and location, spread and shape) of data in a single measure.

Thus, you should always use a range of summary measures, tables and graphs when attempting to explore and describe a particular set of data.

Further resources

Black, K 2008, *Business statistics for contemporary decision making*, 5th edn, Wiley, NJ.

Anderson, DR, Sweeney, DJ & Williams, TA 2008, *Statistics for business and economics*, 10th edn, South-Western Thomson Learning, Cincinnati, Mass.

Selvanathan, A, Selvanathan, S, Keller, G & Warrack, B 2006, *Australian business statistics*, 4th edn, Nelson Thomson Learning, Melbourne.

TOPIC 4

Analytical techniques for discovering relationships in data

Contents

Introduction

So far we have only considered one variable at a time, but clearly once we find significant variation in a variable, we want to understand that variation. Why don't all investment funds make the same return? Why don't all Conrobar employees perform to the 100% productivity level? What are some of the underlying factors explaining or causing these variations?

Thus, in this and later topics, we look at relationships between two or more variables. For example, is there a relationship between employee salaries and job satisfaction? Or between salary and gender? Do sales results for a firm depend on the amount it spends on advertising? Do the selling prices of houses depend on the number of rooms? Is there a relationship between incidence of skin cancer and gender?

The term 'relationship' clearly means two variables are related in some sense. If a relationship exists, we call the variable that is dependent on the other variable the *dependent* variable. (We can also say it is the explained variable.) A variable which helps predict or explain outcomes for the dependent variable is called an *independent,* or *explanatory* (explaining) or *predictor variable*.

For example, age and income are related in a positive or direct way. Generally—within a given range—the older people are, the more income they earn. In this instance, age is the independent variable and helps explain variation in income. We can also say that income *depends* on age.

Car prices and age are generally related in a negative or inverse manner. We generally find the older a car, the lower the price. Price is the dependent variable, while age is the independent/explanatory variable.

These few examples are cases of relationships between *two numerical variables*.

Exploring for relationships is one of the most interesting areas of statistics—the statistician gets the chance to play detective.

In detecting relationships, our aim is to help explain and understand the variation in the dependent variable, and/or use the independent variable to help predict or even control likely outcomes for the dependent variable. For example, if there is a strong relationship between unemployment and interest rates, then by varying interest rates up or down, the government can have some control over the movement in the unemployment rate.

Objectives

At the completion of this topic you should be able to:

- identify possible relationships or dependence between two or more variables whether they be categorical or numerical
- utilise computer software to display relationships or dependences between two or more variables.

Cross-tabulations for analysing two categorical variables

The main technique you learn about here is the *cross-tab,* a two-way *cross-classification* table also called a *contingency table*. All three names are acceptable. Plotting data from a cross-tab can also be very effective.

Cross-tabulations are used to detect whether a relationship exists between two categorical variables. Consider an example of comparing customer satisfaction ratings (poor, good, excellent) across two store locations (Upfield, Downton).

Frequency Count		**Location**	
Customer Satisfaction	**Downton**	**Upfield**	**Total**
Poor	62	22	84
Good	67	83	150
Excellent	9	57	66
Total	138	162	300

The table summarises 300 customer responses with location shown in the columns and satisfaction rating shown in the rows (note that the choice of which variable to place in the rows/columns in arbitrary). In this case, we can quickly obtain a good overview of the data. For example, half (150 out of 300) of the customers rate the service as good; only nine customers at Downton rate the service as excellent; 22 customers at Upfield rate the service as poor; and so on.

Note the following:

- A cross-tab provides a cross-classification of a particular situation according to two variables or characteristics.
- Public opinion polls published in the daily press are generally in cross-tab form, for example, a table might show voting intention (by party) down the left side and Australian state along the top. The cells in the table tell the number of cases possessing the two characteristics classified on the left and top sides, e.g. a cell might tell us that of 1000 people surveyed, 20 lived in South Australia *and* would vote for the ALP.
- As well as showing the actual frequencies, the table can also be used to show percentages in various categories.
- A cross-tab is not limited just to cross-tabulating two categorical variables. You can also have numerical variables on one or both sides of the table, where that variable has been grouped (either into values such as 1, 2, 3, or into classes such as '$0 to less than $50,000', '$50,000 to less than $100,000', etc.).
- As a normal rule, limit the number of rows and columns in your table, otherwise you will find many cells have zero or very low counts in them.

As well as showing absolute frequencies in a cross-tab, it is often useful to show percentages, either as an overall percentage or row or column percentages. See below.

Overall Percentage	Location		
Customer Satisfaction	**Downton**	**Upfield**	**Total**
Poor	20.7%	7.3%	28.0%
Good	22.3%	27.7%	50.0%
Excellent	3.0%	19.0%	22.0%
Total	46.0%	54.0%	100.0%

Row Percentage	Location		
Customer Satisfaction	**Downton**	**Upfield**	**Total**
Poor	73.8%	26.2%	100.0%
Good	44.7%	55.3%	100.0%
Excellent	13.6%	86.4%	100.0%
Total	46.0%	54.0%	100.0%

Column Percentage	Location		
Customer Satisfaction	**Downton**	**Upfield**	**Total**
Poor	44.9%	13.6%	28.0%
Good	48.6%	51.2%	50.0%
Excellent	6.5%	35.2%	22.0%
Total	138	162	100.0%

- The 'Per cent of Row' and 'Per cent of Column' tables are very important in identifying these relationships/dependencies.
- If there tends to be no relationship between the two classified variables then the percentages *across rows* will tend to be similar or the percentages *down columns* will tend to be similar. Quite dissimilar percentages (across rows or down columns) indicate a potential relationship (or dependent situation).
- The percentage table chosen to analyse should be the table where the independent variable is located as the independent variable is the explanatory variable. That is, if the independent variable is placed in the row, you should use 'Per cent of Row' to compare percentages across the rows. Likewise, if the independent variable is located in the column, you should use 'Per cent of Column' to compare percentages across the columns.

In the customer satisfaction example, we are interested in comparing the satisfaction levels (dependent variable) across the two store locations (independent variable). The column percentages show that although the two locations both have approximately half of their customers who rate the service as good, the percentages that rate the service as poor or excellent are very different. Thus we would conclude that customer satisfaction appears to be dependent on store location. Customers at Downton are more likely to rate the service as poor and at Upfield they are more likely to rate the service as excellent.

EXERCISE 4.1

A survey of 1200 residents in the suburb of Downton collected data about work status and gender. A partially completed cross-tabulation showing the frequency counts is shown below:

Frequency Count	**Gender**		
Work Status	Female	Male	Total
Casual		199	356
Part Time	118	159	
Full time		211	567
Total	631	569	1200

(a) Complete the missing fields in the cross-tabulation.

(b) Which of the variables is the dependent variable?

(c) Would you use a Row Percentage or Column Percentage cross-tabulation? Explain your choice.

(d) Complete the appropriate cross-tab and comment on whether the variables are related.

Scatter diagrams for analysing two numerical variables

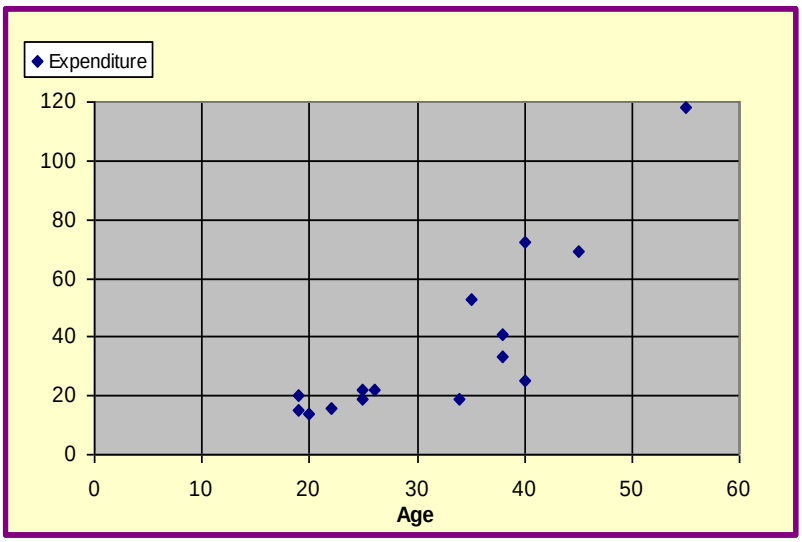
A *scatter diagram,* or *scatter plot*, uses an *XY* coordinate system to measure the independent *X* variable on the horizontal axis and the dependent *Y* variable on the vertical axis, with points plotted at each intersection of the *X* and *Y* values for each observation.

Consider the following example of comparing the amount spent by a customer at a supermarket versus the customer's age.

Expenditure	Age
$19	25
$19	34
$20	19
$25	40
$22	26
$41	38

$22	25
$15	19
$72	40
$118	55
$69	45
$33	38
$53	35
$14	20
$16	22

In this case, expenditure is the dependent variable and age is the independent variable. The graph below shows the data plotted as a scatter diagram.



- Scatter diagrams (or scatter plots) are normally used for plotting two numerical variables against each other.
- In drawing scatter diagrams it is the normal procedure to plot data points with dots or small crosses. However, a refinement is to plot a characteristic of the point, such as M for male and F for female, or to plot the actual value. Plotting a line or drawing a freehand curve through the points may also help the analysis.
- The points in a scatter plot may or may not show a useful pattern. With two numerical variables, straight-line and curved relationships are of interest. Clusters of data, and their location, may also be meaningful.

  In the above example, the scatter diagram shows a straight-line relationship. Thus there is a positive linear relationship between the two variables. There appears to be a moderate relationship between expenditure and age. In general, the older a customer is, the more they tend to spend in a single transaction.

- As it is often linear relationships which we are looking for, often we plot a straight 'line of best fit' on a scatter plot. This concept *line of best fit* will be

explained in detail later in the unit. Until then, consider the *line of best fit* to be the 'line that is closest to all the points' on the scatter plot.

The XY coordinate system

Most graphs are plotted with a *horizontal axis* and a *vertical axis*, using what is known as an *XY coordinate system*.

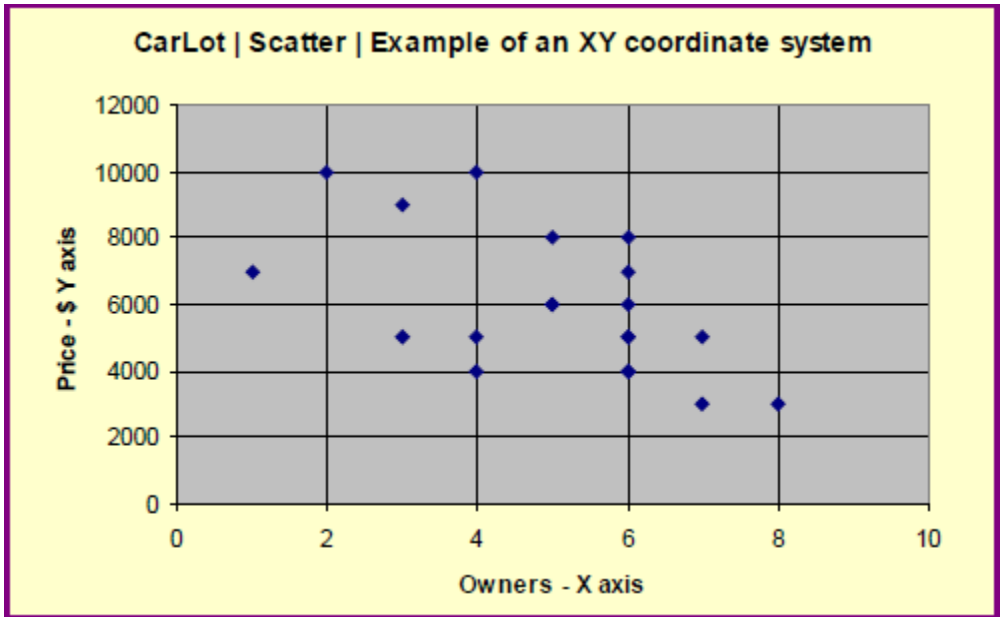
The *horizontal axis* is known as the *X axis*, because this is the axis on which we tend to plot the independent variable which, by convention, we denote by X.

The *vertical axis* is known as the *Y axis*, because this is the axis on which we tend to plot the dependent variable which, by convention, we denote by Y.

For example, if we wanted to observe people's *weight* according to their *age*, then clearly *weight* is the dependent variable. ('Your weight depends on your age'. It does not make sense to say: 'Your age depends on your weight'.)

Exhibit 4.1 (scatter diagram) is an example of an XY coordinate system. In it we have plotted the Price of cars (from exhibit 3.1) on the vertical axis against Owners on the horizontal axis.

Exhibit 4.1: Scatter plot of Price v Owners on an XY coordinate system



This is a coordinate system with an *arithmetic scale* on both axes.

An *arithmetic scale*, or *standard scale*, means that individual lengths on the scale have actual meaning. Thus, if you are 2 centimetres from zero, and move a further 2 centimetres up the Y axis or to the right on the X axis, you are doubling the value of the variable plotted on that axis. For example, in exhibit 3.2 $4000 is plotted 2 grid marks from zero on the vertical axis, and the value at 4 grid marks is exactly double, at $8000. On the horizontal axis, every extra grid mark corresponds to 2 extra owners.

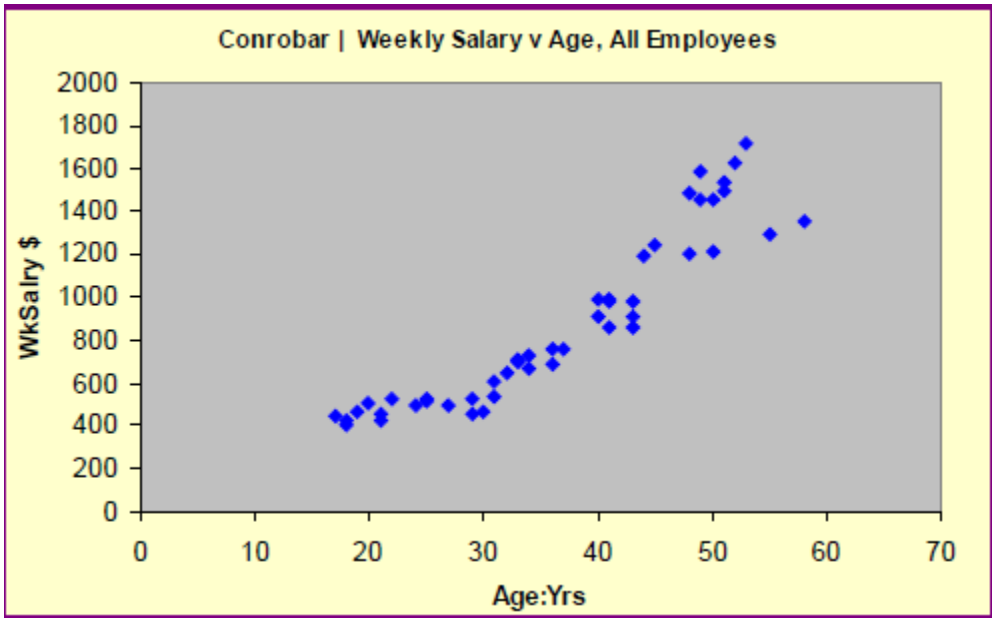
With an XY coordinate system, we plot a point at each combination of X and Y. For example, Car No. 1 in exhibit 3.1 has had 8 previous owners, and is priced at $3000. Thus, we can plot the (X, Y) point on the graph as (8, $3000).

The X and Y axes cross at (0, 0). Larger values of X are graphed to the right—that is, the positive numbers are graphed to the right of 0—while negative values are graphed to the left of 0. Similarly, positive values of Y are graphed above 0 and negative values below 0. In statistics it is not often necessary for us to use the negative sections of the XY coordinate system.

Scatter diagram examples

The following are some examples from the Conrobar data set. Exhibit 4.2 plots Weekly Salary (on the vertical axis) against Age on the horizontal axis.

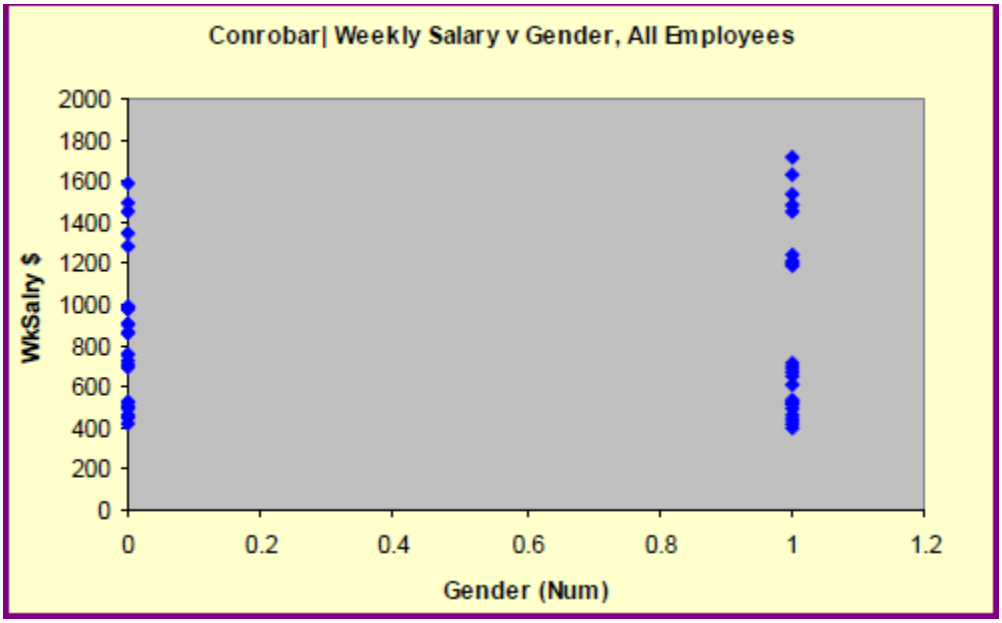
Exhibit 4.2



Here, we note (as we would expect) a positive relationship between age and salary. Note the slight upward curve. A relationship does not have to be a straight line to be significant. A distinct curved pattern is just as meaningful, and potentially just as useful, though we will see (when we study regression) we generally prefer to work with linear relationships, or relationships that can be 'linearised'.

Scatter diagrams can also be useful for analysing one numerical variable against a categorical variable. This is an example of where the coded form of a categorical variable can be used. Exhibit 4.3 is also a plot from the Conrobar database. It plots Weekly Salary (on the vertical axis) against Gender on the horizontal axis.

Exhibit 4.3



This diagram indicates that Female salaries cover a wider range than Male salaries. There appear to be clusters for both genders, although more pronounced for females. There are certainly interesting patterns in this graph, but there is no evidence to suggest that income is dependent on gender, that is, neither gender tends to earn more than the other.

EXERCISE 4.2

The data below represent the total volume (litres) and the energy consumption per year (kWh/annum) of 18 single door refrigerators available in both Australia and New Zealand.

Brand	Model	Total Volume (litres)	Energy Consumption (kWh/annum)
HELLER	BFH12	100	275
SAMSUNG	SRG149PT	125	300
TELMANN	TEL170R	125	300
LG	GR-051SSF	50	200
WESTINGHOUSE	Refrigerator RP372	275	325
HAIER	HBF80	75	250
HAIER	BC-76	75	275
WESTINGHOUSE	Refrigerator RA110T	100	300
WESTINGHOUSE	Refrigerator/Freezer RA122T	100	275
WANBAO	WB-70D	50	250
HAIER	BC-145	125	300
HAIER	BC-117SS	100	300

	Upright Refrigerator		
WESTINGHOUSE	RP142	125	300
ELECTROLUX	Refrigerator CS370	200	300
FISHER & PAYKEL	C373	300	325
DEC	D-60FR	50	225
CENTRAX	CTBF47 (nil)	50	275
SIGNATURE	SBR-120	125	275

a) Construct a scatter diagram with the volume on the horizontal axis and energy consumption on the vertical axis.

b) Does there appear to be a relationship between volume and energy consumption?

Exploring for relationship between one numerical variable and one categorical variable

The cross-tab is used when comparing two categorical variables (one or both of which can be categorised numerical variables). The scatter diagram is used when we have two numerical variables. However, often we need to see if there is a relationship (or dependency) when one of the variables is numerical and the other is categorical. Two such techniques include comparing summary measures for subsets of data and comparing multiple plots on one chart.

Table of comparative summary measures

Previously we produced and analysed summary measures for a single numerical variable. You will see that we often wish to compare two or more groups on the basis of summary measures: comparing the averages, variability, etc. For example, comparing productivity performance of males and females at Conrobar. A very useful table for this purpose is a table of *comparative summary measures*.

You will be developing such tables in the next exercise below. There you will see that by comparing measures for each group we may be able to detect a potential relationship or dependency. These summary measures may suggest that there is a relationship between the dependent variable and the grouping categorical variable. Any suggestion of relationship/dependency can be tested statistically using inferential techniques developed later in this course. At this stage, we use the table of comparative summary measures to give some indication of possible relationships.

The following comparative summary measures represent the time taken in minutes to solve problems of a sample of 20 customers at two different Internet Service Providers ("Easy' and 'Mighty').

	Easy	*Mighty*
Mean	2.21	2.01
Median	1.54	1.51
Mode	1.48	3.75
Standard Deviation	1.72	1.89

Sample Variance	2.95	3.58
Skewness	1.13	1.47
First Quartile	0.93	0.60
Third Quartile	3.93	3.75
Minimum	0.52	0.08
Maximum	6.32	7.55
Range	5.80	7.47
IQR	3.00	3.15
Sum	44.28	40.23
Count	20.00	20.00

Although there appears to be no real difference in the average (Median) time taken for the two providers, 'Easy' appears to be slightly more consistent (smaller IQR) in the time taken and 'Mighty' has a higher skewness coefficient. Overall there is not much difference between the two providers and we would conclude that the two variables are not related.

EXERCISE 4.3

Below are comparative summary measures comparing salaries of male and female employees at Conrobar. Is there a relationship between salary and gender? Explain the differences (if any).

WkSalry	*Male*	*Female*
Mean	$852	$900
Standard Error	$68	$97
Median	$810	$680
Mode	$760	#N/A
Standard Deviation	$348	$457
Sample Variance	$120,936	$208,695
Kurtosis	-0.40	-1.38
Skewness	0.69	0.54
Range	$1,170	$1,320
Minimum	$420	$400
Maximum	$1,590	$1,720
Sum	22153	19807
Count	26	22
Q_1	$530	$523
Q_3	$984	$1,234
IQR	$454	$711

Multiple box plot

In the previous topic, you produced a box plot for a single numerical variable. As discussed earlier, the box plot is used in explanatory analysis and is based around the five summary measures.

A principal advantage of the box plot is for comparative purposes, since two or more box plots can be drawn on the one graph. This is an extremely useful aid in comparing central tendency/location, variation and shape.

To draw a multiple box plot, you simply draw each box plot on the same axis and compare (see lecture notes for examples).

EXERCISE 4.4

Using the information in Exercise 4.3, draw a multiple box plot showing weekly salary vs. gender.

Summary

If we find a variable of interest (often this will be our dependent variable) shows a certain amount of variation, the obvious question is 'why'? That is, what could be an explanation or cause or predictor of this variation? For example, humans don't all have the same resting pulse rate. But we know that there are many variables that explain this variation between individuals, including age, height, weight, fitness level, gender, ethnic background and diet.

Thus we often wish to see if any differences exist between sub-groups, or if there are any dependencies or relationships that may assist us.

We have introduced several descriptive methods for exploring for relationships or dependencies between two (or more) variables. These relationships may be between numerical variables only, categorical variables only or between numerical and categorical variables. We use different techniques in each case.

In this topic, we introduced a number of general methods for exploring for relationships:

- *Cross-tab (contingency table)*, and the per cent of row and per cent of column tables, and associated column charts.
- *Scatter plot*, with the *line of best fit* and R^2 for two numerical variables.
- *Table of comparative summary measures* whereby equivalent summary measures for different groups can be directly compared.
- *Multiple box plots* for comparing two or more groups for a numerical variable, permitting us to explore possible similarities or differences among the groups.

It should be emphasised, however, that the techniques we look at here belong to descriptive statistics, their aim being to help us understand the data better, and to search out potential relationships. Refinements of these techniques are taken up in later topics where we will present techniques which allow us to determine, in a more definite manner, whether such relationships do, in fact, exist.

Further resources

Black, K 2008, *Business statistics for contemporary decision making*, 5th edn, Wiley, NJ.

Anderson, DR, Sweeney, DJ & Williams, TA 2008, *Statistics for business and economics*, 10th edn, South-Western Thomson Learning, Cincinnati, Mass.