1331.0

Australian Bureau of Statistics

Statistics — A Powerful Edge!

Second Edition

W. McLENNAN
Australian Statistician

AUSTRALIAN BUREAU OF STATISTICS

ABS Catalogue No. 1331.0
ISBN 0 642 25744 2

Produced by the Australian Bureau of Statistics

INQUIRIES

For information about other ABS statistics and services, please refer to the back page of this publication.

For further information about the content of this publication, contact Soo Kong on Melbourne (03) 9615 7360.

CONTENTS

PREFACE

Statistics - A Powerful Edge! Second Edition is a resource book from the Australian Bureau of Statistics (ABS) about getting the most from statistics.

It is published for secondary students of Mathematics and Information Studies, although it is expected that the book will find a wider use amongst other students, teachers and general readers.

The book aims to assist students:

gain confidence in using statistical information to complete study requirements;

appreciate the importance of statistical information in today's society; and

make critical use of information that is presented to them, whatever its source.

All these goals are at the heart of the ABS mission to assist informed decision-making in the Australian community.

Along with extensive text, the book contains exercises to help students consolidate their understanding of presented material.

This edition has updated much of the data that was used in the original version so as to keep examples and exercises as relevant and topical as possible. Developments in technology, such as the growing use of the Internet and CDRoms, also led to sections being rewritten.

The release of this edition continues the ABS's commitment to provide service and appropriate resource materials to the education sector in Australia.

I trust that ***Statistics - A Powerful Edge! Second Edition*** will assist Australia's students to strengthen their understanding and use of statistics, both in their current studies and in future years.

W. McLennan
Australian Statistician
July 1998

GENERAL INFORMATION

THIS PUBLICATION:

General enquiries concerning this publication should be directed to the Manager, Education Services, ABS Victoria on (03) 9615 7360.

Comments on ways to improve this and other ABS publications for schools are welcome; and should be addressed to the Manager, Education Services, ABS Victoria, GPO Box 2796Y, Melbourne 3001.

FURTHER READING:

For further studies in sample surveys the ABS recommends:
An Introduction to Sample Surveys: A User's Guide (ABS Cat. No. 1202.2).
This publication is expected to be re-issued in 1999 as ABS Cat. No. 1202.0.

This publication is intended as a basic guide on the use of sample surveys, for the conduct of all types of research. Major topics include: survey objectives, data collection methods, questionnaire and sample design, sources of error, survey testing, data collection, processing, analysis and presentation of results.

SYMBOLS:

In all tables the following symbols mean:

n.a.	not available
n.y.a.	not yet available
p	preliminary
..	not applicable

INTRODUCTION

DATA, INFORMATION & STATISTICS

DATA, INFORMATION & STATISTICS

As the world approaches the 21st Century we are facing new and challenging problems. More than ever before, governments, industry and the wider community need information to help them to make decisions to tackle these problems.

The need for an informed society is one reason why Australian secondary education is developing an emphasis on students gathering data and presenting information to complete work requirements. In some cases students gather data, process it and present it as statistics. Before undertaking such activities it is important to have a sound understanding of the terms *data, information* and *statistics.* They are often misunderstood.

DEFINITION

Before one can present and interpret information there has to be a process of gathering and sorting data. Just as crude oil is the raw material from which petrol is distilled, so too, data can be viewed as the raw material from which information is obtained. Therefore, a good definition of data is:

Data are observations or facts which when collected, organised and evaluated become information or knowledge.

6 3.33 0.1
Daisy Elizabeth
(1) Yes

Data can take various forms, but is often numerical. As such, data can relate to an enormous variety of events, for example: the number of twins born every day, or the number of times Australia has beaten England in a one-day cricket match. Other forms of data exist; such as radio signals, digitised images and laser patterns on a compact disc.

The Australian Bureau of Statistics (ABS) collects data every day to provide information. For the 1996 Census of Population and Housing the ABS collected nearly seven hundred million (696,000,000) separate observations! An example of one of these observations is shown at the top of the next page.

19	How well does the person speak English?	(1) Very well ( ) Well ( ) Not well ( ) Not at all

Once data have been collected and processed they are ready to be organised into information. Indeed, it is hard to imagine reasons for collecting data other than to provide information. This information leads to knowledge about issues, and helps individuals and groups to make informed decisions.

In practice, informed decision-making can save countries millions of dollars, for example: through accurate targeting of government spending. It can also lead to life saving breakthroughs in medicine, and can help conserve the earth's natural environment. Therefore, a good definition of information is:

Information is data that has been organised to serve a useful purpose.

Information, like data, can take various forms. The first known artefacts containing information date back 40,000 years; animal bones are believed to have been etched with information about phases of the moon. Astronomers are using information in the form of organised radio waves to explore space. In recent years, visual information through the medium of television and video has become very common.

The 17th Century, English philosopher Francis Bacon recognised the importance of knowledge many years ago. His quotation below is probably more true today than it has ever been, and is an important reason why the general public should have access to information. This theme is explored in more detail in the *Information Studies* section of this publication.

"KNOWLEDGE ITSELF IS POWER".

Francis Bacon

Statistics represent a common method of presenting information. In general, statistics relate to numerical data, and can refer to the science of dealing with numerical data or the numerical data itself. Above all, statistics aim to provide useful information by means of numbers. Therefore, a good definition of statistics is:

S tatistics are numerical data that have been organised to serve a useful purpose

AUSTRALIA	ENGLAND
289 and	210 and
5 for 432 dec.	332

A major role of the ABS is to provide the Australian community with statistics that will help them make informed decisions. Statistical information provided by the ABS is used widely in Australia: by governments, business people, doctors, farmers, teachers and students.

The provision of accurate and authoritative statistical information strengthens modern societies. It provides a basis for decisions to be made on such things as where to open schools and hospitals, how much money to spend on welfare payments and even which football players to replace at half-time! An example of decision-making from statistical information is given below.

In May 1993 the Victorian Transport Accident Commission began a radio, television and newspaper advertising campaign about motorcycle awareness. They did so in response to the previous year's statistics on road fatalities in Victoria.

These showed that despite an overall 15 per cent decrease in fatalities in metropolitan areas, the number of motorcyclists killed in metropolitan areas increased by almost 30 per cent!

+30%

The next four pages contain examples of statistical information. As you will see, statistical information can be presented in different ways, including: graph, table or illustration. Presenting statistics graphically is discussed in detail in the *Stats Maths* section of this publication.

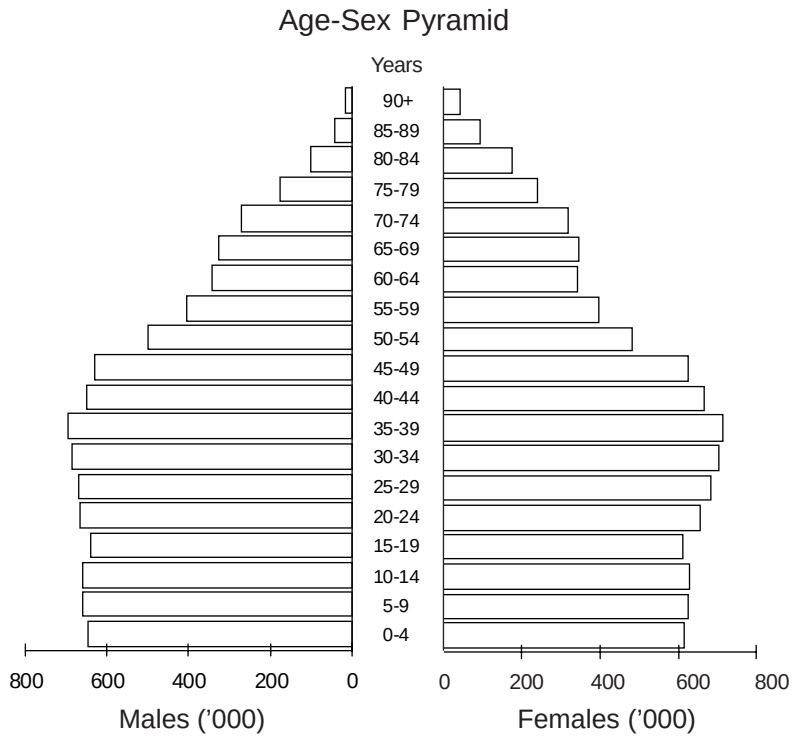
EXAMPLE 1	SELECTED OCCUPATIONS, VICTORIA, 1881 CENSUS	
	Males	***Females***
Boarding house keeper	162	458
Capitalists	378	103
Farm servant	5394	2160
Hotel keeper	3102	848
Jack-of-all-trades	1	0
Loafers	2	0
Mesmerist	1	0
Mudlarker	1	0
Music teacher	190	732
Nurse	0	981
Opium sellers	59	1
Scarecrow-on-a-farm	1	0
Statistician	1	0
Ventriloquist	1	1

EXAMPLE 1

This is a table of statistical information about occupations in Victoria last century. It shows the number of people in Victoria who had particular occupations at the time of the 1881 Census. Note how some occupations are referred to differently today!

(The total number of males and females in the table is 14,577.)

EXAMPLE 2 | POPULATION, AUSTRALIA, 1996 CENSUS



EXAMPLE 2

This chart is an *age-sex pyramid* from the 1996 Census. It shows statistical information on Australia's population by age-group and sex .

Age-sex pyramids are commonly used to present statistical information on the composition of a population. In the chart the population of Australia totals 17,892,423.

EXAMPLE 3	AUSTRALIAN FOOTBALL LEAGUE, GRAND FINAL, 1997																							
	Kicks				Marks				Handball				Free		Scr		ho	tk	sh					
ADELAIDE	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**F**	**A**	**G**	**B**								
D. Jarman	4	3	3	5	1	1	2	1	2	0	1	1	2	2	6	2	1	7	1					
K. Kosier	4	1	2	2	1	0	0	0	3	1	0	1	1	4	0	1	0	7	0					
N. Smart	3	2	2	2	2	1	1	1	2	1	1	0	0	2	1	0	1	3	0					
T. Edwards	0	4	2	3	0	0	1	0	0	1	1	0	0	1	0	1	0	1	1					
M. Robran	3	0	1	1	2	0	0	2	2	1	0	1	1	0	0	0	5	4	1					
B. James	0	2	5	2	0	0	2	0	0	0	4	2	0	1	0	0	0	5	1					
S. Ellen	3	0	2	3	1	0	2	3	1	1	0	2	1	0	5	1	1	3	1					
M. Connell	2	1	1	5	1	0	1	0	0	2	0	1	1	0	0	0	0	2	1					
D. Pittman	0	2	1	0	0	1	1	0	1	0	0	0	1	1	0	0	2	1	2					
T. Bond	3	1	2	1	3	1	1	0	0	1	0	0	1	2	4	0	0	0	0					
A. Keating	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	3	1	0					
A. McLeod	8	2	2	6	4	1	2	4	1	4	4	4	0	1	0	0	0	3	1					
C. Sampson	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0					
M. Bickley	2	2	1	1	0	1	0	1	1	1	0	0	0	0	0	1	0	6	0					
K. Johnson	2	3	2	3	0	0	1	1	3	2	4	2	1	1	0	0	1	2	0					
B. Hart	3	1	3	2	0	1	1	1	1	1	0	1	0	2	0	0	0	0	2					
R. Jameson	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0					
S. Goodwin	5	2	4	5	3	0	1	0	0	0	3	0	0	2	1	0	0	5	0					
C. Rintoul	1	0	3	2	0	0	1	0	0	2	1	2	3	0	1	1	0	3	0					
P. Caven	3	2	2	4	3	1	0	2	1	0	2	1	2	1	1	0	0	0	1					
S. Rehn	3	2	4	3	1	0	3	2	0	1	2	2	0	0	0	0	16	4	0					

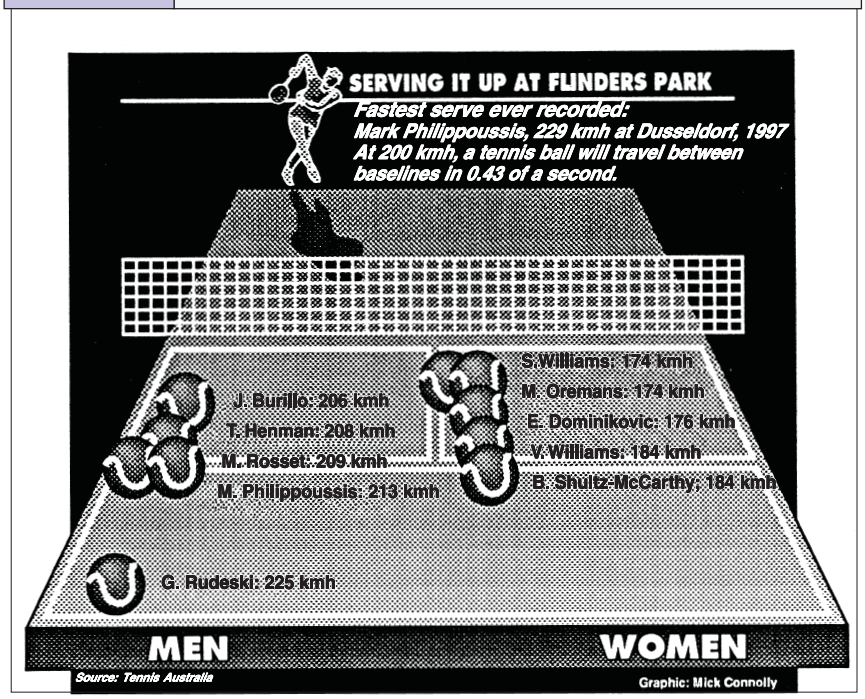
EXAMPLE 3

This table of statistical information shows how the Adelaide players performed in the 1997 AFL Grand Final.

The match statistics are shown quarter by quarter under the bold column headings 1, 2, 3, 4.

The last four column headings of the table indicate:
Scr = Score, ho = hit-outs, tk = tackles and
sh = shepherds.

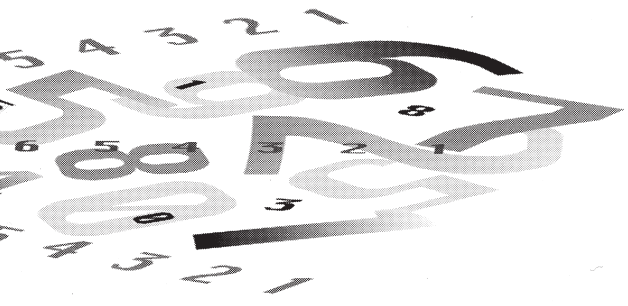
EXAMPLE 4 | FASTEST TENNIS SERVE, AUSTRALIAN OPEN, 1998



EXAMPLE 4

This is an illustration of statistical information. It refers to players who made the fastest serves at the 1998 Australian Tennis Open.

The illustration shows information about the speed of tennis serves, and the fastest serve recorded since measurements began.



EXERCISES

1. In your own words define the terms: *data, information and statistics.* Give examples of each.

2. Put the following terms in correct logical order:
KNOWLEDGE — DATA — INFORMATION

3. Identify three current political, economic or social issues for which information is necessary. Then describe the information that is needed for each issue.

4. On the previous four pages there are examples of statistical information. Which example contains the fewest observations and which example contains the most?

5. In Example 1, one occupation has no male units (zero observations). Does this mean that no information is available for males in that occupation?

6. In Example 2, which male and female age group has the largest population?

7. In Example 3, which Adelaide players recorded no marks for the entire match?

8. In Example 4, information on how many units (tennis players) are illustrated?

9. Who do you think might need the information in Example 1 and for what purpose?

10. Who do you think might need the information in Example 2 and for what purpose?

11. Does the information in Example 3 tell you accurately which players performed better than others in the Grand Final? Explain.

12. Which example required a scientific instrument to collect the data?

13. Which example shows all the individual observations collected?

14. These examples also illustrate the variety of ways in which statistics can be presented. Look in newspapers or journals for other ways in which statistics are presented. Be careful to distinguish between data, information and statistics as in question 1.

The quotation below is probably the most famous one about statistics. Later in this publication you will see how easy it is to be misled by statistical information. If used wisely, however, statistics can be a powerful tool in decision-making.

"THERE ARE LIES, DAMNED LIES, AND ... STATISTICS".
Mark Twain

INFORMATION STUDIES

DATA COLLECTION

Individuals and organisations collect data because they or someone else require information. They may want information to keep records, make decisions about important issues, or be required to pass information on to others. So, whatever the specific reason, data is collected to provide information.

But who in society wants or needs information? The answer to this question could be simply stated as 'many people and organisations'. Some of the groups that use statistics include:

GOVERNMENTS Federal, state and local govemments need information on the population and economy, among other things. This information helps them to make decisions on issues such as where to build hospitals, locate services, or how much money to raise through taxation. It also allows the public to hold a government to account by measuring its performance.

BUSINESSES Most Australian businesses require information. This information may be about the economy, the profile of a local population or various social trends. It helps them to make decisions about employing people, where to market their products and where to open new offices, warehouses and factories.

COMMUNITY GROUPS These organisations need information about a wide variety of subjects, for example: Aboriginal health and population distribution, or the number and location of people with poor English proficiency. Sporting clubs may want information about attendances at matches or the number of young people in their local area.

INDIVIDUALS

Everyone, from students to pensioners, needs information at some time. It may be needed to complete an essay, a major project or simply to satisfy one's curiosity.

TYPES OF DATA COLLECTION

There are three main types of data collection: census, sample survey, and administrative by-product. Each has advantages and disadvantages over the other. As students you may well be required to collect data at some time. The method you choose will depend on a number of factors.

CENSUS

A census refers to data collection about *everyone* or *everything* in a group or population. So, if you collected data about the height of *everyone* in your class that would be regarded as a class census. There are various reasons why a census may be chosen as the method of data collection:

ADVANTAGES	
Accuracy:	Everyone in a group has had data collected about them, resulting in a high degree of accuracy.
Detail:	Detailed information about small subgroups of the population can be made available.

DISADVANTAGES	
Cost:	In money terms, conducting a census can be expensive for large populations.
Speed:	Time taken to do a census can be long compared to a survey.

SAMPLE SURVEY

In a sample survey, only *part of the total population* is approached for data. So, if you collected data about the height of 10 students in a class of 50, that would be a sample survey of the class rather than a census. Reasons to select a survey include:

ADVANTAGES	
Cost:	A survey costs less than a census because only part of a group has had data collected about it.
Speed:	Results are obtained far more quickly than for a census. Fewer people are contacted, and less data needs processing with a survey.

DISADVANTAGES

Accuracy: Depending on sample size, results have a degree of inaccuracy.

Detail: Information on small population sub-groups or small area geography is not usually obtainable, unlike a census.

ADMINISTRATIVE BY-PRODUCT

Administrative by-product data is collected as a by-product of an organisation's day to day operations. Examples include data on: births, deaths, marriages, divorces, airport arrivals, and motor vehicle registrations. For example, prior to a marriage license being issued, a couple must provide the registrar with information about their age, sex, birthplace, whether previously married, and where they live.

ADVANTAGES

Accuracy: Data is collected about everyone who uses that organisation's service, resulting in a high degree of accuracy.

Time series: Data is collected on an on-going basis, allowing trend analysis.

Simplicity: Administrative data may eliminate the need to design a census/survey (and associated work), and saves the public having to complete further forms.

DISADVANTAGES

Flexibility: Data items may be limited to essential administrative information, unlike a survey.

Control: The agency which controls the data may restrict access to outsiders or charge for access.

EXAMPLE

1. Census. The ABS's largest data collection exercise is its five yearly *Census of Population and Housing.* The information provided by this Census is used by all four groups referred to on page 15.

Before every Census of Population and Housing, the ABS invites data users to submit topics they would like included. Not every submission for new information is successful. So how does the ABS decide which topics data is gathered on by the Census?

For the 1996 Census the ABS used the following principles:

- Whether the topic was of major national importance. The Census is a large and costly operation imposing a burden on householders who are required to answer questions. It is essential that every question has a specific purpose.
- Whether the topic was suitable for inclusion. Census topics should not cause an adverse reaction from people by unacceptably invading their privacy. They should not require an overlong explanation or instruction to ensure an accurate answer, and should not refer to things people are unlikely to remember.
- Whether the Census was an appropriate method of collecting the data. Consideration should be given to the alternatives to a Census. In some cases the information being sought may already exist; it may be collected by another organisation, or administrative records may provide the required data.
- Need to limit the number of questions on the form so it won't take too long for the public to complete.

Some topics have been included in every National Census since 1911. From the table opposite you can see what they are. For example, *Name, Age, Sex, Birthplace, Citizenship, Religion* and *Occupation* have always been included. An asterisk in the table means the topic was covered in the Census.

However, other topics have not been covered in every census. After the table opposite, some topics have been chosen to explain the reasons in more detail.

SELECTED CONTENT OF CENSUSES (a), 1911 TO 1996

TOPIC	1911	1921	1933	1947	1954	1961	1971	1976	1981	1986	1991	1996
Name	*	*	*	*	*	*	*	*	*	*	*	*
Age	*	*	*	*	*	*	*	*	*	*	*	*
Sex			*									
Orphanhood			*									
Birthplace	*	*	*	*	*	*	*	*	*	*	*	*
Birthplace of parents		*					*	*	*	*	*	*
Citizenship	*	*	*	*	*	*	*	*	*	*	*	*
Disability								*				
Ethnic origin										*		
Blindness, deaf-mutism	*	*	*									
Religion	*	*	*	*	*	*	*	*	*	*	*	*
Educational qualifications							*	*	*	*	*	*
Holidays								*				
Income			*					*	*	*	*	*
Occupation	*	*	*	*	*	*	*	*	*	*	*	*
Journey to work							*	*	*	*	*	*
Mode of travel to work								*	*	*	*	*

(a) Excludes 1966 Census.

DISABILITY ■ A general question on the effect of disabilities was asked in 1976. Questions relating to specific disabilities among the population were included in the 1911, 1921 and 1933 censuses. The questions asked people to indicate whether they were deaf, dumb or blind.

Information on disabilities is required by Federal and State Governments to help develop policies for disabled people . For example, where to best place services and how much money to set aside.

The 1976 Census found that information on disabilities was unreliable. This has also been the experience of Censuses in other countries, and has been confirmed in recent testing in Australia. Because of the sensitivity involved in answering a question on disability, and the

difficulty in deciding what constitutes being disabled, many people under-report their situation while others inappropriately identify as disabled when this is not how others would classify them.

ETHNIC ORIGIN

■ A question on the ethnic origin (or ancestry) of the population has only been asked once, in 1986.

Information on ethnicity is needed to identify changing patterns of cultural diversity within the population. It is also used in community relations programs, targeting government access and equity strategies, and measuring well-being of ethnic groups.

The ABS retained all ethnicity-related 1991 Census questions for the 1996 Census, including: birthplace, citizenship, birthplace of parents, language and religion.

However, the 1986 Census question, *"What is each person's ancestry?"* was not included. The ABS did not believe the additional ethnicity information the question provided was sufficient to justify inclusion in the 1996 Census.

INCOME

■ A question on income was first asked in the 1933 Census to measure effects of the Great Depression. It was re-included in 1976 and all subsequent Censuses. Income information helps locate the disadvantaged for social service planning.

Collecting data on income has its problems. There is a general tendency among respondents to under state income, particularly social welfare payments and interest earned on financial investments. Pensioners sometimes state they receive no income, as they do not regard their pension as income.

HOLIDAYS

■ The only census question on holidays was asked in 1976. At the time, there was interest to know if people on lower incomes took as much time off for holidays as those on higher incomes.

However, the ABS found it necessary to reduce the size of the Census questionnaire in following censuses. People find it burdensome when a questionnaire is large, so the question on holidays was dropped as it was considered to be of less national importance than the other retained topics.

OCCUPATION

■ A question on occupation has been asked in all Censuses of Population and Housing since 1911. Governments at all levels (Federal, State and Local) need detailed information on Australia's occupation patterns. Information on occupation is vital to government policies and programs in the fields of education, training, immigration and industry.

The Census provides information about occupation patterns in small geographic areas, whereas sample surveys do not. For example, you can compare the proportion of working population who are *tradespersons* in, say, Brunswick, Bondi and North Adelaide.

CENSUS: A GLOBAL HISTORY

BC		
3800	BABYLON	Carried out every six or seven years. It counted asses, oxen, butter, milk, honey, and wool.
2500	EGYPT	Carried out to assess the labour force available for building pyramids.
1491	ISRAEL	Carried out to count people liable for military service and taxation purposes.
550	CHINA	Carried out by Confucius to obtain information on the nation's agricultural, industrial and commercial state
AD		
1719	PRUSSIA	Europe's first systematic census.
1790	USA	America's first census.
1801	ENGLAND FRANCE	England's and France's first census.

CENSUS: AN AUSTRALIAN HISTORY

1788	'Musters' of convicts began. During this year they were held weekly because of the colony's dependence on public stores and dread of famine. These 'musters' were also held on Sundays at the church parade. Families of convicts and free settlers were required to attend. Punishment of up to 500 lashes could be enforced for non-attendance.
1828	The colony's first regular census was held in NSW.
1881	The colony's first Australia-wide census was held. This was part of a simultaneous census of all British Empire colonies.
1911	The first census held under the Census and Statistics Act of 1905. This was followed by censuses in 1921, 1933, 1947, 1954, 1961 and thereafter at five yearly intervals.

EXAMPLE

2. **Labour Force Survey**. The ABS conducts many sample surveys each year. One of these is the monthly Labour Force Survey, from which information is produced on Australia's employed and unemployed persons. The survey obtains information from 65,000 persons each month, through a sample of about 29,000 private dwellings, and a further sample of non-private dwellings such as hotels, motels and caravan parks.

As with the Census of Population and Housing, many people in the Australian community need and use information from the Labour Force Survey. Governments need the information to assess whether their economic policies are changing the level of employment; and service providers need to know the areas where unemployment is highest, to be able to target assistance to people.

Sometimes there is a need to obtain more information than just employment and unemployment. Therefore, the ABS may ask additional questions in the Labour Force Survey. Topics can vary, and in the past have included job search experience of unemployed persons and number of people with more than one job.

Many important items of information are produced from the Labour Force Survey, for example: youth participation in the labour force (proportion of all teenagers aged 15-19 years who are either employed or unemployed). Information on this topic is shown below, and the figures are for August of each year.

Do you think the information below is useful in describing the general situation for teenagers? What other data should the survey collect for this purpose?

LABOUR FORCE PARTICIPATION RATES (%): 15-19 YEAR OLDS, AUGUST, AUSTRALIA

1967	1972	1977	1982	1987	1992	1997
Females						
61.1	56.2	57.6	56.1	53.6	54.2	53.6
Males						
64.7	58.5	62.1	62.4	57.4	54.0	52.9

METHODS OF DATA COLLECTION

PERSONAL INTERVIEW:

- *Face to face:* involves trained interviewers visiting people to gain data. It is good for ensuring a high response rate to a sample survey or census, and trained interviewers should be able to gather accurate data. The ABS conducts face to face interviews for its Labour Force Survey during the respondent's first month, and by telephone for subsequent months. However, it is costly to train interviewers, occasionally respondents are unavailable, and travel costs could be high.

- *Telephone:* involves trained interviewers phoning people to gain data. It is quicker and cheaper than face to face interviewing. However, only people with phones can be interviewed, and the interviewed person can end the interview very easily!

- *Computer Assisted Telephone Interviewing (CATI):* involves a phone interview as above, but with the interviewer keying respondent answers directly into a computer. This saves on time involved in processing data, but can be expensive to set up, and needs interviewers with computer and typing skills. The ABS uses CATI for its Retail Survey.

SELF-ENUMERATION:

- *Postal survey:* a common method of conducting ABS economic surveys. It is a relatively inexpensive method of collecting data, and one can distribute large numbers of questionnaires in a short time. It provides the opportunity to reach difficult to contact people, and respondents are able to complete the questionnaire in their own time. Postal surveys do require an up-to-date list of names and addresses. Added to this is the need to keep the questionnaire simple and straightforward.

  A major disadvantage of a postal survey is that it usually has a lower response rates than other data collection methods. This may lead to problems with data quality, and therefore reliability of results. People with a limited ability to read or write English may experience problems.

- *Hand-delivered questionnaire:* a self-enumerated survey where questionnaires are hand-delivered to people and collected later. This method usually results in better response rates than a postal survey, and is particularly suitable when information is needed from several household members. This method is used for the ABS's Census of Population and Housing.

- *Hand-delivered mail-back questionnaire:* a combination of hand-delivered and postal methods, which reduces the cost of collecting completed forms. It gives a greater sense of privacy for respondents concerned with someone entering their home or business to collect forms.

ROLE OF BIAS

The effect of bias is to prejudice or unfairly influence data quality (see definition of bias on page 179). In practice, bias can be deliberate or unconscious. This is loooked at in more detail in the section *Information - Problems with Using.*

For now, it is worth noting that the method of data collection itself can bias or unfairly influence data collection results. For instance, television program polls in which viewers call either a “yes” or “no” number to register their opinion are open to bias. The survey population may not be representative of the wider community, the program may bias callers before voting with one-sided information, and there is nothing to stop individuals voting many times to sway results.

ROLE OF DATA COLLECTORS

The role of data collectors is very important. The process of interviewing people to collect data involves a number of skills. Without these skills the quality obtained data can be reduced. Therefore, when someone is employed to collect data they should have, for example:

- use of a car and telephone,
- good communication skills,
- a confident and professional appearance, and
- the freedom to work evenings and weekends.

The ABS employs a large number of interviewers to collect data. Interviewers are trained before collecting data. This training emphasises that the interviewer's opening remarks and manner in which they are made have a strong influence on a respondent's reaction and willingness to co-operate. Because of this, data collectors should ensure certain things are carried out before they ask people questions, including:

- give the respondent their name,
- explain that a survey is being conducted and by whom,
- provide identification and give the person time to read it,
- explain that the respondent's household or business has been selected in the survey sample, and
- explain the survey's purpose.

In addition, it is important that the data collector is familiar with correct interview technique. Such technique includes:

- the ability to listen attentively,
- keeping the interview short,
- asking questions the same way for each respondent interviewed, and
- NOT suggesting any answers for the respondent.

EXERCISES

1. Suggest reasons why data would be collected on the following topics:

   a) Burglaries b) Causes of death c) Climate

   d) Forests e) Immigration f) Schools

2. When collecting data, why is it sometimes better to conduct a sample survey than a census?

3. List some of the reasons why the ABS may decide not to include a question in the Census of Population and Housing on:

   a) Health b) Marital status c) Sports participation

4. List some of the things you would need to consider when choosing a data collection method.

5. Given some of your answers to question 4, decide as a class which method of data collection you would employ to gather data on the following topics:

   a) favourite rock or pop group in your class,

   b) average height of your class,

   c) time your parents spend each week doing housework, and

   d) attitude of Australians toward the environment.

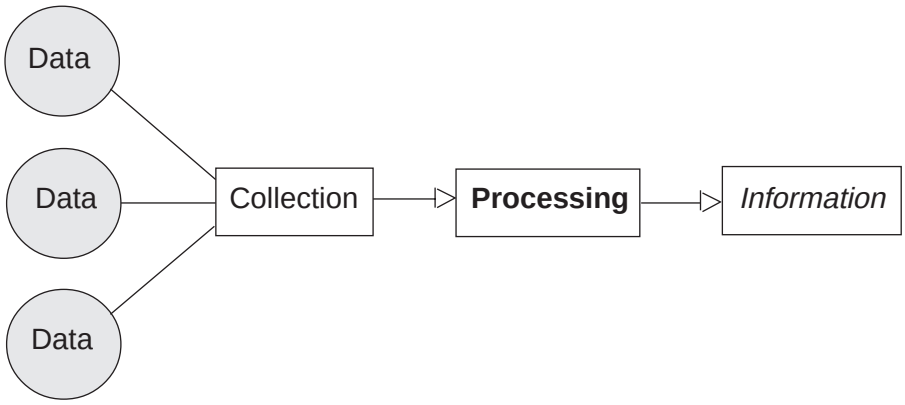
6. Are there some topics for which data should not be collected? For example, data on people's health or political beliefs? Discuss as a class where you would 'draw the line' in deciding if issues are too sensitive to ask about.

   What factors would help you make your decision? For example, is the information you need from the data of national importance?

DATA PROCESSING

Data are raw facts. When organised and presented properly, they become information. Turning data into information involves several steps. These steps are known as *data processing*. This section looks at data processing and the use of computers to do it easily and quickly.

The diagram below shows a simplified view of the procedure for turning data into information. Data, in a range of forms and from various sources, may be entered into a computer where it can be manipulated to produce useful information (output).



Data processing includes the following steps:

- data coding,
- data input,
- data editing, and
- data manipulation.

DATA CODING

Before raw data is entered into a computer it may need to be coded. Coding involves labelling the responses in a unique and abbreviated way (often by simple numerical codes). The reason raw data are coded is that it makes data entry and data manipulation easier. Coding can be done by interviewers in the field or by people in an office.

A *closed question* implies that only a fixed number of predetermined responses are allowed, and these responses can have codes affixed on the form. An *open question* implies that any response is allowed, making subsequent coding more difficult. One may select a sample of responses, and design a code structure which captures and categorises most of these.

DATA INPUT

The keyboard of a computer is one of the more commonly known input, or data entry, devices in current use. In the past, punched cards or paper tapes have been used.

Other input devices in current use include light pens, trackballs, scanners, mice, optical mark readers and bar code readers. Some common everyday examples of data input devices are:

- bar code readers used in shops, supermarkets or libraries, and
- scanners used in desktop publishing.

The ABS gathers data from censuses and surveys. The method of data entry varies, depending on the type or method of collection.

- The 1996 *Census of Population and Housing* used optical mark readers to read the forms.
- Data from surveys using computer-aided telephone interviews (CATI) are entered by means of a keyboard while survey staff telephone interview the respondent. The ABS's *Retail Trade Survey* uses CATI.

- Tests are being made for household surveys using computer assisted personal interviewing (CAPI). Interviewing staff enter data directly into notebook computers during the interview at a respondent's house. Laptop computers are considered too bulky for this type of work, but hand held computers which use lightpens for data entry are also being tested.

DATA EDITING

Before being presented as information, data should be put through a process called editing. This process checks for accuracy and eliminates problems that can produce disorganised or incorrect information. Data editing may be performed by clerical staff, computer software, or a combination of both; depending on the medium in which the data is submitted.

Some editing processes are:

Validity check: ensures that data fall within set limits. For example, alphabetic characters do not appear in a field that should have only numerical characters, or the month of year is not greater than 12.

Verification check: checks the accuracy of entered data by entering it again and comparing the two results.

Consistency check: checks the logical consistency of answers. For example, an answer stating *never married* should not be followed by one stating *divorced.*

Data editing should detect and minimise errors such as:

- questions not asked by interviewers,
- answers not recorded, and
- inaccurate responses.

Inaccuracy in responses may result from carelessness or a deliberate effort to give misleading answers. Answers needing mental calculations may result in errors, for example: when converting days into hours, or annual income into weekly income.

EXAMPLE

1. This example of data to be edited shows an inaccurate response. By carefully reading this section of the ABS Labour Force Survey form you should be able to detect a very inaccurate response!

34A. ON WHICH DAYS DID *Person 1* WORK LAST WEEK (IN ALL JOBS)?

	MON	TUES	WED	THU	FRI	SAT	SUN
Yes	☑	☑	☑	☑	☑	☑	☑
No	☐	☐	☐	☐	☐	☐	☐

34B. DID *Person 1* HAVE ANY TIME OFF FROM *this* JOB(S) ON THOSE DAYS?

Yes ☐
No ☑

34C. DID *Person 1* WORK ANY PAID OR UNPAID OVERTIME ON ANY DAY LAST WEEK?

Yes ☑
No ☐
Other ☐

34D. HOW MANY HOURS DID *Person 1* ACTUALLY WORK LAST WEEK (LESS THE TIME OFF) (BUT) (COUNTING THE OVERTIME)?

35 hours or more ☐

1 - 34 hours ☐

Less than 1 hour/ no hours ☑

Question 34A shows that Person 1 said they worked on every day of the previous week. Question 34B shows there was no time off, and Question 34C says some overtime was also worked. However, Question 34D says that all of this amounted to less than one hour of time worked!

The answers to individual questions look acceptable. It is only by comparing them with each other that you find if one or more are wrong.

This cross-checking is only one type of edit. It could be performed either by clerical staff or editing software. It indicates that further action should be taken. In the previous example, the interviewer will get in touch with the household and re-check how many days and hours were worked by Person 1.

DATA MANIPULATION

After editing, data may be manipulated by computer to produce the desired output. The software used to manipulate data will depend on the form of output required.

Software applications such as word processing, desktop publishing, graphics (including graphing and drawing), databases and spreadsheets are commonly used. Following are some ways that software can manipulate data:

- *Spreadsheets* are used to create formulas that automatically add columns or rows of figures, calculate means and perform statistical analyses. They can be used to create financial worksheets such as budgets or expenditure forecasts, balance accounts and analyse costs.

- *Databases* are electronic filing cabinets: systematically storing data for easy access to produce summaries, stocktakes or reports. A database program should be able to store, retrieve, sort, and analyse data.

- *Charts* can be created from a table of numbers and displayed in a number of ways, to show the significance of a selection of data. Bar, line, pie and other types of charts can be generated and manipulated to advantage.

Processing data provides useful information called *output.* Computer output may be used in a variety of ways. It may be saved in *storage* for later *retrieval* and use. It may be laser printed on paper as tables or charts, put on a transparent slide for overhead projector use, saved on floppy disk for portable use in other computers, or sent as an electronic file via the internet to others.

Types of output are limited only by the available output devices, but their form is usually governed by the need to communicate information to someone. For whom is output being produced? How will they best understand it? The answers to these questions help determine one's output type.

EXERCISES

1. Place the following in correct logical order:
   PROCESSING — COLLECTION — INFORMATION — DATA

2. List the steps involved in data processing and write a brief description of each.

3. Investigate the different types of data input devices that are present in your school. Do any require special skills to be able to use them?

4. If data editing did not take place, what effect might this have on information produced from the data?

5. The following responses to sample survey questionnaires contain inaccuracies or errors. Can you list what they are?

   a) Lambing during year ended 31 March 1995

	Number
Lambs marked	1054
Ewes mated to produce above lambs	yes

   b) What is your present marital status?

(-) Never married
( ) Married
( ) Separated
(-) Divorced
( ) Widowed

   c) How did you get to work on 5 April?
   (If you used more than one method mark all relevant boxes)

( ) Train
( ) Bus
( ) Ferry or tram
( ) Car
(-) Motorbike
( ) Bicycle
( ) Walked
(-) Worked at home
(-) Did not go to work

DATA AND COMPUTERS

Professional organisations have been using computers to process data and provide information for many years now. Computers and computer systems have been growing in sophistication and complexity, but their basic characteristics remain unaltered.

Computer systems are a combination of the following ingredients: *hardware, software* and *users* (people). Each is necessary for the system to function. The development history of computing systems has stressed the importance of each ingredient in turn, but it is important to remember that *all* of them have a necessary place in the system's operation.

COMPUTERS — A SELECTED HISTORY

BC 5000	EGYPT	Abacus developed, world's first calculating machine.
AD 1622	ENGLAND	Slide rule invented by William Oughtred.
1642	FRANCE	Mechanical calculator invented by Blaise Pascal.
1830s	ENGLAND	Charles Babbage conceived computing components such as input and output units and storage systems.
1859	ENGLAND	George Boole developed symbolic logic; his work is the basis for binary switching, upon which modern computing depends.
1886	USA	Herman Hollerith of the US Census Bureau developed mechanical device to use punched cards for compiling and tabulating data.
1946	USA	The very first electronic computer, ENIAC (Electronic Numerical Integrator and Calculator) was completed at the University of Pennsylvania. Its computing power was far less than that of a current notebook computer.
1951	USA	The first commercially available computer delivered to the US Census Bureau.
1981	USA	First IBM personal computers were introduced.

The electronic computer was introduced to the Australian business community and the world in the 1960s. It was only gradually adopted, as early models were very expensive and slow, and there was little expertise available. Commercial computing skills were not taught at university, and business software applications could not be bought, but had to be developed in-house by programming staff. Good programmers were scarce and expensive.

The following table shows development of computer usage in the ABS, which mirrored a growing use of computers in the large business community.

COMPUTERS — AN ABS HISTORY

1963	A computer running paper tape programs, with a memory of only 512 bytes, was used for initial data entry from punched cards and punched paper tape.
1964	A CDC 3600 mainframe computer was installed for statistical processing. At the time it was one of the biggest in Australia, with 32,768 words of memory, about 20 MB of disk storage and a number of magnetic tape units.
1966	The ABS's largest collection of data, the Census of Population and Housing, was processed using software developed in-house.
1972	Data entry terminals, connected to the mainframe, were installed to replace electro-mechanical devices such as punched tape, punched cards, or key-to-tape equipment.
1976-79	Advances in data communications and increasing computing power enabled centralisation of data resources and the ability to access that data from all State offices. Large hard disk units gradually replaced 100,000 or so magnetic tapes.
1979	A Desktop publishing package was introduced for the production of statistical publications.
1982	A Fujitsu "IBM compatible" mainframe computer was installed, replacing our CDC mainframe computers.
1986	Typing pools were equipped with personal computers (PCs) for routine word processing, replacing electronic "memory" typewriters. Spreadsheet and other PC software was available on these machines
1988	The ABS distributes population statistics on CD-ROM, and other information such as publications catalogues on diskette.

1989	A PC based Local Area Network was installed which provided an ABS-wide electronic mail system. Typing pools were disbanded and staff did their own typing on desktop PCs.
1991	Unix based mid-range computers were installed to supplement our large Fujitsu mainframe computer. Applications like Computer Aided Telephone Interviewing were used.
1993	Lotus Notes based document databases and electronic mail system introduced. More and more staff had a PC on their desk.
1994	Our Fujitsu mainframe computer, the size of a large room and which required large industrial airconditioning systems, was replaced by a new Fujitsu mainframe computer the size of a vending machine. The new machine has the power of the old computer, but only requires normal office airconditioning and just plugs into a wall socket.
1995	ABS has its own Internet site, which contains contact details for ABS offices and news about latest releases of products and services. Clients can contact us by E-mail.
1995	Data warehousing system, which consolidates many ABS data holdings, used as a client servicing tool.

COMMERCIAL COMPUTER INSTALLATIONS

Large companies don't have just one large computer, but usually a network of specialist computers. Depending on company size, there could be one or two very powerful mainframe computers, a number of small, medium and large Unix based mid-range computers, and individual PCs for all staff who work with information.

These computers are networked together. The network connecting the computers in any one location is usually referred to as a Local Area Network (LAN), while all the individual LANs are connected together by a Wide Area Network (WAN). Using this network, authorised personnel can access data and make use of computer facilities anywhere in the system.

Until about 10-15 years ago, the mainframe was the most common sort of computer in business. Mainframe computers today typically run older computer applications (often called "legacy" systems) written a few years ago; and also applications which require the highest levels of reliability, security and processing power. If the highest performance is not required, these days corporations often find it easier and more economical to use mid-range computers, usually running under the Unix operating system.

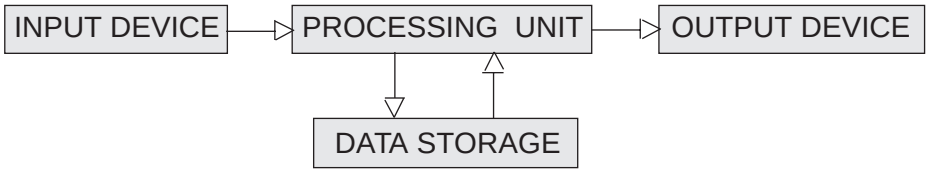
Regardless of the type of large computer being used, users interact with their computer systems, and obtain data using their desktop or notebook PC.

THE INTERNET and INTRANETS

Businesses, academic institutions and people at home use the Internet to send E-mail messages around the world or to assist in research. However, most businesses consider that the Internet is too slow, unreliable and not secure enough for widespread internal use. To solve this problem, they set up their own private internal Internets, known as *Intranets*. This gives them the ease of use and convenience of the Internet, but with the performance and security of an in-house system. Intranets can be connected to the public Internet via secure gateways, which have "firewalls" to prevent unwanted external access to internal systems.

COMPUTER HARDWARE

To function properly, a computer system needs the following hardware components:



An input device allows a user to enter data or program the computer. The processing unit controls all activities within the system. Data Storage holds databases, files and programs. Output devices present the finished information product to the user.

Input device. Data can be supplied for input to a computer in a range of ways, including: scanned or keyed from paper forms, as files on a floppy disk or CD-ROM, magnetic tape, light pen or bar-code readers, microphone, digital camera, communication and phone lines, or via Internet e-mail and Internet databases.

Processing unit. The Central Processing Unit (CPU) is the heart of a computer system. In most modern computers the CPU consists of just one or two silicon chips that are small enough to hold in one hand, but which contain many millions of logic circuits. A CPU would typically execute millions of instructions per second.

Associated with the CPU is the read access memory (RAM) memory. The RAM has to be big enough to hold all programs and data that are being worked on at a given time. RAM size ranges from a few million bytes to a few hundred million bytes.

Data storage. Disk drives are used on almost all computers to hold data. A current PC might have one 2.0 GB disk drive (two billion bytes), while large computers might have a number of disk drives holding tens or hundreds of GB.

Data can be copied onto magnetic tapes or CD-ROMs for backup, transfer between computers, or long term storage. A CD-ROM can hold 600 MB, while tape units can range from about 500MB to 24GB. Data can also be copied onto Diskettes (floppy disks), but their small size (1.44MB) limits their usefulness with large data sets.

Output devices. Output devices include video monitors, various sorts of printers, magnetic disks and tapes, CD-ROMs, data communication and phone lines, and stereo speakers. Computers can also be used to drive industrial processes, control chemical plants, and lock/unlock security doors. Modern car engine management systems, office lifts, VCRs, and numerous other domestic and industrial systems are now controlled by miniature computer systems.

STORAGE AND RETRIEVAL

Much of the computer's power comes from its ability to store, sort and classify data. Over the past few years, disk systems have become very cheap and reliable, and it is now possible to obtain disk systems that will store billions of bytes of data for just a few hundred dollars. This thousand-fold drop in data storage price, and the development of sophisticated database systems has greatly improved the usefulness of computers. It is now possible to hold all company information, going back for years, on a computer. Laws are being changed to drop the legal requirement for paper storage.

SOFTWARE

There are two basic types of computer software: **systems software** which controls the operation of the computer, and **applications software** which performs useful tasks for the user. Computers are purchased in order to run the application software, while system software assists to make this job easy and convenient.

Systems software: is dividied into two classes, operating systems and tools and utilities.

An **operating system** typically has two levels:

The *lower level basic input-output system (BIOS)* controls the most basic functions, such as: reading and writing RAM (memory), and input to and output from peripherals such as mouse, keyboard, printer and screen.

The *higher level main operating system* (eg Windows) acts as a platform to host

programs. It provides the user interface to control the computer's operation, and the environment to effectively operate application software. For example, it provides a file sub-system with its structure of drive names, directories, folders, files, and indexes; and file handling facilities such as creating, copying and deleting.

Typical *operating systems* are DOS, Unix, Mac OS and Windows 98. Mac OS and Windows 98 have a "user friendly" Graphical User Interface (GUI) which enables computer control by means of windows, menus, icons and a mouse. DOS and UNIX require the user to type precise commands, which can be hard to remember. Windows 98, Mac OS and Unix can run many programs at one time (multiprogramming), which makes for more efficient computer hardware use and user convenience.

Tools and utilities are usually necessary to make productive use of a computer. Some are provided with the operating system and others are purchased separately. Typical system utilities are Internet browsers, anti-virus software, program compilers, editors and file backup systems.

Applications Software can also be divided into two classes: personal productivity tools and other computer applications.

Personal productivity tools are commercial products designed to handle standard computing tasks such as word processing, numerical analysis, data manipulation and storage, and data presentation. Typical products are:

- *Word processing:* Word processing software is designed for the creation of documents: letters, reports, newspaper articles and books. They were one of the first applications for personal computers, designed to streamline large amounts of routine typewriting. They succeeded because they allowed text to be edited without having to retype the whole document. MS Word, MS Works, Word Perfect, and Word Pro are typical word processing products.

- *Spreadsheets:* These packages are an electronic development of accounts used by bookkeepers to organise business information, and are very useful for handling tabular data. Addition, subtraction, division, multiplication and totalling can be done very quickly, and all results can be automatically recalculated later if new data is inserted. Formatting and graphing facilities are used to aid analysis and presentation. Hundreds of functions are included to enable typical statistical, engineering, economics and business calculations to be performed automatically. Examples are functions for the calculation of compound interest and standard deviation. Typical Spreadsheets packages include MS Excel, MS Works and Lotus 123.

- *Databases:* Database packages are a convenient way of organising and storing data in a uniform fashion. Data can be quickly and systematically searched, sorted and presented. They can be used by people with no special training to create mailing lists or record store inventories, but they can also be used by professional programmers to produce complex applications to assist with

running a business. MS Access, MS Works and Lotus Approach are typical database packages.

- *Presentation:* Presentation packages are used to illustrate talks and lectures. Presentation packages have almost replaced hand-drawn or typed overhead projector slides. In many presentations and lectures given in industry today, the presenter plugs a Notebook computer directly into a projector to show slides on a screen. MS Powerpoint and Lotus Freelance are typical presentation packages

- *Graphics:* Graphics packages enable any user to create drawings, paint pictures and enhance or manipulate scanned pictures. MS Paint is a typical graphics package.

- *Desktop publishing:* These packages are intended to enhance the final appearance or layout of text and graphics to make them suitable for publishing. Graphics can come from a library of clip art be created by drawing or paint package, or be a scanned image such as a photo. MS Publisher is a typical desktop publishing package.

Other Computer Applications: The above "personal productivity tools" are commonly used by most computer users at home, school and in the office. They are fairly cheap and available on many computers. However, businesses and organisations usually buy computers to automate major business functions, and this is not usually done on personal productivity software.

Some software applications can cost many thousands of dollars to buy or develop, while a major banking or airline reservation system could cost millions. Application software can either be purchased "off the shelf" or developed for a specific purpose. An accounting package is a typical example of a purchased application, while a system to handle parking fines might be designed and written from scratch.

A database application could be used for retail stock control: recording sales and setting up replacement orders. An airplane's autopilot navigation system is an example of software that receives information (eg. compass heading and global positioning system [GPS] position) and outputs data that controls rudder and flap hydraulics which adjust the course. If it is decided to develop software to automate a task, the work is done by systems analysts and programmers.

SYSTEMS ANALYST

System analysis is the process of breaking down a data processing problem into functional components to determine the best method of handling the problem.

The systems analyst must, with consultation:

- define the system problems of an organisation,
- analyse the results of the investigation to determine new system requirements,
- design a new system that is practical, efficient, cost effective and makes best use of available hardware and software,
- communicate the new system to all parties concerned, and
- assist in implementing the new system.

PROGRAMMER

Programming is the process of producing a set of instructions to make a computer perform a specified activity. The programmer takes system analysis results and develops computer programs to solve the problem. A programmer must:

- understand problems and plan solutions,
- design programs using data flow diagrams and other design tools,
- write programs to implement the design,
- test programs and correct any problems,
- write detailed documentation of programs and their operation.

USER

The user is the final judge of whether a computer system is meeting the needs it was designed to fulfil. The better the link between automated system component and user, the more likely it is that the system will be effective. Modern system designers consult widely with the user in order to design systems that meet user needs, and put considerable effort and ingenuity into designing the interface between a system and its users.

EXERCISES

1. Computer systems are a combination of three ingredients; what are they?

2. Name three hardware components.

3. Give two reasons why many large organisations set up their own internal intranet system.

4. Why would someone doing data analysis find spreadsheets useful?

5. Why is it important that a systems analyst consult with the system user before trying to design a new system?

INFORMATION - USE IN SOCIETY

You have probably heard of the term *the information age*. Modern society has come to depend more and more on information. In many fields such as politics, economics, environment and entertainment, information is relied upon to help make decisions or recommendations.

This section will outline some examples of how information is used by people and organisations. Without reliable information wrong decisions can be made. In some cases the consequences of a wrong decision can be serious.

CASE STUDY 1

QUIT SMOKING CAMPAIGNS

The Victorian Smoking & Health Program or Quit Campaign has been working since the 1980s to encourage Australians to give up smoking. This organisation makes use of statistical information to decide particular campaign strategies. Details of two such campaign strategies and the statistics that influenced them follow.

1. Between the years 1983 and 1987, survey data was collected on the characteristics of young adults (16-19 years) who smoked. Some of the major findings of the surveys were that:

   - smoking rates among students were 11.6%;
   - smoking rates among young full-time workers were 45.7%; and
   - smoking rates among young people looking for work were 47.4%.

Considering these major findings, Quit Campaign organisers decided that a campaign should be targeted toward young adults in the workforce.

To design the campaign, data were collected from young people at their place of work. As a result of this research, Quit Campaign organisers were able to conclude the following (among other things) before designing their campaign:

- smoking was very much a life-style issue for young adults at work. It was tied in with social activities and friendship groups of this age; and
- smoking patterns change from school to work: a transition from occasional to habitual smoker tends to occur.

2. Quit Campaign data (1988-90) showed smokers over the age of 50 were least likely to believe health facts about smoking and least likely to intend quitting. The Quit Campaign's annual household and telephone surveys found:

   - 19% of people aged over 50 years who smoked believed they had *no personal risk of dying* from smoking;
   - this compared to 3% for people aged 16-29 and 6% for people aged 30-49; and
   - people aged 50 plus were most likely to report that no one favoured their quitting.

   As a result of the above information, Quit Campaign organisers launched a campaign in 1994 designed to increase awareness of Quit's services in smokers over 50 years of age. The hope was that the numbers intending or making attempts to quit would increase significantly.

3. The 1995 Victorian Quit campaign targeted people aged 18-24 years, based on 1994 US Surgeon General research on preventing tobacco use among young people and 1993 research on the incidence of Victorian secondary school students smoking.

   Identified groups included: established smokers aged 18-24, experimental smokers aged 12-19, and potential smokers aged 12-19. Post-campaign research found that the advertisements were seen by 89% of 16-24 year olds, and that 39% of this group's smokers felt encouraged to quit.

CASE STUDY 2

CAR POOLING

In 1991 the Royal Automobile Club of Victoria (RACV) wanted to know if their staff would be interested in a car pooling scheme. Their head office in Melbourne is difficult to reach by public transport and their car park had room for 600 cars with a staff size of 650.

Consultants were employed to survey staff about their attitudes to car pooling. A sample of 80 staff were asked about car pooling and what would motivate them to join in such a scheme. Some of the major findings of the survey were that:

- 27% of men and 28% of women would be pleased to car pool,
- 43% of men and 55% of women had mixed feelings about car pooling, and
- 26% of men and 15% of women were unhappy about car pooling.

In general, the survey recognised that most staff preferred to travel to and from work alone by car. A high proportion believed they needed to travel in their own cars at least one day per week. However, the survey also revealed that staff did see benefits in a car pooling scheme. These benefits in priority order were:

- social benefits,
- environmental benefits, and
- personal benefits.

The RACV introduced a car pooling scheme as a result of the survey. The scheme has proved effective in meeting the needs of interested employees and reduced demand for car parking space. Indeed, it is hoped the success of the scheme will encourage other large employers to introduce their own car pooling schemes.

CASE STUDY 3

HOUSEHOLD EXPENDITURE AND MARKETING DECISIONS

The Australian Bureau of Statistics conducts an extensive survey of how Australian households spend their money. While the main purpose of this survey is to measure change over time in the cost of living, the survey also allows businesses to gain an understanding of the type of people who buy their products. In turn, this allows businesses to develop better advertising and marketing strategies.

In other words, information on household expenditure is valuable market research for any business.

1. A meat products promotional body wanted to find out the levels of expenditure on different types of meat products by households with differing income levels. The ABS survey covered 18 different types meat products, and the income spent upon them by households. The promotional body found that:

   - high income households spent far less on its type of product compared to other meat products.

   As a result of the above, the promotional body redesigned its product for the high income market.

   It also used expenditure information to produce recipe ideas that would promote its product to specific markets in each state.

2. A home delivery pizza company wanted to identify the best location at which to open a new shop. The company examined ABS information on household expenditure on take-away food and eating out. Information was provided to the company on a suburb by suburb basis.

   This allowed the company to establish a new shop in an area with high take-away food expenditure.

CASE STUDY 4

OZONE LAYER DEPLETION AND THE MONTREAL PROTOCOL

The ozone layer is an important part of the global atmosphere-climate system. It limits the amount of ultraviolet radiation from the sun to levels necessary for life on Earth. A depleted ozone layer is likely to have serious consequences, such as: increased rates of sunburn and skin cancer, eye damage and other diseases, and reduced plant growth.

Human developed chemical compounds are the main cause of ozone layer depletion. These are compounds such as chlorofluorocarbons (CFCs) and halons, among others. In the past, they had been commonly used in refrigerators, air-conditioning systems and fire-retardant chemicals. In general, a fall of 1% in atmospheric ozone is equivalent to an increase of 1-2% in UV radiation at ground level.

The problem of ozone layer depletion became prominent in the 1980s, as scientific measurements began to show significant global decreases in ozone. Some of the general results follow:

- for mid latitudes, Europe and North America, annual ozone losses of 2-4% over the 1980s were reported,
- for Australia, ozone losses during the 1980s ranged from 5% over Hobart to 0.5% over Darwin, and
- for Antarctica, the ozone hole has become a regular feature of each southern hemisphere spring with total ozone losses of 60-70% reported since 1985.

The seriousness of the problem has led to global agreement to reduce and control the production of ozone depleting substances. At Montreal in 1987, 149 countries signed an agreement to reduce the use of ozone depleting substances. Some decisions taken were:

- for CFC-11s: freeze consumption at 1986 levels by 1989, and
- for CFC-12s: reduce consumption by 20% by 1 July 1993.

Australian efforts to meet Montreal targets can be seen below.

Australian domestic use of major ozone depleting substances (tonnes)			
	1986	1989	1992
Total CFC	14,633	14,293	5,540
Car air-conditioning CFC	1,765		2,372
Halon-1211	690		15
Halon-1303		220	39
Methyl chloroform		8,537	4,680 (a)
(a) Estimated			

There are probably thousands of decisions made every day based on statistical information. Some are complex and require an in-depth study of the statistics before any decision is made. Others are straightforward and can be made with a quick look at the statistics.

The preceding examples and those below should make you appreciate just how important a role statistics play in modern society.

In January 1986 the space shuttle *Challenger* blew up killing those on board. After the accident, NASA re-examined the joints in the shuttle's booster rockets. Each joint had better than 97% reliability. However, probability calculations showed that six joints working together were much less dependable. Such probability calculations played a major role in identifying the problem and getting the shuttle back into space.[1]	

A leading advertising agency analysed ABS information from the 1991 Census. They were particularly interested in the average number of children in young families in Australia. After making certain assumptions they calculated that the average young family in Australia has 2.3 children. This finding was the basis of a national television advertising campaign about a car aimed at Australia's young families.	

In the early 1980s over 300 Hispanic agents took the FBI to court in America. They claimed the FBI was discriminating against them over promotion, and in the hiring and firing of agents. To support their case they presented statistics on the promotion and hiring and firing of FBI agents. Their presentation persuaded the judge to order the FBI to adopt new policies to correct the problem.[2]	

EXERCISES

1. Choose a night and a particular television station to watch the news. List the first ten news stories and note how many had statistical information as part of the story. Were any decisions in the stories based on statistics?

2. Compare the front pages of three Australian newspapers on the same day. Count the number of different stories or reports on each front page, and calculate for each the proportion of stories that mention statistics. Do the same thing over a period, say once a week for four weeks. Present the results in the form of a table.

3. Imagine you are a politician who wants to lower the voting age of the population from 18 to 16. What statistical information might you use to argue your case? Would you argue that the decision in favour of lowering the voting age should be based on the statistics alone?

4. Draw up a list of what you would take into account, other than statistics, when making a decision about:

    a) increasing taxation,
    b) reducing traffic congestion,
    c) quitting smoking,
    d) buying a computer,
    e) using public transport, and
    f) moving interstate.

5. Decide your favourite city in Australia or the world and use one item of statistical information to argue in favour of your choice.

6. Can you think of situations where the same statistical information could be used to justify opposite decisions?

7. Write an essay entitled 'Using Statistics in today's society'.

Footnotes 1 & 2 (previous page). From *Statistics - Decisions through Data*. Video produced by the Consortium for Mathematics and its Applications (COMAP) Inc, America. Available from the Australian Association of Mathematics Teachers Inc., GPO Box 1729, Adelaide, SA 5001.

8. Carefully study the following table on world population figures and answer the questions afer it.

UNITED NATIONS WORLD POPULATION FORECAST		
	(millions)	
	1996	**2050**
China	1232	1517
India	945	1533
Pakistan	140	357
Nigeria	115	338
Indonesia	200	318
Iran	70	170
USA	269	347
Ethiopia	58	213
Brazil	161	243
Bangladesh	120	218
Kenya	28	66
Mexico	93	154
Russian Federation	148	114
Philippines	69	130
Uganda	20	66

Source: United Nations Dept for Economic and Social Information & Policy Analysis. *World Population*. 1996.

a) Which organisations might want the above information?

b) To what issues would the above information be relevant?

c) Based on these issues, what decisions would you take considering the above information? (Discuss as a class.)

INFORMATION - PROBLEMS WITH USING

The previous section should have given you an idea of just how important statistical information is in modern society. Decisions that affect the lives of all Australians are often made by taking statistics into account. This places a large responsibility on people who make decisions. They should be aware of the traps one can fall into when using statistics.

This section will outline some of the problems you may encounter if you are not careful in using statistics. The quotation below from H.G. Wells was made at the beginning of the 20th century, and few would disagree that it is relevant today. The modern citizen needs to have an awareness of the problems with using statistical information.

> "STATISTICAL THINKING WILL ONE DAY BE AS NECESSARY FOR EFFICIENT CITIZENSHIP AS THE ABILITY TO READ AND WRITE."
>
> H. G. Wells

MISINTERPRETATION OF STATISTICS

Misinterpretation is a good example of a common problem in the use of statistical information. It may be caused by a number of factors such as:

- *Ignoring definitions.* You should always familiarise yourself with the definition of concepts surrounding statistical information you are using. If you are examining labour force issues, you should familiarise yourself with the definition of unemployment, participation rate, etc. If you are examining environmental issues, you should consider the definition of forest, woodland, extinct or endangered species, or even the definition of a National Park (which differs between States). An example of how ignoring a definition can lead to misinterpretation of data follows:

  The ABS released a labour force publication in November 1992 with the following main feature:

  > "AN ESTIMATED 25 PER CENT OF ALL FAMILIES HAD NO FAMILY MEMBER EMPLOYED."

  Based on the above, a headline in a leading Australian newspaper read:

  > **"UNEMPLOYMENT AFFLICTS ONE IN FOUR FAMILIES."**

  This headline does not logically follow from the main feature above it. The headline represented a lack of understanding about the definition of *unemployed.* If you are not employed you may be *unemployed:* that is, in the labour force and actively seeking a job; OR, you may not be in the labour force, for example, you may be a student, retired or not actively looking for work.

  Just because a family has no member employed does not necessarily mean that those members are *unemployed,* because to be unemployed you have to be in the labour force: that is, you have to be actively seeking work. The headline showed a misinterpretation based on lack of understanding of an underlying definition.

- *Comparing statistics inappropriately.* A great advantage of using statistics is that one can compare information to assess beliefs, ideas or thoughts about issues and topics. For example, you can compare: Sydney's weather with Melbourne's, past sporting results with the present, or whether males and females do the same amount of unpaid household work.

  However, there can be real problems in comparing statistics when the definitions, classifications or methods of collection underpinning them are *different*. Nowhere is this more apparent than with environmental statistics. Consider the table below.

FOREST COVER, AUSTRALIA, 1980	
Source	*Per cent of Australia*
CSIRO	4
WORLD BANK	14

  The definitions of forest used by CSIRO and World Bank in the above table are very different. The World Bank has included 'woodland' in their estimate, and this explains the large difference in figures. Therefore, it is wrong to compare the two figures in any way! It would be worse still to compare a World Bank estimate for 1980 with a CSIRO estimate for 1981, and conclude that Australia had logged most of its forests!

- *Deliberate misrepresentation.* In the modern information age, it is certainly important to recognise that information can have integrity, and be objective, accurate and factual. However, it must also be recognised that information can sometimes be flawed; by being subjective, inaccurate or fictional. Consider the quotation below:

> "POLITICAL TACTICIANS ARE NOT IN SEARCH OF SCHOLARLY TRUTH OR EVEN SIMPLE ACCURACY. THEY ARE LOOKING FOR AMMUNITION TO USE IN THE INFORMATION WARS. DATA, INFORMATION, AND KNOWLEDGE DO NOT HAVE TO BE TRUE TO BLAST AN OPPONENT OUT OF THE WATER."
>
> Alvin Toffler

You might say this is an overly cynical quotation, but one does need to realise that information is open to manipulation by various forces, for example:

- According to a United Nations report, in 1986 the South African authorities ceased publishing information on South Africa's imports and exports classified by countries that supplied and received them. This was an attempt to head off trade sanctions being imposed because of apartheid policy.

This section outlined some problems you may encounter trying to understand and compare statistical information. Of course, you also have to be careful about how accurately statistics were collected in the first place. This leads to you being aware of sampling and non-sampling error, concepts outlined in the following pages.

SAMPLING ERROR

In any sample survey that you undertake you will experience sampling error. Sampling error refers to:

> THE DIFFERENCE BETWEEN AN ESTIMATE DERIVED FROM A SAMPLE SURVEY AND THE 'TRUE' VALUE THAT WOULD RESULT IF A CENSUS OF THE WHOLE POPULATION WAS TAKEN.

Sampling error can be measured mathematically and is influenced by:

Size of sample. In general, the larger the sample size (the number of people being surveyed) the smaller the sampling error.

Many people are surprised by the small size of well-known sample surveys. Opinion polls about which party people will vote for are taken with sample sizes ranging from 600 to 2,000 people, with samples of about 1,000 the most likely. Television ratings of different programs and channels are taken from a sample survey of about 1,900 homes, out of an Australian total population of 6.5 million homes. Despite a perception that such polls are accurate, some statisticians would question their accuracy due to the small sample sizes.

Design of sample. The method of sampling can also affect the size of sampling error. This concept is looked at in detail on pages 175-182.

NON-SAMPLING ERROR

This concept refers to error apart from sampling error. Non-sampling error can occur at any stage of a sample survey or census, and unlike sampling error it is not generally easy and inexpensive to measure. There are two main types of non-sampling error: *systematic error* and *variable error.* Variable error is less serious than systematic error because, on average, it tends to balance out . Systematic error does lead to distortion of survey results, so it is important to be aware of how it occurs.

SYSTEMATIC ERROR (BIAS)

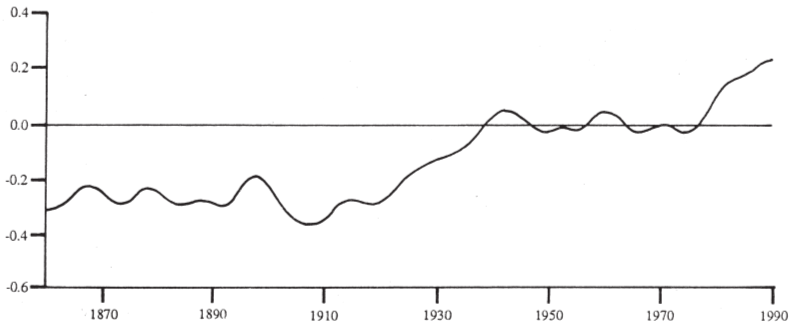
Later in this publication you will come across a technical definition of bias (page 179). For the purposes of this section, bias is defined as any influence that unreasonably affects or sways the results of a sample survey or census. There are a number of different sources of bias:

- *Inappropriate estimation.* The ABS and other data collection agencies spend much time designing and monitoring sample surveys to ensure that non-sampling error is kept to an absolute minimum. However, even after sample survey results are finalised, bias can be introduced. This occurs if estimation (see page 186) is inappropriate.

  An example of inappropriate estimation relates to the issue of global warming (greenhouse effect). The graph below is the most common portrayal of global temperature change. In general, it shows an average increase in the last 160 years of between 0.3 Celsius and 0.6 Celsius in global temperatures.

  However, some scientists have questioned the accuracy of this chart. This is because they feel that estimates from the sample survey are biased.

CHANGES IN GLOBAL TEMPERATURE
(Degrees Celsius)



Source: United Nations Working Group

The measurements that make up the graph have been taken at various weather stations around the world. You can regard the world's surface as the *population* from which a sample survey can be taken.

Scientists argue, therefore, that measurements should be taken to reflect the ratio of the world's land mass to its sea mass. For example, if the land mass is half the sea mass, then twice as many measurements should come from the world's seas as opposed to the land.

In fact, in the graph above, there have been very few measurements taken from the world's sea surfaces, whereas the great majority of measurements were taken from weather stations on land.

But why might this bias the estimates from the sample survey? The reason is that temperatures on land tend to be naturally higher than on sea surfaces. This is due to a phenomenon known as *urban heat island effect.* Hence, if the sample is too heavily weighted towards land based temperatures, and the estimates do not take account of this (as some scientists claim), the results may not reflect a true global average.

- *Poor questionnaire design.* You should always be careful with the questions you ask in a sample survey or census. Otherwise, bias may be introduced. If questions are leading, misleading, ambiguous or difficult to understand, the survey or census results may be distorted.

  An example of poor question design is shown by a pilot test the ABS conducted for the 1986 Census of Population and Housing. A pilot test checks that questions in a forthcoming census will be easily understood.

  For the 1986 Census it was requested that the ABS gather data on ethnicity. Initially a question: *'What is your cultural background'* was framed.

  One of the replies to this question was simply *'none'*. When the respondent was contacted he was asked what he meant. He replied, *'Look, leave me alone, I'm a regular sort of a bloke, I go to the footy every now and then, but I've never been to the opera and I've never taken up a musical instrument in my life.'*

  This example shows that people may interpret broad concepts such as 'cultural background' quite differently. The question was reframed and written as: *'What is each person's ancestry?'*. E.g. Greek, Armenian, English... etc.

- *Non-response bias.* If a significant number of people do not respond to a mail-out for a sample survey, then results may be biased. This is because the characteristics of non-respondents may differ from those who have responded. Some questions may be difficult to understand for certain people.

To reduce this form of bias, care should be taken in the design and testing of questionnaires, and following up non-respondents to a survey.

- *Interviewer bias.* An interviewer can unfairly influence the way a respondent answers questions. This may occur if the interviewer is too friendly, aloof or prompts the respondent. Interviewers therefore need to be trained correctly (see page 26).

- *Processing errors.* These can arise through miscoding, mispunching, incorrect computer programming and inadequate checking (see data processing, page 29).

SUMMARY

It is useful to have a checklist of questions ready for whenever you are presented with statistical information. This is not because there are always going to be problems with the statistics, but rather because it will give you confidence in judging their reliability. Some questions you might ask include:

- What is the source of the information? Is it from a primary source (organisation that collected the data) or a secondary source?

- If the information is from a secondary source, is it possible it may have been altered for whatever reason?

- Has the primary source of information got a possible reason for misrepresenting the information?

- Do you need to find out the method of data collection, sampling technique or response rate to the survey? Were the questions asked easy to understand?

- If the information is from a sample survey, was the sample size adequate?

- Do you understand the definitions of variables or topics talked about in the survey or census? Are definitions consistent?

These are just some questions you may consider when presented with statistical information. You may feel that some of them would be difficult to answer, but if the source cannot provide you with answers, then the information's reliability should be questioned!

EXERCISES

1. Can you list some possible problems with the statistics in the following statements?

   a) The average income of Australians is $83,000 according to a survey carried out in the Sydney suburb of Double Bay.

   b) A large majority of rural people oppose dropping the wool floor price according to a television phone-in poll carried out by a regional television station.

   c) Tests reveal that half (50%) of our nation's school leavers are below average in reading and writing.

   d) Youth unemployment is over 30%; therefore, 30% of Australia's 15-19 year olds are unemployed.

   e) A leading environmental group recently claimed that only 3% of Australia's land mass was covered by forest, whereas a leading business organisation claimed the figure was 7%.

2. Examine the statistics you have gathered from Exercise 2 on page 55 and list some problems you think might exist with the information.

3. As far as presenting and debating ideas, statistical information can have limitations: it often needs to be explained or interpreted with words. Conduct a class debate about the above sentence or about the quotation below!

"ORATORY IS DYING, A CALCULATING AGE HAS STABBED IT IN THE HEART WITH INNUMERABLE DAGGER-THRUSTS OF STATISTICS."
Sir Keith Hancock

STATISTICS - PRIVACY AND SECURITY

As modern society comes to depend more and more on information, new problems of individual rights and privacy arise. People want information about many things, such as the latest figures on jobs for school leavers, or up to the minute accuracy on personal bank account balances from automatic teller machines. At the same time, many people are concerned that too much data about individuals is being stored on computer databases and accessed by persons or organisations unknown to them.

Writers and film-makers have painted terrifying pictures of futuristic societies where the thought and action of human beings are controlled by all-powerful computers. Such fears are often exaggerated in the interests of good fiction. However, there are important issues that society must handle to ensure that rights of the individual are protected in the information age. This section discusses some of these issues.

PROVIDING INFORMATION

In daily life, nearly all people provide information about themselves to many different organisations. To get borrowing rights from a video shop or local library it is usually necessary to complete a personal particulars form. Taxpayers and users of government services must provide details about themselves to government.

Likewise, banks and big retailers will only issue credit cards to a customer if they know something about the income, occupation, family status and other details of that customer. Health services often collect and store a lot of data about each client they treat.

These are just a few examples of where the individual is required to provide personal information, much of which is entered into computer databases. There is a concern that it may be possible for authorities, commercial organisations and others to access and link such databases. In this way, data profiles (or information pictures) on individuals could be put together and perhaps used in a way that disadvantages the individual.

If one accepts the principle of personal privacy, then it follows that personal information should not be used for a purpose other than that which its collection was authorised, without the permission of the person to whom the information relates.

There is also concern that individual privacy and corporate confidentiality may be breached by *hacking* into computer databases. 'Hacking' usually applies to computer users who gain unauthorised access to large databases, and even amend the datafiles. Such activity is highly illegal.

> "The need for the protection of privacy is more evident in modern Australian society than ever before. Increasing powers given to public officials to intrude on people's lives, new intrusive business practices such as credit reporting and direct mailing, and new computer and surveillance technology which allows information to be manipulated, matched, compared and profiled both locally and internationally, have altered the level of control that individuals have over their own affairs."
>
> Australian Geographic Society (Sydney, 1988)

These concerns about information privacy are legitimate; however, there are many beneficial aspects to the provision of information. Earlier sections have emphasised the role of information in the decision-making processes of society. Information is necessary for many aspects of modern society to function efficiently. Therefore, people who provide information need to have confidence that their privacy and security are protected.

PRIVACY AND SECURITY STEPS

Today's young people will need more data, and greater skills in data handling, to assist in making decisions in the years ahead. As the information age continues to expand, it is also important to remember that the desire for privacy remains an essential issue to be considered.

In Australia, government and community organisations are increasingly taking steps to ensure that privacy of statistical information can be safeguarded in the computer-based information age. Some examples of these steps are:

- the Commonwealth Privacy Commissioner's examination of data exchange arrangements between government agencies;
- groups such as the Australian Privacy Foundation and other community watchdog organisations have been working to ensure that individual rights and liberties are not unduly compromised by large scale data collections;
- holders of data are encouraged to maintain tight security procedures, so that only those with the correct authorisation can access databases;
- much more secure telecommunications links are now available for transmission of data between different locations of a particular government agency or business organisation; and

- tighter procedures are also in place to counter the spread of computer viruses, whereby important information can be destroyed or distorted by a piece of rogue programming introduced into a computer network.

ABS, PRIVACY AND SECURITY

As the national statistical agency, the Australian Bureau of Statistics takes measures to ensure that the privacy and security of data provided by individuals and organisations is carefully protected.

These measures involve:

- strict access controls to databases where information is stored: even for people working at the ABS, passwords are needed for entry to specific databases;

- strengthened security arrangements on buildings from which the ABS operates: access to floors where census and survey forms are processed is denied to the general public;

- for household based surveys and censuses, the ABS does not store names and addresses of people on its databases;

- forms provided by individual respondents to ABS census and survey collections are destroyed under careful supervision once processed;

- care is taken by the ABS to ensure that no individual respondent can be identified from the statistics it publishes; and

- the Census and Statistics Act sets stiff penalties for any officer or employee of the Bureau who divulges confidential data without authority.

The ABS and its staff have a strong commitment to ensuring that privacy rights of individuals and confidential affairs of organisations are not compromised through the collection and publication of official statistics.

EXERCISES

1. List the various organisations where personal data about yourself might be stored. List some of the uses that might be made of the data. Which uses do you think are appropriate? Are there any that you consider inappropriate?

2. Does your school hold data and information about you on computer? If so, what sort of data and information is available for you to look at? Is any of the data or information private, and if so, why?

3. Imagine you are in charge of computer security in a government agency or commercial organisation. What steps would you take to ensure that personal information held in the organisation's databases is not accessed without authority?

4. Do you think computers will ever be more powerful or intelligent than human beings? Discuss as a class.

STATS MATHS

ORGANISING DATA

After collection and processing, data need to be organised to produce useful information. It helps to be familiar with some definitions when organising data. This section outlines those definitions and provides some simple techniques for organising and presenting data.

VARIABLES

The word *variable* is often used in the study of statistics and so it is important to understand its meaning. A *variable* is:

DEFINITION

ANY TRAIT THAT IDENTIFIES DIFFERENT VALUES FOR DIFFERENT PEOPLE OR ITEMS

Height, age, amount of income, country of birth, grades obtained at school and type of housing are examples of variables. Variables may be classified into various categories, some of which are outlined in the following pages.

NOMINAL VARIABLES

A nominal (also called categorical) variable is one that describes a name or category.

EXAMPLE

1. The method of travel to work by people in Darwin at the time of the 1996 Census was:

Method of travel to work	Number of people
CAR AS DRIVER	23,617
CAR AS PASSENGER	3,699
BICYCLE	1,335
WALKED	1,703
BUS	1,335
WORKED AT HOME	1,012
MOTOR BIKE/SCOOTER	577
TAXI	284
TRAIN	25
FERRY/TRAM	11

In this case the variable ‘method of travel to work’ is nominal because it describes a name.

NUMERIC VARIABLE

A numeric variable is one that describes a numerically *measured* value. However, not all variables described by numbers are numeric. For example, the age of a person is a numeric variable, but their year of birth, despite being a number, is a nominal variable.

Numeric variables may be either continuous or discrete:

CONTINUOUS VARIABLE

A variable is said to be continuous if it can take *any value within a certain range.* Examples of continuous variables may be distance, age or temperature.

The measurement of a continuous variable is restricted by the methods used, or by the accuracy of the measuring instruments. For example, the height of a student is a continuous variable because a student may be 1.6321748755... metres tall.

However, when the height of a person is measured, it is usually only measured to the nearest centimetre. Thus, this student's height would be recorded as 1.63m.

Note that continuous variables are usually grouped using *class intervals* (explained shortly). They are grouped to make them easier to handle as part of the general process of organising data into information.

DISCRETE VARIABLE

Any variable that is not continuous is discrete. A discrete variable can only take *a finite number of values within a certain range.* An example of a discrete variable would be a score given by a judge to a gymnast in competition: the range is 0-10 and the score is always given to one decimal place.

Discrete variables may also be grouped. Again, this is done to make them easier to handle.

NOTE: measurement of a continuous variable is always a discrete approximation.

ORDINAL VARIABLE

An ordinal variable is one that can be placed in order. Numeric variables are always ordinal, while only some nominal variables are ordinal.

1. A teacher may rank a class of students in order according to their behaviour:

Behaviour	Number of Students
Excellent	5
Very Good	12
Good	10
Bad	2
Very bad	1

In this case the variable 'behaviour' is *nominal* and *ordinal.*

FREQUENCY DISTRIBUTION TABLES

The frequency (f) of a particular observation is *the number of times the observation occurs in the data.* The distribution of a variable is *the pattern of values of the observations.*

Frequency distribution tables can be used for both nominal and numeric variables. (For continuous variables they should only be used with class intervals, explained on the next page.)

EXAMPLE

1. Twenty people were asked how many cars were registered to their households. The results were recorded as follows:

   1, 2, 1, 0, 3, 4, 0, 1, 1, 1, 2, 2, 3, 2, 3, 2, 1, 4, 0, 0.

   Present this data in a frequency distribution table.

Number of cars (x)	Tally	Frequency (f)
0	IIII	4
1	~~IIII~~ I	6
2	~~IIII~~	5
3	III	3
4	II	2

A tally mark is placed in the appropriate row in the table as the data are read from left to right.

The first result is a '1', so a tally mark is placed in the row where 1 appears in the **'number of cars'** column in the table.

The next result is a '2', so a tally mark is placed in the row where 2 appears in the **'number of cars'** column, and so on.

The fifth tally mark is drawn through the preceding four marks to make final calculations of frequency easier.

Thus, it can be seen that the number of households with no car is 4, the number of households with 1 car is 6 and so on.

CLASS INTERVALS

When a variable takes a large number of values it is easier to present and handle the data by grouping the values in class intervals. Continuous variables are always presented in class intervals; discrete variables can also be grouped and presented in class intervals. In the example below, we set out age ranges for a study of young people, but allow that some older people may fall in-scope for our study.

The *frequency* of a class interval is the *number of observations* that occur in a particular pre-defined interval. If 20 people aged 5-9 appear in our result, the frequency is 20 for this interval.

The end-points of a class interval are the lowest and highest values that a variable can take. Therefore, if the intervals are 0-4 years, 5-9, 10-14, 15-19, 20-24, and 25+; the end-points of the first interval are 0 and 4 if the variable is *discrete*, and 0 and 4.999 if *continuous*.

Class interval *width* is the *difference between lower end-point of the interval and lower end-point of the next interval*. If the intervals (continuous) are 0-4, 5-9, ... , etc.; the width of the first 5 intervals is 5, and the last interval is open. The intervals could also be written as 0-<5, 5-<10, 10-<15, 15-<20, 20-<25, and 25+.

The basic rules to follow when constructing a frequency distribution table for a data set containing a large number of observations are:

- find the lowest and highest values of the variable,
- decide on the width of the class intervals, and
- make sure that all possible values of the variable are included.

EXAMPLE

1. Thirty AA size batteries were tested to determine how long they lasted. The results, to the nearest minute, were recorded as follows:

423, 369, 387, 411, 393, 394, 371, 377, 389, 409, 392, 408, 431, 401, 363, 391, 405, 382, 400, 381, 399, 415, 428, 422, 396, 372, 410, 419, 386, 390.

Construct a frequency distribution table.

The lowest value is 363 and the highest value is 431.

For the given data, and choosing a class interval of 10, the first class interval should be 360-369 to include 363 (the lowest value). There should be enough class intervals until the highest value has been included to give the following table:

Class Interval (x) (Battery life, mins)	Tally	Frequency (f)
360-369	II	2
370-379	III	3
380-389	𝍸	5
390-399	𝍸 II	7
400-409	𝍸	5
410-419	IIII	4
420-429	III	3
430-439	I	1
Total		**30**

RELATIVE AND PERCENTAGE FREQUENCY

Analysts studying this data may not be only interested in how long batteries last, but also what proportion fall in each class interval.

The relative frequency of a particular observation or class interval is found by dividing the frequency (f) by the number of observations (n): that is, (f/n). Thus:

RELATIVE FREQUENCY	= FREQUENCY ÷ NUMBER OF OBSERVATIONS

The percentage frequency is found by multiplying each relative frequency value by 100. Thus:

PERCENTAGE FREQUENCY	= f/n x 100

1. Using the previous example of battery life, set up a table giving the relative frequency and percentage frequency of each interval.

EXAMPLE

Class Interval (x) (Battery life, mins)	Frequency (f)	Relative frequency	Percentage frequency
360 - 369	2	0.07	7
370 - 379	3	0.10	10
380 - 389	5	0.17	17
390 - 399	7	0.23	23
400 - 409	5	0.17	17
410 - 419	4	0.13	13
420 - 429	3	0.10	10
430 - 439	1	0.03	3
Total	**30**	**1.00**	**100**

The analyst might now be able to say that:

- 7 per cent of AA batteries have a life from 360 minutes up to, but less than, 370 minutes; or that
- the probability of any randomly selected AA battery having a life in this range is approximately 0.07.

Note: these statements assume a representative sample has been drawn. For completeness, an estimate of variability should be referred to as well (see section ‘Measures of Spread’, page 155).

Nevertheless, in summary, frequency distribution tables are important in providing information about the population from which the sample is drawn.

STEM AND LEAF PLOTS

The use of a stem and leaf plot, or stemplot, is a technique to classify either *discrete* or *continuous* variables.

In the previous example on battery life, it can be seen that there are two observations that lie in the interval 360-369. However, it cannot be seen from the table what those actual observations are.

The two values, 363 and 369, can only be found by searching through all the original data. The main advantage of a stemplot is that the data are grouped whilst also displaying all the original data.

Each observation may be considered as consisting of two parts: a stem and a leaf. To make a stemplot, each observation must first be separated into its two parts:

- a *stem* is the first digit or digits;
- a *leaf* is the final digit of a value;

- each *stem* can consist of any number of digits; and
- each *leaf* can only have a single digit.

So — for example:

- if the value of an observation is 25: the stem is 2 and the leaf is 5; and
- if the value of an observation is 369: the stem is 36 and the leaf is 9.

Where observations are accurate to one or more decimal places, such as 23.7, the stem is 23 and the leaf is 7. (The number 23.7 could be rounded off to 24 to limit the number of stems if the range of values is too great.)

In stemplots, tally marks are not required as the actual data are used.

1. The numbers of books ten students read in one year were as follows:

EXAMPLE

12, 23, 19, 6, 10, 7, 15, 25, 21, 12.

Prepare a stemplot for the data.

Stem	Leaf
0	6 7
1	2 9 0 5 2
2	3 5 1

In the table:

- the stem '0' represents the class interval 0-9,
- the stem 'l' represents the class interval 10-19, and
- the stem '2' represents the class interval 20-29.

Note that the number '6' can be written as 06 thus having a stem of 0 and a leaf of 6 .

Usually, a stemplot is placed in order, which simply means that the leaves are arranged in ascending order from left to right. Also, commas that separate the leaves (digits) are not necessary since the leaf is always a single digit.

Using the above table, the resultant *ordered* stemplot is shown below:

Stem	Leaf
0	6 7
1	0 2 2 5 9
2	1 3 5

SPLITTING STEMS

If the leaves are crowded on too few stems, then it is useful to split each stem into two or more components. Thus, for the interval 0-9, a stem split in two would create one interval of 0-4 and another interval of 5-9. A stem split in five would create the intervals 0-1, 2-3, 4-5, 6-7 and 8-9.

EXAMPLE

1. Fifteen people were asked how often they drove to work over ten working days. The number of times each person drove were as follows:

   5, 7, 9, 9, 3, 5, 1, 0, 0, 4, 3, 7, 2, 9, 8.

   Prepare an ordered stemplot for this data.

   The stemplot could be drawn as follows:

Stem	Leaf
0	0 0 1 2 3 3 4 5 5 7 7 8 9 9 9

This stemplot's organisation does not give much information about the data. Having only one stem creates an overcrowded leaf. In this case it is useful to split the stem. The stemplot is then displayed as follows:

Stem	Leaf
$0^{(0)}$	0 0 1 2 3 3 4
$0^{(5)}$	5 5 7 7 8 9 9 9

- The stem $0^{(0)}$ means all the data within the interval 0-4.
- The stem $0^{(5)}$ means all the data within the interval 5-9.

2. A swimmer training for a competition recorded the number of 50 metre laps she swam each day for thirty days. The numbers of laps recorded each day were as follows:

22, 21, 24, 19, 27, 28, 24, 25, 29, 28, 26, 31, 28, 27, 22, 39, 20, 10, 26, 24, 27, 28, 26, 28, 18, 32, 29, 25, 31, 27.

a) Prepare an ordered stemplot. Make a brief comment on what the stemplot shows.

b) Redraw the stemplot by splitting the stems into five-unit intervals. Make a brief comment on what the new stemplot shows.

---

a) The observations range in value from 10 to 39 so the stemplot should have stems of 1, 2 and 3. The ordered stemplot is shown below:

Stem	Leaf
1	0 8 9
2	0 1 2 2 4 4 4 5 5 6 6 6 7 7 7 7 8 8 8 8 8 9 9
3	1 1 2 9

It is obvious from the stemplot that the swimmer usually swims between 20 and 29 laps in training each day.

b) Splitting the stems into five-unit intervals gives the following stemplot:

Stem	Leaf
$1^{(0)}$	0
$1^{(5)}$	8 9
$2^{(0)}$	0 1 2 2 4 4 4
$2^{(5)}$	5 5 6 6 6 7 7 7 7 8 8 8 8 8 9 9
$3^{(0)}$	1 1 2
$3^{(5)}$	9

Note that $1^{(0)}$ means all the data between 10 and 14, $1^{(5)}$ means all the data between 15 and 19, and so on.

The revised stemplot shows that the swimmer usually swims between 25 and 29 laps in training each day. The values $1^{(0)}\,0 = 10$ and $3^{(5)}9 = 39$ are *outliers:* a concept that is described shortly.

3. The weights (to the nearest tenth of a kilogram) of 30 students were measured and recorded as follows:

59.2, 61.5, 62.3, 61.4, 60.9, 59.8, 60.5, 59.0, 61.1, 60.7, 61.6, 56.3, 61.9, 65.7, 60.4, 58.9, 59.0, 61.2, 62.1, 61.4, 58.4, 60.8, 60.2, 62.7, 60.0, 59.3, 61.9, 61.7, 58.4, 62.2.

Prepare an ordered stemplot for the data and briefly comment on what the analysis indicates.

In this case, the stems will be the whole number values and the leaves will be the decimal values. The data ranges from 56.3 to 65.7 so the stems should start at 56 and finish at 65.

Stem	Leaf
56	3
57	
58	4 4 9
59	0 0 2 3 8
60	0 2 4 5 7 8 9
61	1 2 4 4 5 6 7 9 9
62	1 2 3 7
63	
64	
65	7

It is not necessary to split stems because the leaves are not crowded on too few stems; nor is it necessary to round the values as the range of values is not large. The stemplot reveals that the group with the highest number of observations recorded is the 61 to 61.9 group.

OUTLIERS

An outlier is an *extreme value* of the data. It is an observation value that is significantly different from the rest of the data. There may be more than one outlier in a set of data.

Sometimes, outliers are significant pieces of data and should not be ignored. In other instances, they occur as a result of an error or misinformation and should be ignored.

In the previous example, outliers are 56.3 and 65.7, as these two values are quite different from the other values.

By ignoring these two outliers, the previous example's stemplot could be redrawn as below:

Stem	Leaf
58	4 4 9
59	0 0 2 3 8
60	0 2 4 5 7 8 9
61	1 2 4 4 5 6 7 9 9
62	1 2 3 7

Outliers: 56/3 and 65/7

When using a stemplot, it is often a matter of judgement to spot an outlier. This is because, except when using boxplots (explained on page 158), there is no strict rule to specify how far removed a value must be from the rest of a data set to qualify as an outlier.

FEATURES OF A DISTRIBUTION

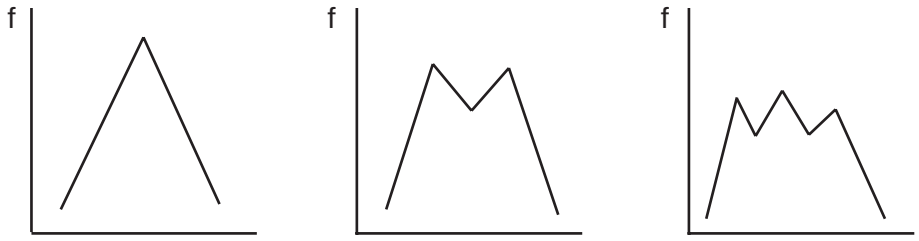
When assessing the overall pattern of any distribution, the features to look for are the number of peaks, general shape (skewed or symmetric), centre and spread.

NUMBER OF PEAKS

The first characteristic that can be readily seen from a line graph is the number of high points or peaks the distribution has.

While most distributions that occur in statistical research have only one main peak *(unimodal),* other distributions may have two peaks *(bimodal)* or more than two peaks *(multimodal).*

Examples of unimodal, bimodal and multimodal line graphs are shown below:



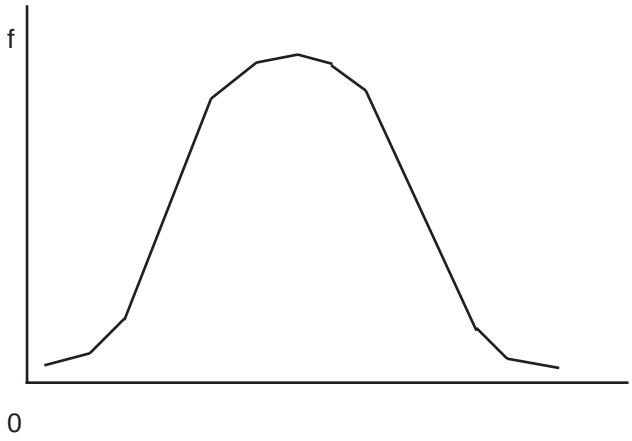
Unimodal Bimodal Multimodal

GENERAL SHAPE

The second main feature of a distribution is the extent to which it is *symmetric*.

A perfectly symmetric curve is one in which both sides of the distribution would exactly correspond if the figure was folded over its central point.

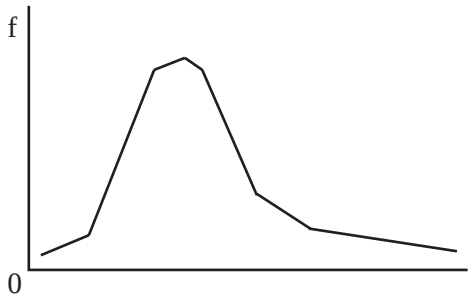
It should be noted, though, that it is unusual for a distribution to be *perfectly* symmetric. An example of a symmetric distribution is shown below:



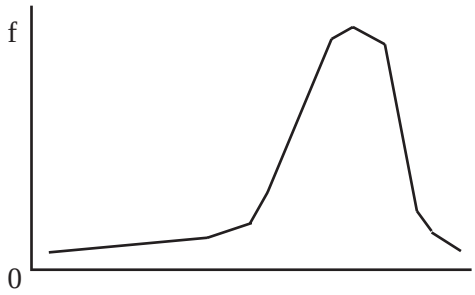
A symmetric, unimodal, bell-shaped distribution - a relatively common occurrence - is called *a normal distribution.*

If the distribution is lop-sided, it is said to be *skewed.*

A distribution is said to be skewed to the right, or *positively skewed,* if most of the data are concentrated on the left of the distribution. The right tail clearly extends further from the centre than the left tail as shown below:



A distribution is said to be skewed to the left, or *negatively skewed,* if most of the data are concentrated on the right of the distribution. The left tail clearly extends further from the centre than the right tail as shown below:



CENTRE AND SPREAD

Locating the centre (median) of a distribution can be done by counting half the observations up from the smallest. Obviously, this method is impracticable for very large sets of data. A stemplot makes this easy, as the data are arranged in ascending order. (A more precise technique of finding this mid-point is described in a later section.)

The amount of distribution spread and any large deviations from the general pattern (outliers) can be quickly spotted on the graph.

USING STEMPLOTS AS GRAPHS

A stemplot is a simple kind of graph that is made out of the numbers themselves, and is a means of displaying the main features of a distribution. By turning a stemplot on its side, it will resemble a histogram and provide similar visual information.

EXAMPLE

1. The results of forty-one students' Maths tests (out of 70) are recorded below:

   31, 49, 19, 62, 50, 24, 45, 23, 51, 32, 48, 55, 60, 40, 35, 54, 26, 57, 37, 43, 65, 50, 55, 18, 53, 41, 50, 34, 67, 56, 44, 4, 54, 57, 39, 52, 45, 35, 51, 63, 42.

   a) Is the variable discrete or continuous? Explain.

   b) Prepare an ordered stemplot for the data and briefly describe what the stemplot shows.

   Are there any outliers? If so, what are they?

   c) By turning the stemplot on its side (or rotating the page 90 degrees left), describe the distribution's main features such as:

   number of peaks,
   symmetry, and
   value at the centre of the distribution.

Answers:

a) A test score is a discrete variable. It is not possible to have a test score of 35.74542341... for example.

b) The lowest value is 4 and the highest is 67. Therefore, the stemplot for Maths test results that covers this range of values is as follows (next page):

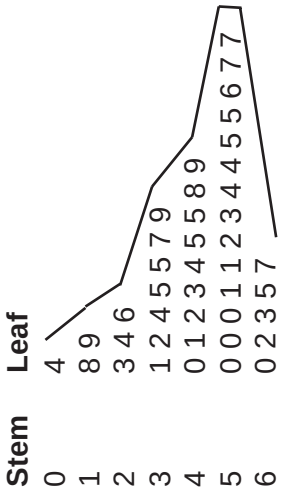
Stem	Leaf
0	4
1	8 9
2	3 4 6
3	1 2 4 5 5 7 9
4	0 1 2 3 4 5 5 8 9
5	0 0 0 1 1 2 3 4 4 5 5 6 7 7
6	0 2 3 5 7

2|4 represents 24

The stemplot reveals that most students obtained a mark in the interval between 50 and 59. The large number of students who obtained high results could mean the test was too easy, most students knew the subject being tested, or a combination of both.

The result of 4 could be an outlier, as there is a gap between this and the next result, 18.

C) If the stemplot is turned on its side, it will look like the following:



The distribution has a single peak in the 50s interval.

Although there are only 41 observations, the distribution shows that most data are clustered at the right. The left tail extends further from the data centre than the right tail. Therefore, the distribution is skewed to the left or *negatively skewed.*

As there are 41 observations, the distribution centre will occur at the 21st observation. By counting 21 observations up from the smallest, the centre is 48. (Note: the same value would have been obtained if 21 observations were counted down from the highest observation. Measures of centre or location are discussed in detail on pages 133-145.)

EXERCISES

1. Indicate which of the following are discrete or continuous variables:

   a) The time taken for you to get to school.

   b) The number of couples who were married last year.

   c) The number of goals scored by a women's hockey team.

   d) The speed of a bicycle.

   e) Your age.

   f) The number of subjects which you can choose to do next year.

   g) The time of a phone call between two people.

   h) The annual income of an individual.

   i) The number of people working at the Australian Bureau of Statistics.

   j) The number of brothers and sisters you have.

   k) The distance between your house and school.

   I) The number of pages in this book.

2. Give two examples, different to any of those given in Question 1, of:

   a) a discrete variable, and

   b) a continuous variable.

3. a) Copy and complete the frequency distribution table for the following set of data:

   2, 5, 4, 3, 4, 3, 1, 3, 3, 2, 3, 4.

Score (x)	Tally	Frequency (f)
1		
2		
3		
4		
5		

   b) Which score occurs the most frequently (the mode)?

4. A local milkbar owner records how many customers enter the store over a 25 day period. The number of customers is as follows:

   20, 21, 23, 21, 26, 24, 20, 24, 25, 22, 22, 23, 21, 24, 21, 26, 24, 22, 21, 23, 25, 22, 21, 24, 21.

   a) What type of variable is used?

   b) Present this data in a frequency distribution table by tallying the data.

   c) Which observation occurs the most frequently (the mode)?

   d) Set up a table to include the relative frequency and percentage frequency of the data.

   e) Comment briefly on what conclusions you can make from the tables.

5. The wind speed (measured to the nearest knot) of the *Fremantle Doctor* was recorded for 40 days.

   15, 22, 14, 12, 21, 34, 19, 11, 13, 0, 16, 4, 23, 8, 12, 18, 24, 17, 14, 3, 10, 12, 9, 15, 20, 5, 19, 13, 17, 11, 16, 19, 24, 12, 7, 14, 17, 10, 14, 23.

   a) What type of variable is used?

   b) Choose an appropriate class interval and present this data in a frequency distribution table by tallying the data.

   c) Which class interval occurs the most frequently?

   d) Set up a table to include the relative frequency and percentage frequency of the data.

   e) Comment briefly on what conclusions you can make from the tables.

6. Copy and complete the stem and leaf table below for the following set of data:

21, 35, 27, 2, 18, 25, 10, 4, 43, 14, 29, 24, 15, 9, 26, 31, 41, 1, 28, 38, 40, 22, 37, 26, 19, 0, 33, 12, 16, 23.

Stem	Leaf
0	
1	
2	
3	
4	

Redraw the table so that it is an ordered stem and leaf table.

7. a) Prepare an ordered stem and leaf plot for the data in Question 5.

b) Do any outliers exist? If so, can you explain the reason for their presence?

c) Describe the distribution's main features:

i) number of peaks,

ii) general shape, and

iii) approximate value at the distribution's centre.

8. The number of road fatalities in the A.C.T. from 1960 to 1992 was as follows:

10, 7, 8, 8, 17, 15, 17, 23, 14, 26, 31, 20, 32, 29, 31, 32, 38, 29, 30, 24, 30, 29, 26, 28, 37, 33, 32, 36, 32, 32, 26, 17, 20.

a) What type of variable is used?

b) Prepare an ordered stem and leaf plot for the data.

c) Expand the stemplot by using five-unit intervals

d) Do any outliers exist? If so, can you explain the reason for their presence?

e) Describe the distribution's main features:

i) number of peaks,

ii) general shape, and

iii) approximate value at the distribution's centre.

9. The mean July daily minimum temperature (Celsius) for Sydney from 1972 to 1992 is recorded as follows:

   6.1, 8.9, 6.9, 7.2, 7.0, 6.2, 5.7, 6.2, 6.8, 6.4, 6.8, 6.4, 7.6, 7.8, 7.3, 6.8, 8.8, 7.8, 8.1, 8.1, 7.9.

   a) What type of variable is used?

   b) Prepare an ordered stem and leaf plot for the data.

   c) Is it necessary to expand the stemplot? Why or why not?

   d) Do any outliers exist? If so, can you explain the reason for their presence?

   e) Describe the main features of the distribution.

10. Fifty company staff were surveyed and asked what their weekly salary was to the nearest dollar. The results follow:

    514, 476, 497, 511, 484, 513, 471, 470, 441, 466, 443, 481, 502, 528, 459, 548, 521, 517, 463, 478, 473, 514, 542, 519, 522, 523, 546, 487, 486, 473, 527, 470, 440, 564, 499, 523, 484, 463, 461, 437, 555, 525, 461, 539, 466, 470, 486, 490, 543, 519.

    a) What type of variable is used?

    b) Choose an appropriate class interval and present this data in a frequency distribution table by tallying the data.

    c) Which class interval occurs the most frequently?

    d) Set up a table to include the data's relative frequency and percentage frequency.

    e) Comment briefly on what conclusions you can make from the tables.

    f) Prepare an ordered stem and leaf plot for the data.

    g) Do any outliers exist? If so, can you explain the reason for their presence?

    h) Describe the main features of the distribution such as:

       i) number of peaks,

       ii) general shape, and

       iii) approximate value at the distribution's centre.

CLASS ACTIVITIES

1. Accurately draw a straight line measuring exactly 10 centimetres long. Without measuring, put a mark where you think halfway is (exactly). Now measure the length of each segment. By how many millimetres was your estimate short of the halfway (5cm) mark? Record this value. Find out how much the rest of the class deviated from halfway.

   With this data, construct a frequency table including relative frequency and percentage frequency.

   Which result occurred the most?

   Prepare a stem and leaf plot for the data.

   Do any outliers exist?

   How many peaks does the distribution have?

   What is the distribution's general shape?

   What is the distribution's approximate centre?

   What conclusions can you make from the analysis?

2. Ask your teacher to give you a class set of results from a recent test or assignment. Perform a detailed analysis on the data similar to that described above. Comment briefly on:

   a) standard of test or assignment,

   b) ability of the class, and

   c) standard of teaching, supporting each answer with evidence based on your analysis.

3. Throw a dice 30 times. Record each result using a frequency table.

   What type of variable is being used? Calculate the relative frequencies and percentage frequencies.

   Which result occurred the most? Would you expect any number to occur more often than the others?

   Prepare a stem and leaf plot for the data.

   Do any outliers exist?

   How many peaks does the distribution have? What is the general shape of the distribution?

   What is the distribution's approximate centre?

   What conclusions can you make from the analysis?

4. Survey teachers in your school to find what colour car they drive. Don't include shades of colours. What type of variable is this? Present the data in a frequency table, including relative frequency and percentage frequency.

   What colour car is the most popular among surveyed teachers? By what percentage is this colour more popular than the second most common colour?

   Why can't you prepare a stemplot for this data?

   How do you think a car manufacturer might use this type of data analysis?

DISPLAYING INFORMATION: GRAPH TYPES

Information that is presented in a graph can be quick and easy to understand. So it is not surprising that the use of graphs has increased in recent years, particularly in the media (newspapers and television). There are times when information is better presented by graph than by table. This is often the case when there is a *trend* or *comparison* to be shown. A graph can do this very effectively.

USING GRAPHS

Knowing how to convey information by graph is important in the presentation of statistics. The following is a list of some general rules to keep in mind when preparing graphs. A graph should:

- be simple and not too cluttered,
- show data without changing the data's message,
- clearly show any trend or differences in the data, and
- be accurate in a visual sense (if one chart value is 15 and another 30, then 30 should appear to be twice the size of 15).

It is also important to know what type of graph to use when presenting statistics. There are several types of graph you can use, which are outlined in the following pages.

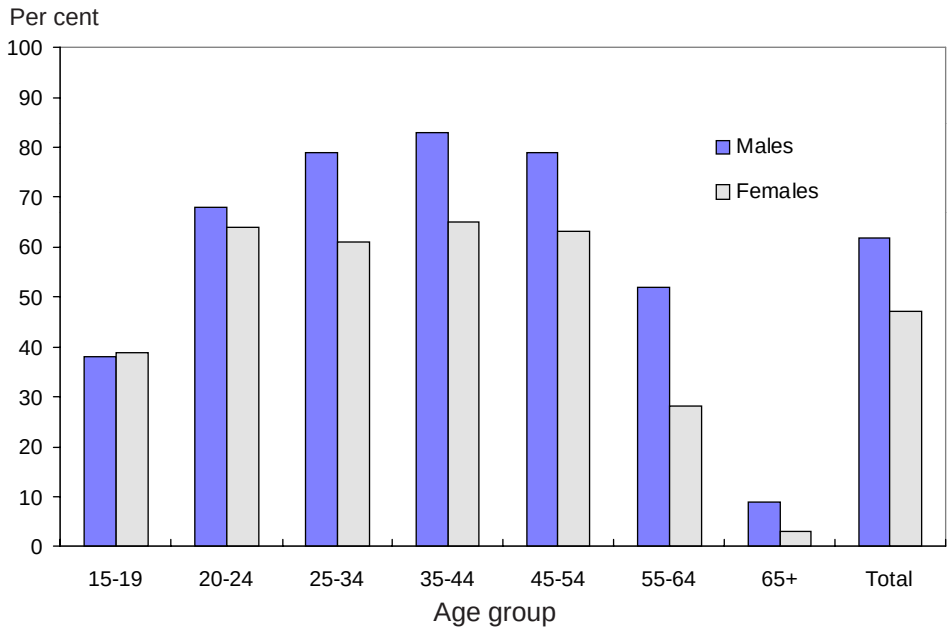
BAR GRAPH

A bar graph may be either horizontal or vertical. To differentiate between the two, a vertical bar graph is called a *column* graph. An important point about bar graphs is the *length* of the bars: the greater the length, the greater the value.

COLUMN GRAPH

Column graphs are good for comparing values. One disadvantage of column graphs is lack room for a written label at the foot of each bar; so it is best to use a column graph when the label is short, as in the example below.

PERCENTAGE OF SAME SEX AND AGE GROUP EMPLOYED, AUSTRALIA, 1996 CENSUS



Notice how a column graph allows you to show more than one series of data in the graph: in the above example, data for males and females.

A careful examination of the column graph on the previous page should allow you to make some basic conclusions about the information shown, for example:

- The graph shows a comparison between male and female employment rates by age groups.
- There is a *lower* percentage of females in employment compared to males in all age groups except one! Which one?
- The graph shows male employment rates rising up to age 35-44 and then falling.
- The graph shows female employment rates rising up to age 20-24, then falling, then rising, then falling! Why is this pattern different to that for males?

In general, column graphs display comparisons between data better than horizontal bar graphs, so you should use them in preference.

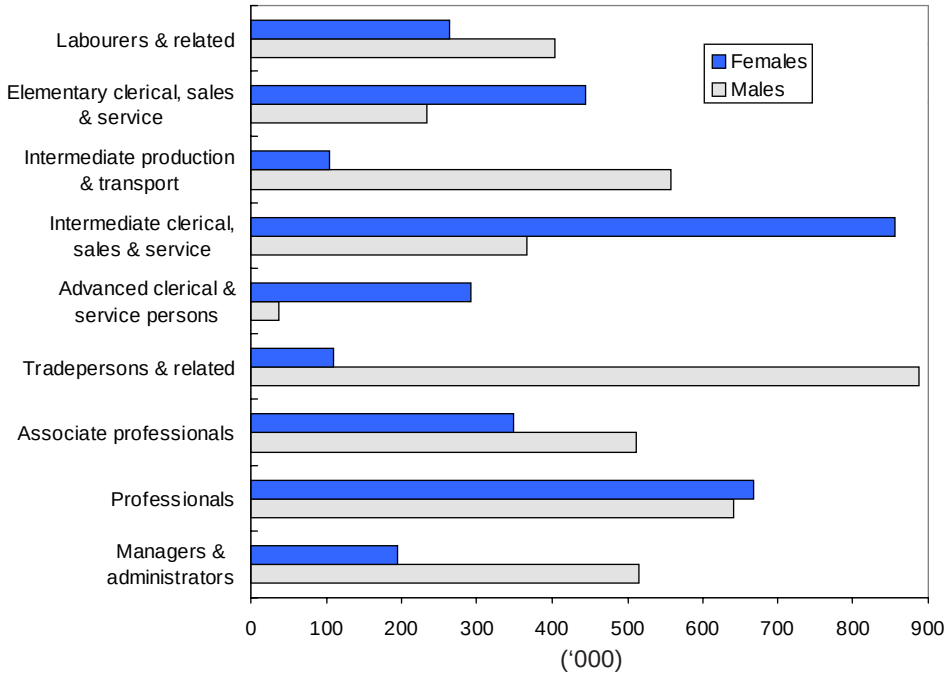
However, when category labels in the graph are long it is better to display information using a horizontal bar graph.

HORIZONTAL BAR GRAPH

There are two advantages of a horizontal bar graph over a column graph:

- category labels in a horizontal bar graph can be fully displayed (try fitting *Advanced clerical and service persons* neatly at the foot of a column!), and
- it is easier to read the scale of a horizontal bar graph.

EMPLOYED PERSONS BY OCCUPATION AND SEX, AUSTRALIA, 1996 CENSUS



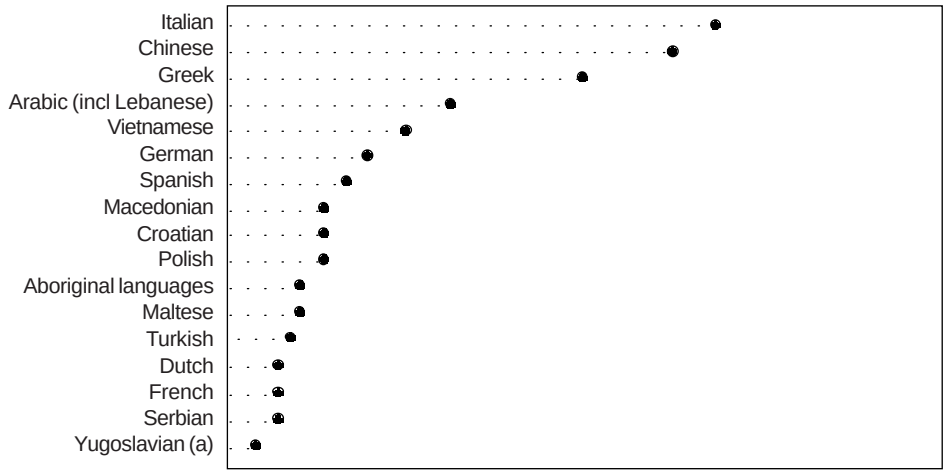
Again, a careful look at the horizontal bar graph will allow you to draw several conclusions:

- The graph compares employed males and females in occupation groups.
- In four of the nine occupation groups more females are employed than males. Which are they?
- One of the occupation groups has a very large difference (nearly 780,000) in the numbers of males and females employed within it. Which is it?

DOT CHART

The dot chart has been adopted by the ABS as the standard type of graph to display information. It is able to convey quite a lot of information in a simple way without clutter. It contrasts values very clearly, and can display many data values.

LANGUAGES OTHER THAN ENGLISH SPOKEN AT HOME, AUSTRALIA, 1996 CENSUS



Per cent of Australian population over 5 years of age

(a) Comprises 'Yugoslav not elsewhere included' and 'Serbo-Croatian'.

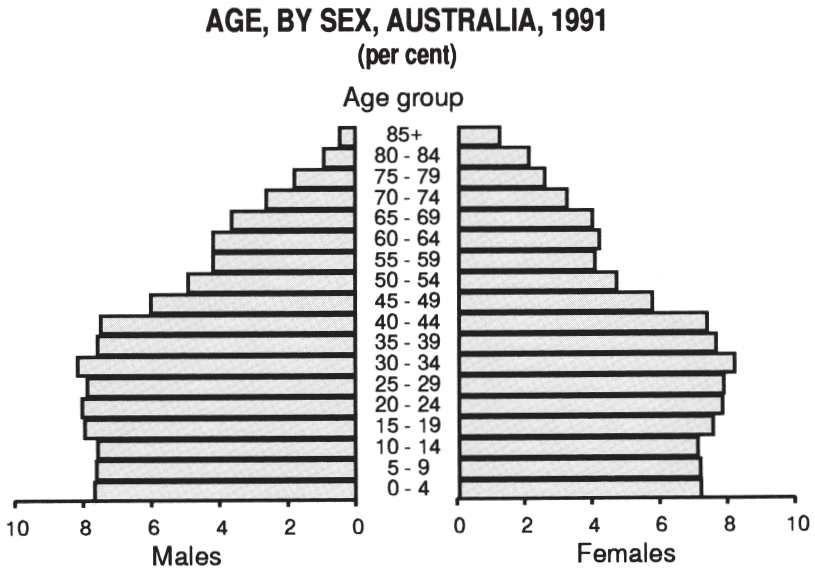
The simplicity of the above dot chart allows you to conclude that:

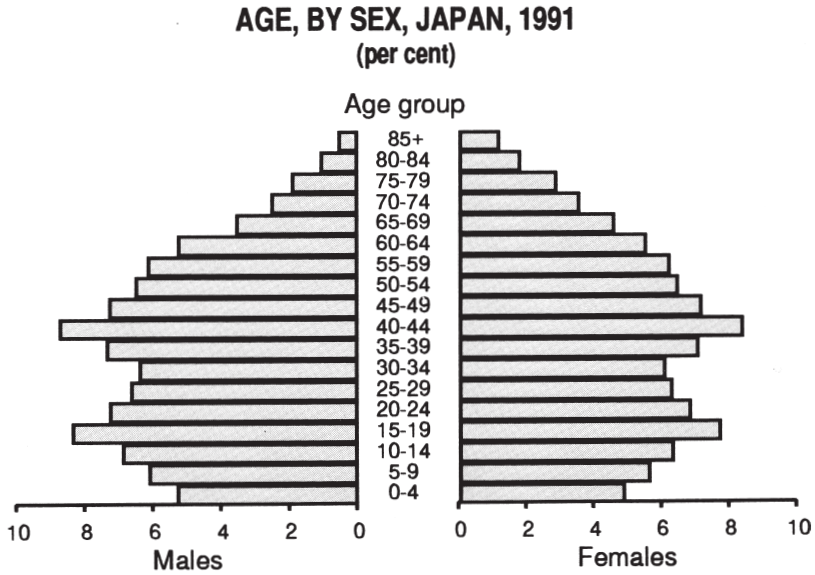
- Italian was the most commonly spoken non-English language in Australian homes: 2.2% of the Australian population aged over 5 speak it at home.
- Italian was followed by Chinese languages, Greek, Arabic, etc.

AGE PYRAMID

These are specially designed to represent the age structure of a population. They are a very effective way of showing change in a country's age structure over time, or for comparing different countries.

The values of age groups may by expressed as numbers or percentages. If you are *comparing* the age distribution of *different* populations, it is better to use percentage than number values .





Carefully study both age pyramids opposite and you should be able to see:

- the male and female age groups with the largest *number* of people.

For Australia:

- the age group with the largest *number* of people is the same for males and females; which is it?

For Japan:

- the age group with the largest *number* of people is also the same for males and females; which is it?

  (Note: the age group with the largest number of people is the same as the age group with the largest percentage of people.)

Why do you think:

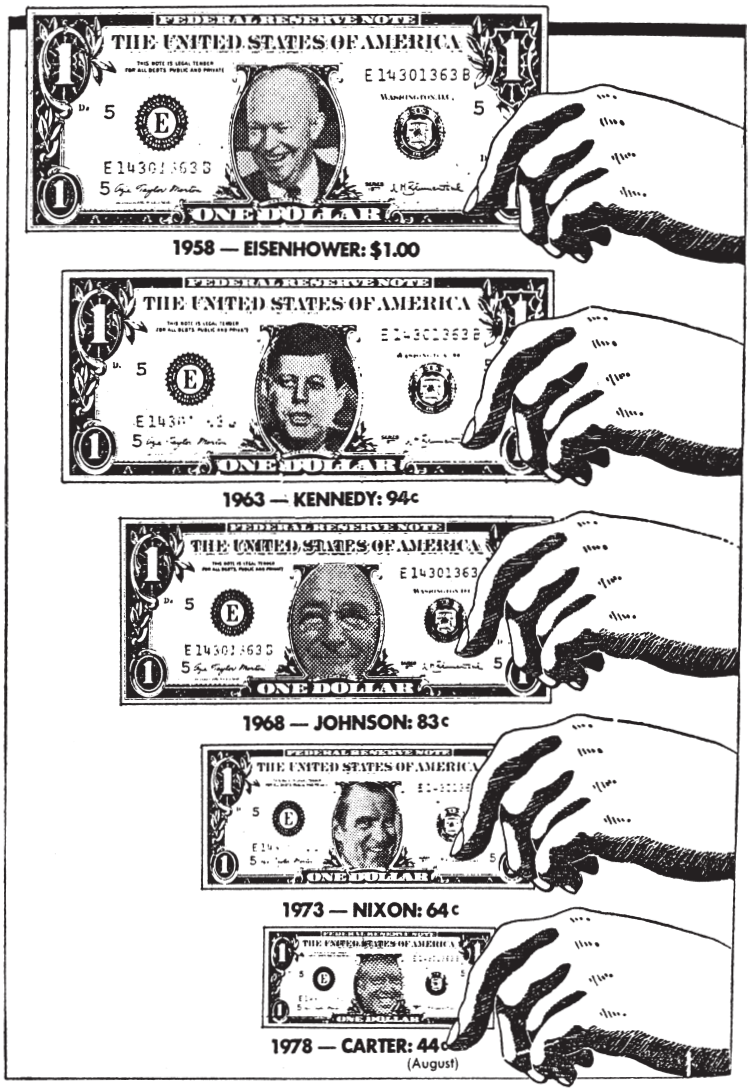
- the answers to the above two questions are different?
- the shape of the age pyramid for Japan is different to that for Australia?

PICTOGRAPH

A pictograph is a graphic illustration of statistical information. Pictographs should be used carefully as they can, either accidentally or deliberately, misrepresent the message the graph is meant to convey.

A rule mentioned at the beginning of the section was that a graph should be accurate in a visual sense. Pictographs, if not drawn carefully, can be quite inaccurate.

PURCHASING POWER OF THE AMERICAN DOLLAR



The pictograph on the previous page shows how one American dollar in 1958 had shrunk to a value of 44 cents in 1978 (due to the effects of rising prices or inflation). If you think carefully, this means that one American dollar in 1978 could buy just under half as much as it could in 1958! So is there any problem with the depiction of statistics in the pictograph?

The size or area (length by breadth) of the dollars shown are in fact misleading. They should reflect the statistics or actual purchasing power of the dollar in the year in question. As 44 cents is just under half of one dollar, so the 1978 dollar area should be just under one half of the 1958 dollar area. This means that the 1978 dollar should be about twice as big as it is.

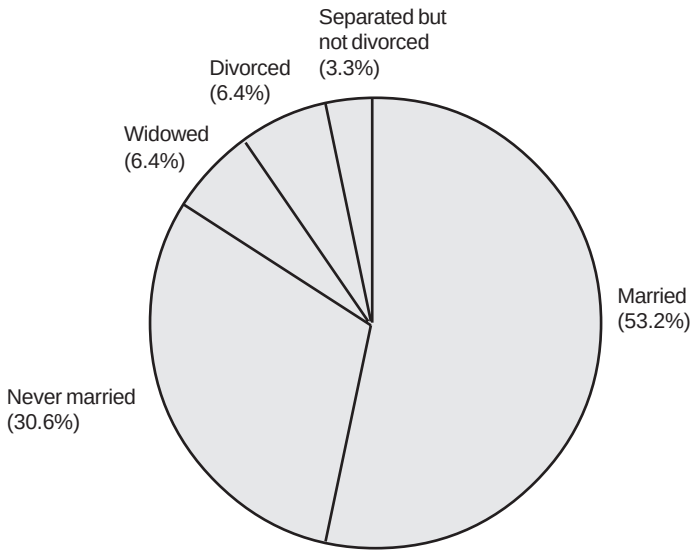
You may argue that this problem goes unnoticed by people when they look at a pictograph like this one, so it is not particularly important. However, the fact is that subconsciously many people interpret the dollar to have lost far more of its value than is the case.

It is also worth noting that the pictograph appeared during an American presidential election campaign in a leading newspaper, and would have been looked at by many voters!

PIE CHART

Pie charts are one of the most commonly used graphs. They have one advantage in that they are simple. However, one disadvantage is that it can be very difficult to see the difference in slice sizes when their values are similar. This is why it is important to label the slices with actual values, as in the example below.

MARITAL STATUS OF AUSTRALIA'S POPULATION (a)
1996 CENSUS



(a) Population aged 15 and over

A pie chart is constructed by converting the percentage share of each category into the same percentage of 360 degrees. In the previous chart, for example:

- the *married* category is 53.2%,
- 53.2% of 360° is 191.52°, and
- using the radius in the 12 o'clock position as the origin, the angle of 191.52° (rounded to 192°) is measured with a protractor and a radius marked off .

This procedure is followed with remaining categories until the pie is complete. The final category need not be measured as its radius is already in position.

An important rule when drawing a pie chart is that *segments are ordered by size (largest to smallest) in a clockwise direction*.

It is best that *segments number no more than five,* so the chart does not become too cluttered.

The simplicity of the pie chart opposite tells you quickly that:

- the majority of Australia's population aged 15 and over were married at the time of the 1996 Census; and
- just under a third were never married.

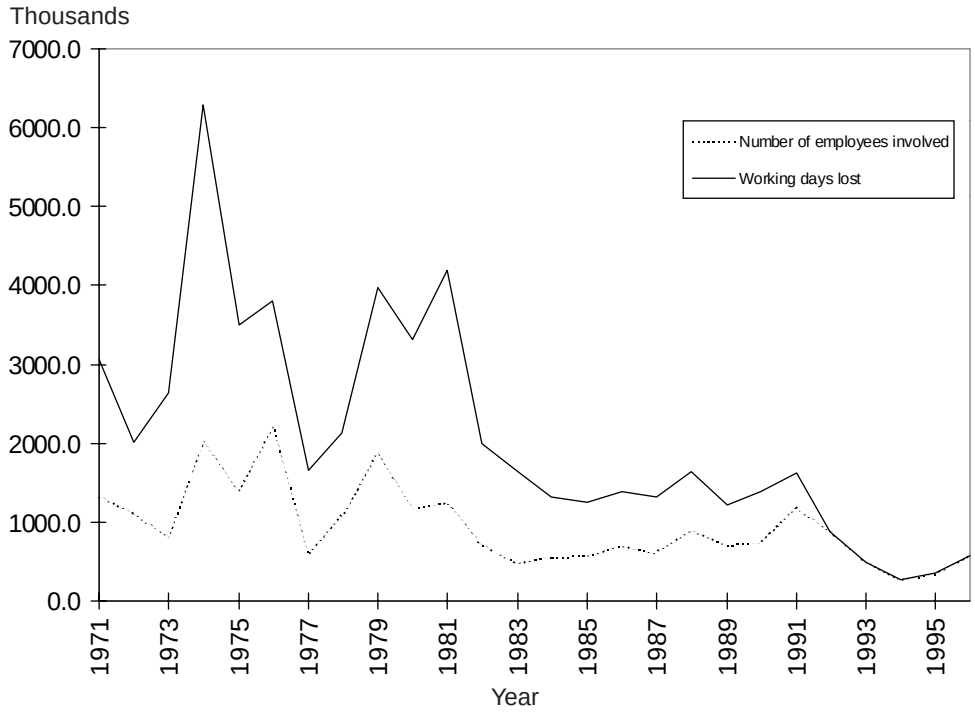
Note that if the *Widowed* and *Divorced* segments were not labelled with percentage values it would be difficult to tell quickly which segment was bigger.

NOTE: Many computer packages will draw pie charts for you quickly and easily. However, research has shown that many people can make mistakes when trying to compare pie chart values. In general, bar charts get the same type of information across to people with less possibility for misunderstanding.

LINE GRAPH

A line graph is a very common way of presenting statistics. It is particularly useful when you want to display information over a time period. It should always be used when you want to show a trend in data over time.

INDUSTRIAL DISPUTES, AUSTRALIA, 1971-96.



The line chart above shows one obvious trend:

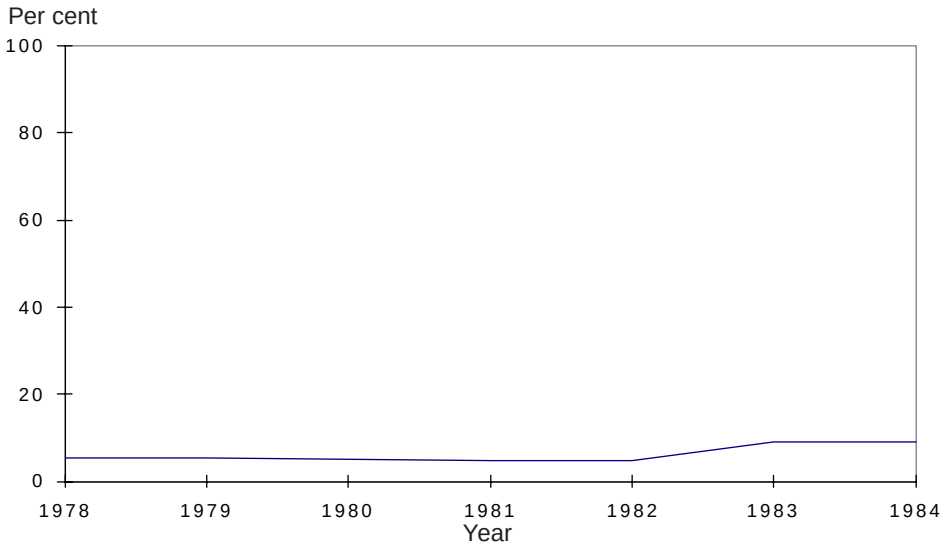
- the number of working days lost through industrial disputes in Australia was far greater in the 1970s than during the 1980s and 1990s.

Can you tell from the graph:

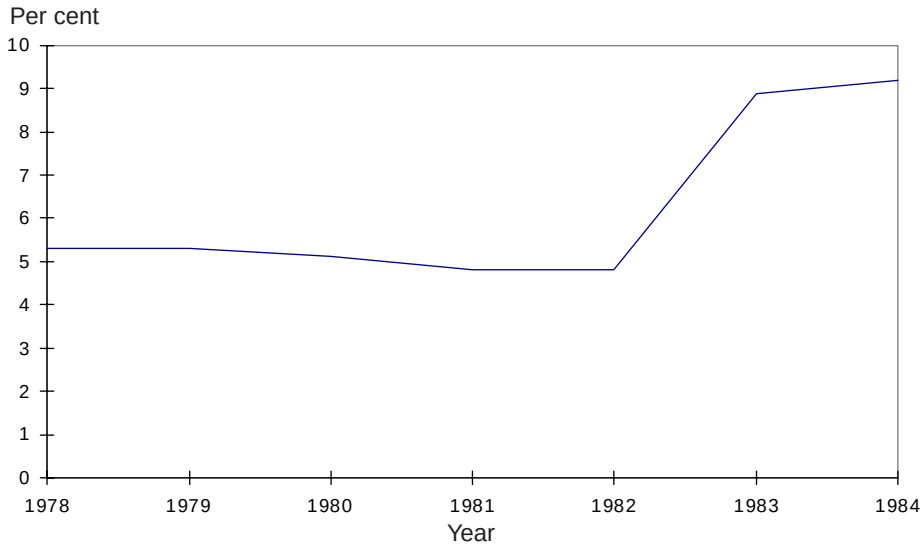
- the year in which the most working days were lost?

It is important when drawing a line graph that you use the correct scale. Otherwise the line's shape can give an incorrect impression about information. Consider the following example:

MALE UNEMPLOYMENT RATES, AUSTRALIA, 1978-84



MALE UNEMPLOYMENT RATES, AUSTRALIA, 1978-84



Using a scale of 0 to 100 (top chart) does not effectively show the doubling of male unemployment rates between 1982 and 1983. However, choosing a scale of 0 to 10 (bottom chart) brings out this important message in the statistics.

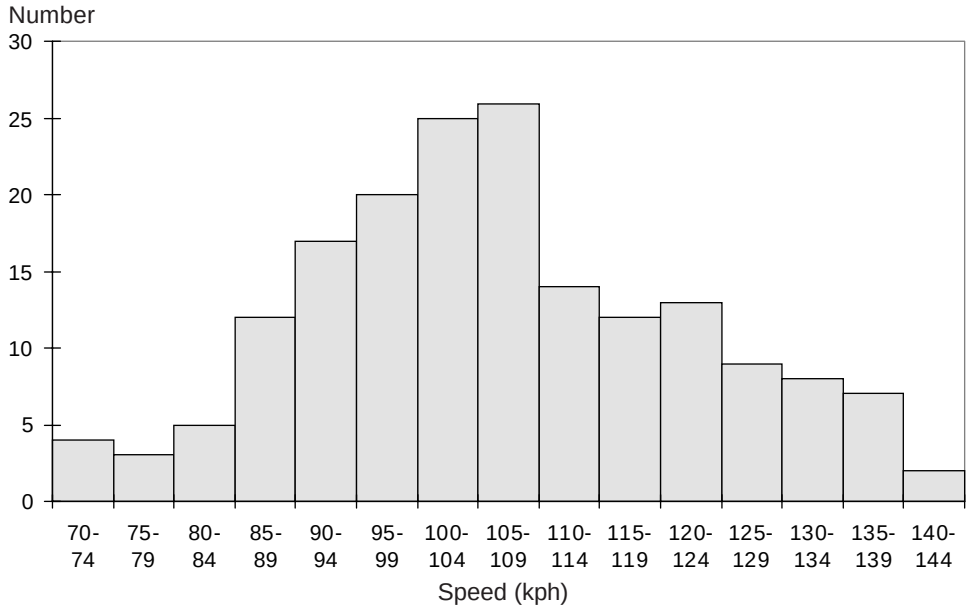
HISTOGRAM

A histogram has a similar appearance to a column graph but no gaps between the columns. It is used to depict data from the measurement of a continuous variable.

Technically, the difference between column graphs and histograms is that:

- in a histogram: frequency is measured by the area of the column, and
- in a column graph: frequency is measured by the height of the column.

DISTRIBUTION OF VEHICLE SPEEDS ON A FREEWAY



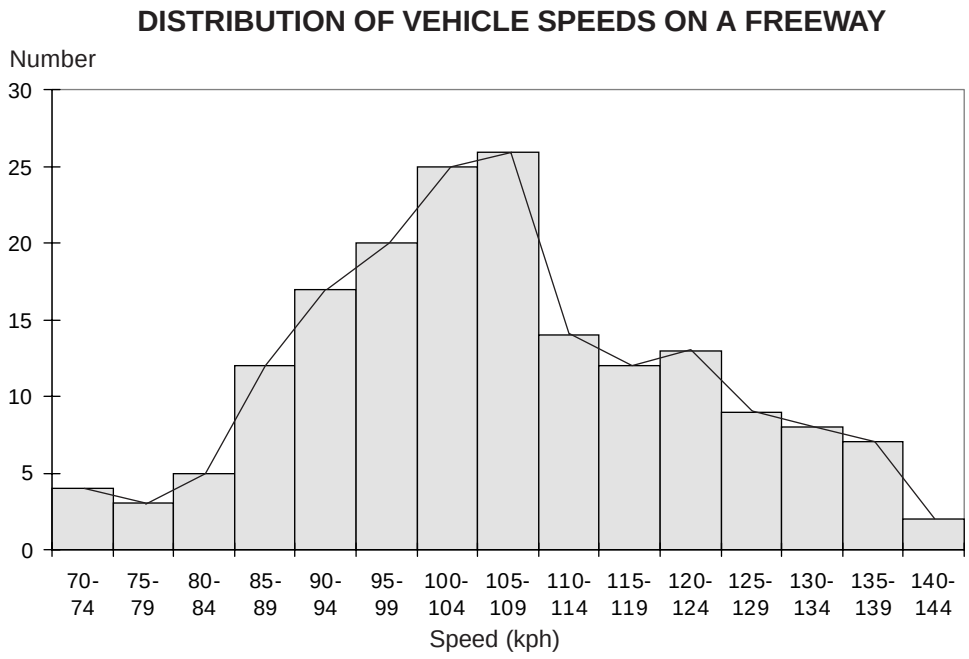
Generally, a histogram will have equal width bars, although when class intervals vary in size this will not be the case.

Choosing the appropriate width of the bars for a histogram is very important.

FREQUENCY POLYGON

A frequency polygon is a graph formed by joining the mid-points of histogram column tops. Obviously, they are only used when depicting data from the continuous variable shown on a histogram.

A frequency polygon smoothes out abrupt changes that may appear in a histogram, and is therefore useful for demonstrating continuity of the variable being studied.



SUMMARY

- **Column graph**: used when comparing data values is important, and there are five or fewer categories. When there are more than five, a dot chart should be used. Column graphs generally display data better than horizontal bar graphs, and are preferred where possible.

- **Horizontal bar graph**: used when category names are too long to fit at the foot of a column. As with the column chart, it is more suited to five or fewer categories. When there are more than five categories, use a dot chart.

- **Dot chart**: used when displaying a comparatively large number of categories and category order is unimportant. It is best used when portraying category values in descending order of size.

- **Age pyramid**: used when representing population age structure.

- **Pictograph**: only used by professional graphic artists, although simple pictorial presentations can be done by students. Care should be taken that comparisons are accurately depicted.

- **Pie chart**: used for simple comparison of a small number of categories. Values should be markedly different, or differences may not be easily seen. Labelling sectors with their actual values overcomes this problem. In some cases, where data values are close to each other, a pie chart's message may be easily misunderstood. A column or horizontal bar chart may be more useful.

- **Line graph**: used for depicting data over time.

- **Histogram**: this should be used with the same advice for a column graph, when depicting continuous variable data.

- **Frequency polygon**: this should be used when depicting continuous variable data, and you want to smooth out abrupt changes that may appear in a histogram.

EXERCISES

1. Over a period, say one week, keep a record of graphically presented statistics in a leading newspaper. Are the graph examples appropriate, or could any of the statistics have been presented better with a different type of graph?

2. What type of graph would you choose to present the following information, and why would it be preferable over other types?

   a) Number of female students in each year level in your school.

   b) Annual road toll (number of fatalities in road traffic accidents) in your State for the last 30 years.

   c) Speed of the world's fastest 20 animals.

   d) Population of China (including males and females).

3. What is the major problem with using a pie chart to present information? Is there a way you might alleviate this problem?

4. When would it be preferable to use a horizontal bar graph rather than a vertical bar graph?

5. Obtain an official copy of the latest music charts for Australia. Produce graphs showing the different information they contain. For example, the number of weeks a record has been in the charts, or the ranking of the 'Top Ten'.

C UMULATIVE FREQUENCY AND PERCENTAGE

Numerical variables can be represented in a variety of ways, including: stem and leaf, frequency distribution, cumulative frequency or cumulative percentage tables. As you will see, the graphs of these are very useful in finding the centres of large data sets.

CUMULATIVE FREQUENCY

Cumulative frequency is used to determine the number of observations that lie above (or below) a particular value.

The cumulative frequency is found from a stem and leaf table or a frequency distribution table by *adding each frequency to the sum of its predecessor.*

The last value will always equal the total for all observations, as all frequencies will have been added.

For *continuous or discrete* variables:

- cumulative frequency is calculated from a frequency distribution table. A stem and leaf plot can be used to construct a frequency distribution table.

DISCRETE VARIABLES

EXAMPLE

1. The number of people who climbed Ayers Rock over a thirty day period were counted and recorded as follows:

31, 49, 19, 62, 24, 45, 23, 51, 55, 60, 40, 35, 54, 26, 57, 37, 43, 65, 18, 41, 50, 56, 4, 54, 39, 52, 35, 51, 63, 42.

a) Set up a stem and leaf plot, and hence calculate the cumulative frequency by adding appropriate columns.

b) Plot a graph of cumulative frequency against number of people.

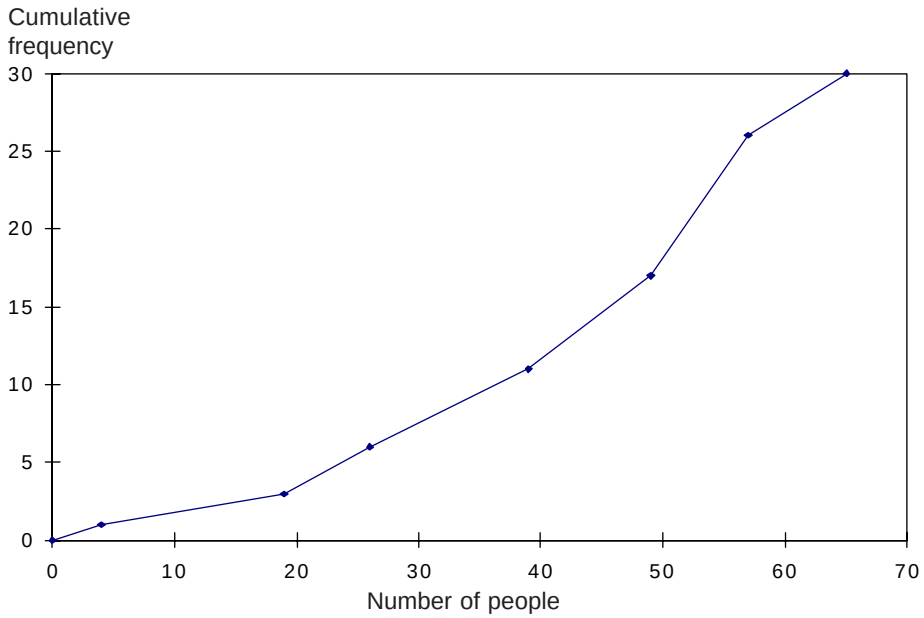
Answers.

a) The data ranges from 4 to 65, so the data is grouped in class intervals of 10 to produce the following table:

Stem	Leaf	Frequency (f)	Upper value	Cumulative frequency
0	4	1	4	1
1	8 9	2	19	1+2 = 3
2	3 4 6	3	26	3+3 = 6
3	1 5 5 7 9	5	39	6+5 = 11
4	0 1 2 3 5 9	6	49	11+6 = 17
5	0 1 1 2 4 4 5 6 7	9	57	17+9 = 26
6	0 2 3 5	4	65	26+4 = 30

b) Because the variable is discrete, the actual upper value recorded in each class interval is used in plotting the graph. Even though the variable is discrete, the plotted points are joined to form a continuous cumulative frequency polygon or curve, known as an *ogive*.

The cumulative frequency is always labelled on the vertical axis and any other variable, in this case the number of people, is labelled on the horizontal axis as shown below:



Some information that can be gained from either graph or table:

- On 11 of the 30 days, not more than 39 people climbed Ayers Rock on a given day.
- On 13 of the 30 days, 50 or more people climbed Ayers Rock.

CONTINUOUS VARIABLES

When a continuous variable or variable taking a large number of values is used, plotting the graph requires a different approach to that for a discrete variable.

EXAMPLE

1) The snow depth at Thredbo in the Snowy Mountains was measured (to the nearest centimetre) for twenty-five days and recorded as follows:

242, 228, 217, 209, 253, 239, 266, 242, 251, 240, 223, 219, 246, 260, 258, 225, 234, 230, 249, 245, 254, 243, 235, 231, 257.

a) Set up a frequency distribution table and hence calculate the cumulative frequency by adding appropriate columns.

b) Plot the graph of snow depth against the cumulative frequency.

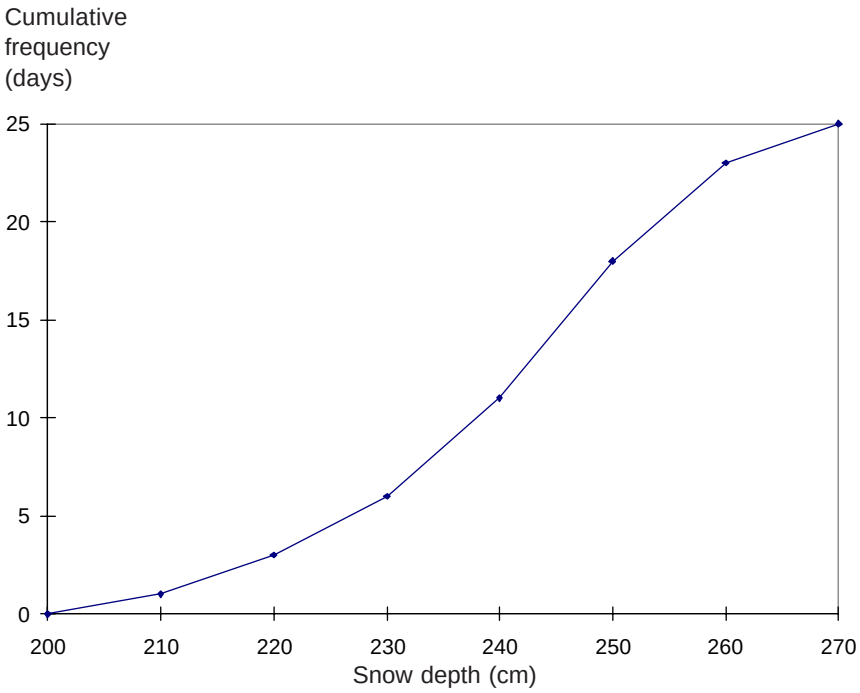
Answers.

a) The data ranges from 209cm to 266cm, so the data are grouped in class intervals of 10 to produce the following table:

Snow depth (x)	Tally	Frequency (f)	End-point	Cumulative frequency
			200	0
200-<210	I	1	210	1
210-<220	II	2	220	3
220-<230	III	3	230	6
230-<240	𝍸	5	240	11
240-<250	𝍸 II	7	250	18
250-<260	𝍸	5	260	23
260-<270	II	2	270	25

b) Because the variable is continuous, the end-points of each class interval are used in plotting the graph. The plotted points are joined to form an *ogive*.

Remember that the cumulative frequency is always labelled on the vertical axis and any other variable, in this case snow depth, is labelled on the horizontal axis as shown below:



Information that can be gained from either the graph or the table:

- None of the 25 days had snow depth less than 200 centimetres.
- On 1 of the 25 days snow depth was less than 210 centimetres.
- On 2 of the 25 days snow depth was 260 centimetres or more.

CUMULATIVE PERCENTAGE

The main advantage of using cumulative percentage rather than cumulative frequency is that it provides an easier way to compare different sets of data.

The cumulative frequency and cumulative percentage graphs are exactly the same, the only difference being the vertical axis scale. In fact, it is possible to have the two vertical axes, cumulative frequency and cumulative percentage, on the same graph.

Cumulative percentage is calculated by dividing the cumulative frequency by the number of observations, n, then multiplying by 100 (the last value will always be equal to 100%). Thus:

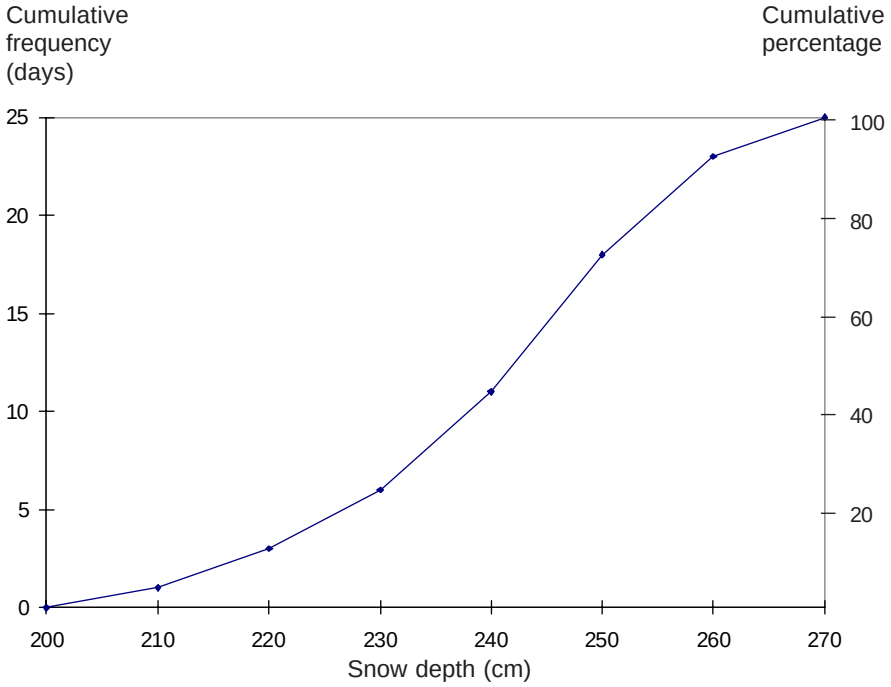
CUMULATIVE PERCENTAGE	= CUMULATIVE FREQUENCY ÷ n x 100

EXAMPLE

1. From the previous example, calculate the cumulative percentage and hence draw a graph with two different vertical axes: one for cumulative frequency and one for cumulative percentage.

Snow depth (x)	Tally	Frequency (f)	End-point	Cumulative frequency	Cumulative percentage
			200	0	0/25x100=0
200-<210	I	1	210	1	1/25x100=4
210-<220	II	2	220	3	3/25X100=12
220-<230	III	3	230	6	6/25X100=24
230-<240	𝍸	5	240	11	11/25x100=44
240-<250	𝍸 II	7	250	18	18/25x100=72
250-<260	𝍸	5	260	23	23/25x100=92
260-<270	II	2	270	25	25/25x100=100

Apart from the extra axis, the graph will be exactly the same as that drawn in the previous example:



Information that can be gained from either the graph or table:

- On 24% of days, snow depth was less than 230 centimetres.
- On 7 of the 25 days, snow depth was at least 250 centimetres.

In summary, most ogives look similar to a stretched ‘S’. They are used to determine the number, or percentage, of observations that lie above (or below) a specified value.

EXERCISES

1. The following set of data gives the length of reign (to the nearest year) of various Kings and Queens of England since the Battle of Hastings in 1066.

   21, 13, 35, 19, 35, 10, 17, 56, 35, 20, 50, 22, 13, 9, 39, 22, 0, 2, 24, 38, 6, 5, 44, 22, 24, 25, 3, 13, 6, 12, 13, 33, 59, 10, 7, 63, 9, 25, 1, 15.

   a) Present the data in the form of an ordered stem and leaf plot.

   b) Do any outliers exist? If so, can you explain the reason for their presence?

   c) Describe the main features of distribution such as:

      i) number of peaks,
      ii) general shape, and
      iii) approximate value at the centre of the distribution.

   d) Calculate cumulative frequency and cumulative percentage.

   e) Draw the ogive with two different vertical axes: one for cumulative frequency and one for cumulative percentage.

   f) How many rulers reigned for less than 10 years?

   g) How many rulers reigned for 50 years or more?

   h) The current Queen of England is Queen Elizabeth II. She has reigned since 1953, and her reign has not been included in the data set. Calculate her length of reign, and briefly comment on this in comparison with the other rulers.

2. At a fast food outlet, *Hungry Stats*, a student often buys a small bag of french fries. Curious to know whether she was getting value for money and how consistent the store was with each bag, she counted and recorded the fries in each bag. The results from 30 different visits were as follows:

   44, 46, 54, 38, 49, 46, 45, 31, 55, 37, 42, 43, 47, 51, 48, 40, 59, 35, 47, 21, 43, 37, 45, 38, 40, 32, 50, 34, 43, 54.

   a) Present the data in an ordered stem and leaf plot. Split the stems if necessary.

b) Do any outliers exist? If so, can you explain the reason for their presence?

c) Describe the main features of distribution such as:

i) number of peaks,
ii) general shape, and
iii) approximate value at the centre of distribution.

d) Calculate the cumulative frequency and cumulative percentage.

e) Draw the ogive with two different vertical axes: one for cumulative frequency and one for cumulative percentage.

f) How many bags had fewer than 40 fries in them?

g) What percentage of bags had 45 or more fries in them?

h) Copy and complete Hungry Stats' promotional saying: 'Fifty per cent of our small bags of french fries contain at least... french fries.'

3. The table below is from the 1996 Census for Darwin. It shows numbers of unemployed females looking for full-time work, by age group.

Age group (a)	Number of females
15-24	339
25-34	273
35-44	147
45-54	121
55-64	22

(a) Age is collected in completed number of years. Thus, the interval 15-24 has an upper end-point of 25 (refer to page 84).

a) Is the variable discrete or continuous?

b) Copy the table and calculate cumulative frequency and cumulative percentage.

c) Draw the ogive with two different vertical axes: one for cumulative frequency and one for cumulative percentage.

d) Why is there no data for females less than 15 years old?

e) In what age group does the cumulative percentage value '50' lie?

f) What percentage of unemployed females looking for full-time work are less than 25 years old.

g) What percentage of unemployed females looking for full-time work is 55 years old or older?

h) How would Australian governments use this sort of information?

4. A survey was taken of 50 ABS employees in Brisbane to determine how long it takes them to travel to work. The results, to the nearest minute, were recorded as follows:

33, 63, 49, 65, 56, 45, 52, 63, 38, 66, 43, 98, 60, 58, 68, 29, 59, 87,
22, 64, 73, 56, 71, 67, 44, 31, 83, 50, 75, 65, 60, 51, 89, 69, 41, 76,
58, 62, 25, 52, 64, 77, 61, 55, 80, 45, 12, 69, 40, 37

a) What type of variable is this?

b) Present the data in a frequency table, using appropriate intervals, including relative and percentage frequencies.

c) Draw a histogram to represent the data and mark in the frequency polygon.

d) Prepare an ordered stem and leaf plot for the data. Do any outliers exist? If so, can you explain the reason for their presence?

e) Describe the main features of distribution such as:
   i) number of peaks,
   ii) general shape, and
   iii) approximate value at the centre of the distribution.

f) Calculate the cumulative frequency and cumulative percentage. Put in the end-points.

g) Draw the ogive with two different vertical axes: one for cumulative frequency and one for cumulative percentage.

h) What was the most common time interval taken for ABS staff to travel to work?

i) What percentage of people took longer than 90 minutes to travel to work?

j) How many staff took less than 40 minutes to travel to work?

CLASS ACTIVITY

1. Survey teachers in your school to find out how long they have been teaching (to the nearest year). What type of variable is this? Present the data in a frequency table, using appropriate intervals, including relative and percentage frequencies.

   For how many years have the majority of teachers taught? By what percentage is this more than the second most common length of service? Draw a histogram to represent the data and mark in the frequency polygon. Prepare an ordered stem and leaf plot for the data.

   Do any outliers exist? If so, can you explain the reason for their presence?

   Describe the main features of distribution such as:

   i) number of peaks,
   ii) general shape, and
   iii) approximate value at the centre of the distribution. Calculate cumulative frequency and cumulative percentage.

   Draw the ogive with two different vertical axes: one for cumulative frequency and one for cumulative percentage

   How many teachers have taught for more than ten years?

   What percentage of teachers has taught for more than ten years?

   What percentage of teachers has taught for less than ten years?

   What is the number of years below which half the teachers have taught?

   Present your analysis and report in a neat project form.

MEASURES OF LOCATION

The centre of a set of data is important. Often, you want to know what most people think, or the average of a set of values. If you have a normal distribution (page 91), a measure of the centre provides information on what the value is for most of the population.

Finding the central location of a data set requires the calculation of the mean, median or mode. All three measures indicate the location of the centre (often called the central tendency). However, each measure has its own definition and application in different situations.

MEAN

DEFINITION

The mean of a numeric variable is calculated by summing the values of all observations in a data set and then dividing by the number of observations in the set. It is often referred to as the average. Thus:

MEAN	= SUM OF ALL THE OBSERVED VALUES ÷ NUMBER OF OBSERVATIONS

DISCRETE VARIABLES

EXAMPLE

1. In 1997 Tony Modra was a leading goal kicker in the Australian Football League. In 10 matches he kicked 7, 5, 0, 7, 8, 5, 5, 4, 5, & 1 goals. What was his mean score?

   Mean = sum of all the observed values ÷ number of observations

   = (7+5+0+7+8+5+5+4+5+1) ÷ 10

   = 47 ÷ 10

   = 4.7

   Therefore, for the above 10 matches Tony Modra kicked an average 4.7 goals per match. The value 4.7 is not a whole number so it only has meaning in a *statistical* sense. In reality it is impossible to kick 4.7 goals (even if you are Tony Modra). *Note:* it is possible to kick 6.17 goals. Why?

Using mathematical notation, for a discrete variable the mean is calculated as follows:

$$\bar{x} = \frac{\sum x}{n}$$

where: **x** stands for an observed value,

$\bar{x}$ stands for the mean value of x,

Σx stands for the sum of all observed x values, and

n stands for the number of observations in a set of data.

2. The number of people killed in road traffic accidents in New South Wales from 1983 to 1996 is given in the following table. What was the average number of people killed per year on New South Wales' roads from 1983 to 1996? How many people died daily in road traffic accidents in New South Wales during this period?

Year	People killed
1987	959
1988	1 037
1989	960
1990	797
1991	663
1992	652
1993	560
1994	619
1995	623
1996	583

$$\overline{X} = \frac{\sum x}{n}$$

$$= 7{,}453 \div 10$$

$$= \mathbf{745.3}$$

This is the average number of people killed per year on New South Wales' roads from 1987 to 1996.

To calculate the daily death rate from road traffic accidents, the average yearly death rate is divided by the number of days in a year (leap years are ignored).

Thus: $745.3 \div 365 = 2.0$ deaths/day approximately

Therefore, on average, approximately 2.0 people died daily in road traffic accidents in New South Wales from 1987 to 1996.

Historical note: the highest road toll recorded in New South Wales was in 1978 when 1,384 people lost their lives.

How do you think road traffic accident statistics can be used to reduce the number of people killed on the roads each year?

FREQUENCY TABLE (DISCRETE VARIABLES)

3. Grouping observations in tables is useful when dealing with a large amount of data. Tony Modra's goal kicking figures can be displayed in a frequency table:

No. of goals (x)	Frequency(f)	xf
0	1	0
1	1	1
4	1	4
5	4	20
7	2	14
8	1	8

Because the observations are grouped, the mathematical notation changes slightly. For a discrete variable in a frequency table the mean is calculated as follows:

$$\overline{X} = \frac{\Sigma xf}{\Sigma f}$$

where: **x** stands for an observed value,
$\overline{\mathbf{x}}$ stands for the mean value of x,
Σ **xf** stands for the sum of all xf values, and
Σ **f** stands for the sum of the frequencies.

Therefore, to calculate the mean for Tony Modra's goal kicking:

$\Sigma xf = (0+1+4+20+14+8) = 47$

$\Sigma f = (1+1+1+4+2+1) = 10$

$\overline{x} = 47/10$

$= \mathbf{4.7}$

GROUPED VARIABLES (CONTINUOUS OR DISCRETE)

NOTE: Determine the mid-point of each class interval for a variable before calculating the mean from a frequency table. This method provides an approximation of the true mean for an ungrouped variable. How good the approximation is depends on how evenly the observed values are spread within each group.

4. The following table shows the heights of 50 randomly selected Year 10 girls in a school. What is the mean height of the girls?

Height (centimetres)	Mid-point (x)	Frequency (f)	xf
150 - <155	152.5	4	610.0
155 - <160	157.5	7	1 102.5
160 - <165	162.5	18	2 925.0
165 - <170	167.5	11	1 842.5
170 - <175	172.5	6	1 035.0
175 - <180	177.5	4	710.0
		50	**8 225.0**

Thus: $\overline{X} = \frac{\Sigma xf}{\Sigma f}$

$= 8{,}225 \div 50$

$=$ **164.5 cm**

Therefore, the mean height of the 50 Year 10 girls is 164.5 cm.

MEDIAN

DEFINITION

If observations of a variable are ordered by value, the median value corresponds to the middle observation in that ordered list. The median value corresponds to a cumulative percentage of 50 per cent. The position of the median is the:

$\frac{n + 1}{2}$th value, where n is the number of values in a set of data.

There are as many values above the median as there are below. After the data have been placed in ascending order:

MEDIAN	= THE MIDDLE VALUE OF A SET OF DATA

The median is usually calculated for numeric variables, but may also be calculated for an ordinal nominal variable.

DISCRETE VARIABLES

EXAMPLE

1. Cathy Freeman is one of Australia's leading Aboriginal athletes. In a typical 200 metre training session she runs the following times:

   26.1, 25.6, 25.7, 25.2 and 25.0 seconds. Find the median time.

   First the values are put in ascending order:

   25.0, 25.2, 25.6, 25.7, 26.1

   Median = (n + 1)/2th value
   = (5+1)/2th = 3rd value
   = **25.6 seconds** (2 values above and 2 below)

2. If Cathy then runs her 6th 200 metre run in 24.7 seconds, what is the median value now?

   Again the data are put in ascending order:

   24.7, 25.0, 25.2, 25.6, 25.7, 26.1

   Median = (6+ 1)/2th = 3.5th value

   Therefore, it lies between the 3rd and 4th values. Since there is an even number of observations, there is no distinct middle value. The median is calculated by averaging the two middle values 25.2 and 25.6.

Thus: = (25.2 +25.6) ÷ 2
= **25.4 seconds**

3. Ordered stem and leaf tables make it simple to calculate the median, particularly if cumulative frequencies have been calculated. Consider the heights of the 50 Year 10 girls.

Using a stem and leaf table:

Stem	Leaf	Cumulative frequency
15 (0)	0 1 1 4	4
15 (5)	5 6 7 7 8 8 8	11
16 (0)	0 1 1 1 1 2 2 2 2 2 2 3 3 3 4 4 4 4	29
16 (5)	5 5 5 5 6 6 6 7 7 8 9	40
17 (0)	0 0 1 2 3 3	46
17 (5)	6 6 7 8	50

15|7 represents 157

There are 50 pieces of data, so the median is the value of the:

(50+ l)/2th = 25.5th observation

Therefore, the median lies between the values of the 25th and 26th observations.

That is, the median lies between 163cm (25th observation) and 164cm (26th observation). The median is found by averaging these 2 values.

Thus: = (163 + 164) ÷ 2
= **163.5 cm**

(Since height is a continuous variable, 163.5cm is an acceptable median value.)

FREQUENCY TABLE (DISCRETE VARIABLES)

4. If the scores from 10 netball matches are placed in a frequency table, what is the median?

No. of goals (x)	Frequency (f)
4	1
5	2
6	0
7	2
8	4
9	1

The median is the (10 + 1)/2th = 5.5th value.

From the frequency column in the above table, it will be either the 5th value (7) or the 6th value (8).

If the average of these is calculated, the result is 7.5.

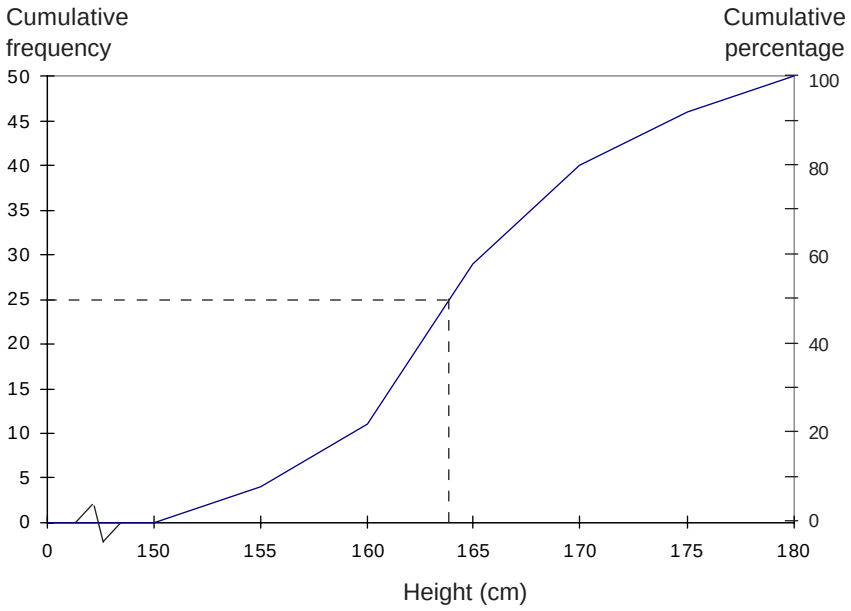
NOTE: Technically, the median should be a possible variable value. In the above example, the variable is discrete and always a whole number. Therefore, 7.5 is not a possible variable value and is strictly not the median. Some argue that 8 is a more appropriate median. For our purposes 7.5 is acceptable.

GROUPED VARIABLES (CONTINUOUS OR DISCRETE)

5. Grouping the data in Example 3 will allow you to find the median using a cumulative frequency graph. The end-points of height intervals, cumulative frequency and cumulative percentage columns are shown in the following table.

Height (centimetres)	Frequency (f)	End-point (x)	Cumulative frequency	Cumulative percentage
		150	0	0
150-<155	4	155	4	8
155-<160	7	160	11	22
160-<165	18	165	29	58
165-<170	11	170	40	80
170-<175	6	175	46	92
175-<180	4	180	50	100

The cumulative frequency graph can now be plotted.



The median obtained from the cumulative frequency graph is a different value to the median obtained from the stem and leaf table. This is because, unless the graph is drawn precisely with all the information used, you can only find an approximation for the median. (Plotting a detailed graph can be time consuming.)

COMPARING THE MEAN AND MEDIAN

It is possible to have the mean and median of a distribution equal to the same value. This is always the case if distribution is symmetric, and the two values will be close together if distribution is roughly symmetric.

In the example of heights of 50 Year 10 girls, the mean (164cm) is very close to the value of the median (163.5cm). This is because the distribution is roughly symmetric (see the previous stem and leaf table).

However, one number can alter the mean without affecting the median.

Consider the following sets of data that represent the number of goals scored by 3 players in 11 netball matches.

EXAMPLE

1. Player 1: 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3

   Mean = 22/11 = **2**

   Median: **2**

   Player 2: 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 4

   Mean = 23/11 = **2.1**

   Median = **2**

   Player 3: 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 14

   Mean = 33/11 = **3**

   Median = **2**

The 3 sets of data are identical except for the last observation values (3, 4 and 14).

The median does not alter, because it is only dependent on the middle observation's value. Whereas, the mean does change, because it is dependent on the *average* value of *all* observations. So, in the above example, as the last observation's value increases, so does the mean.

In the 3rd set, the value of 14 is very different from any other values. When an observation is very different from all other observations in a data set it is called an outlier (see page 89).

In some cases, outliers can occur due to error or deliberate misinformation and, as a result, the measure of central tendency that is used should not include them. In other cases, outliers can be significant pieces of data, so the measure of central tendency used should include them.

2. When house prices are referred to in newspapers, the median price is quoted. Why is this measure used and not the mean?

   There are many moderately priced houses, but also some expensive ones and a few very expensive ones. If the mean figure was given, it could be quite high as it responds to prices of more expensive houses. The median gives a more accurate and realistic value of the prices faced by most people.

3. The ABS uses the median to calculate the centre of a population's age distribution. For the Melbourne Statistical Division (MSD) the median age of a person at the time of the 1996 Census was 33 years. Why is this measure used and not the mean?

   The mean would include all extreme age values, and thus be influenced by them. In this case, the median gives a better indication of centre. (The mean age was 35 years.)

4. In cricket, a batter's average is calculated by adding the number of runs scored and dividing by the number of times they have been dismissed. Consider two batters, X and Y. Both are dismissed five times.

X scores: 0, 0, 0, 0, and 200
Y scores: 34, 36, 39, 42, and 44

   Batter X's average is 40 while that of Y is 39. However, X's average has been influenced by a large score of 200: in this case an outlier. A better indication of batting performance may be the median.

   For X the median is 0, while for Y it is 39. It seems clear from this that Y is a better batter. But what about X's great score of 200? Perhaps a better description of X's batting performance would be to say that, 'X scored 200 runs in one innings, but in the other four innings X's average was 0'.

MODE

In a set of data:

DEFINITION

MODE	= THE MOST FREQUENTLY OBSERVED VALUE

A set of data can have more than one mode. The mode does not necessarily give much indication of a data set's centre. However, it is often close to the mean and median, and will be so if the data has a normal or near normal distribution.

NOMINAL OR DISCRETE VARIABLES

For nominal or discrete variables, the mode is simply the most observed value. To work out the mode, observations do not have to be placed in order, although for ease of calculation it is advisable to do so.

EXAMPLE

1. From Tony Modra's goal kicking figures: 7, 5, 0, 7, 8, 5, 5, 4, 5, and 1 goals in 10 matches, find the mode. The mode is 5, because this value occurred the most often (4 times). This can be interpreted to mean that if one match was selected at random, a good guess would be that Tony would kick 5 goals.

2. In 12 matches a netball player scored 14, 14, 15, 16, 14, 16, 16, 18, 14, 16, 16, and 14 goals. What is the mode? In this case there are two modes, 14 and 16, because both of them occur the most often (5 times).

3. The following data set represents the number of home runs scored by a softball player in 14 matches. Find and compare the mean, median and mode.

0, 0, 1, 0, 0, 2, 3, 1, 0, 1, 2, 3, 1, 0

In order: 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 2, 2, 3, 3

The mode is 0, as this value occurs most often. If one match was selected at random, the mode tells us that a good guess would be that the player would not score a home run.

MEAN

$\bar{X} = \frac{\sum x}{n}$

= 14/14

= **1**

MEDIAN

(14+1)/2 = 7.5th value

= (1+1)/2

=**1**

Therefore, it can be said that on average (mean), the player will score one home run per match; even though the mode indicates he or she doesn't score a home run in a lot of matches. So, in this case, the mode does not provide a useful measure of the data's centre.

GROUPED VARIABLES (CONTINUOUS OR DISCRETE)

When continuous or discrete variables are grouped in tables, the mode is defined as the class interval where most observations lie. This is called the *modal-class interval.*

In the example of heights of 50 Year 10 girls, the *modal-class interval* would be 160-<165cm, as this interval has the most observations in it.

NOTE: For numeric variables the mode is not often used as a measure of central tendency. However, for nominal variables the mode is useful as the mean and median do not make sense.

EXERCISES

1. For the following sets of data find the:

i) Mean ii) Median iii) Mode

a) 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0 (to 1 decimal place)

b) 2, 1, 2, 3, 1, 3, 0, 2, 4, 2, 2

c) 2.4, 3.9, 1.8, 1.7, 4.0, 2.1, 3.9, 1.5, 3.9, 2.6

d) 153.8, 154.7, 156.9, 154.3, 152.3, 156.1, 152.3

2. For the following sets of data find the:

i) Mean ii) Median iii) Mode

iv) Describe briefly the relative positions of the mean, median and mode of each data set.

a)

x	f
-2	3
-1	7
0	8
1	5
2	4

b)

x	f
6.3	2
6.4	1
6.5	6
6.6	5
6.7	13
6.8	4

c)

x	f
1	15
2	5
3	3
4	1
5	2

3. For each of the following stem and leaf tables find the:

i) Median ii) Modal-class interval

a)

Stem	Leaf
2	2 3 8
3	1 1 4 2
4	2 2 3 5 8 9 9
5	2 4 7 7 8
6	0 3 2
7	4

4|2 represents 42

b)

Stem	Leaf
0	2
0	5 6 8
1	0
1	5 5 6 6 7 8 8 9
2	0 0 0 1 1 2 3 3 3 4 4 4
2	6 6 7 8 8 9 9
3	0 4
3	5 6 7 7 8

2|2 represents 22

4. The population increase in Queensland during 1986 to 1995 is given in the table below:

Year	Increase
1986	53 377
1987	52 170
1988	67 000
1989	90 332
1990	72 681
1991	65 226
1992	76 777
1993	83 657
1994	77 753
1995	82 892

a) Calculate the mean population increase for the years 1986-95.

b) Calculate the median population increase for the years 1986-95.

c) Do you think the difference in these two measures is significant? Give reasons for your answer, and explain which result gives a better indication of the data's centre.

d) For what purposes would the Queensland Government use measures such as these?

5. The marks out of 10 for forty students who attempted a maths test were recorded as follows:

9, 10, 7, 8, 9, 6, 5, 9, 4, 7, 1, 7, 2, 7, 8, 5, 4, 3, 10, 7,
3, 7, 8, 6, 9, 7, 4, 2, 3, 9, 4, 3, 7, 5, 5, 2, 7, 9, 7, 1.

a) Prepare a frequency table of the scores.

b) Using the table, calculate the mean, median and mode.

c) How would you interpret these results?

6. The number of people unemployed at the time of the 1996 Census in Tasmania is given in the table below.

Age group	Number unemployed
15-19	3 688
20-24	4 031
25-34	5 432
35-44	4 360
45-54	3 162
55-64	1 702

a) Copy the table and by first finding the mid-point of each interval, calculate the average age of an unemployed person in Tasmania.

b) What is the modal-class interval?

c) In what age group does the median lie?

d) Briefly discuss the comparison between these three results.

e) Why do you think the number of unemployed decreases after the age group 25-34?

f) How might social welfare organisations use these figures?

7. A random analysis of 100 married men gave the following distribution of hours spent per week doing unpaid household work.

Hours	Number of men
0 - <5	1
5 - <10	18
10 - <15	24
15 - <20	25
20 - <25	18
25 - <30	12
30 - <35	1
35 - <40	1

a) Copy the table and include columns to find the end-point of each interval, calculate cumulative frequency and cumulative percentages.

b) Draw the ogive with cumulative frequency as the y-axis.

c) From the curve, find an approximate median value. What does this value indicate?

d) What is the modal-class interval?

e) Calculate the mean. What does this value indicate?

f) Briefly describe the comparison between mean, median and mode values.

g) How might you find out whether women spent more hours per week than men doing unpaid household work?

8. The 1996 Census table below shows annual income of people aged 15 years or more in Western Australia .

Income ($)	Persons
0 - 2 079	114 195
2 080 - 4 159	44 817
4 160 - 6 239	45 862
6 240 - 8 319	139 611
8 320 - 10 399	114 192
10 400 - 15 599	148 276
15 600 - 20 799	123 638
20 800 - 25 999	121 623
26 000 - 31 199	103 402
31 200 - 36 399	73 463
36 400 - 41 599	59 126
41 600 - 51 999	68 747
52 000 - 77 999	56 710

a) What is the modal-class interval?

b) Copy the table into your books and include columns to find the upper end-point of each interval, calculate cumulative frequencies and cumulative percentages.

c) Draw the ogive.

d) From the curve, give an approximate value for the median annual individual income.

e) Calculate the mean annual income. (Hint: in the above table, the interval 2,080 - 4,159 actually represents 2,080 - <4,160, so the mid-point is 3,120.)

f) Describe the comparison of mean, median and mode values.

g) Which measure gives the most accurate picture of the data's centre?

h) What type of organisation would use information such as this?

CLASS ACTIVITIES

1. Measure the height of each student in your class to the nearest centimetre. Are there any outliers? Use an appropriate method to find the mean, median and mode. Compare all three measures. Which value gives the best measure of central location and why? Which organisations or companies would find such statistics useful?

2. Find out what your school's student population or your year level's population has been for the last 10 years. Are there any outliers? Use an appropriate method to find the mean, median and mode. Compare all three measures. Which value gives the best measure of central location and why? How would your school or the Education Department use such statistics?

3. Find from your school's records the final scores of your favourite school sport. Collect the scores, both for and against, for the last ten years. (If the data is not available, use data for your favourite sporting team.)

   What was the mean final score, both for and against, for the last ten years?

   What was the median final score, both for and against, for the last ten years?

   Are any of the mean final scores similar to the corresponding median final score?

   What can be said about the distributions given these values?

   What are some of the problems you might come across in trying to use statistics to compare school or other sports teams of the past with those of today?

4. For ordinal data, can you think of occasions where the mode would be of more use than the median or mean? Discuss as a class.

MEASURES OF SPREAD

Mean, median and mode give locations of a data set's centre, but a data description will be more comprehensive if you also know the spread. (A basic numerical description of a data set requires a measure of both centre and spread.) Measures of spread include range, quartiles, mean and standard deviations, and variance.

RANGE

DEFINITION

Range is the actual spread of data, and hence includes any outliers. Thus, in any data set:

RANGE	= DIFFERENCE BETWEEN HIGHEST AND LOWEST OBSERVED VALUES

The range can be expressed as an interval such as 4-10, where 4 is the lowest value and 10 is highest. Often it is expressed as interval width; that is, the range of 4-10 is 6. The latter convention will be used throughout this section.

The disadvantage of using range is that it does not measure the spread of the majority of values in a data set; rather, it measures spread between highest and lowest values. As a result, other measures are required to give a better picture of data spread.

QUARTILES

Quartiles, as the name suggests, divide data into four equal sets.

When observations are ordered in ascending order according to their value, the first or *lower quartile*, $\mathbf{Q_1}$, is the value of the observation at or below which one-quarter (25%) of observations lie.

The second quartile, $\mathbf{Q_2}$, is the *median* at or below which half (50%) of observations lie.

The third or *upper quartile*, $\mathbf{Q_3}$, is the value of the observation at or below which three-quarters (75%) of the observations lie.

The median divides the data into two equal sets:

- the lower quartile is the value of the middle of the first set, and
- the upper quartile is the value of the middle of the second set.

INTERQUARTILE RANGE

The difference between upper and lower quartiles ($Q_3 - Q_1$) also indicates the spread of a data set. This is called the *interquartile range*. The interquartile range spans 50% of a data set, and eliminates the influence of outliers because, in effect, the highest and lowest quarters are removed. Thus:

INTERQUARTILE RANGE	=	**DIFFERENCE BETWEEN UPPER AND LOWER QUARTILES**

EXAMPLE

1. A computer salesperson, X, sells the following number of computers in 12 months: 34, 47, 1, 15, 57, 24, 20, 11, 19, 50, 28, 37.

Find the: a) range b) median

c) upper and lower quartiles d) interquartile range

Answers.

a) Range = difference between the highest and lowest values

= 57 - 1

= **56**

b) Putting the values in order gives:

1, 11, 15, 19, 20, 24, 28, 34, 37, 47, 50, 57.

Median = (12 + 1) ÷ 2 = 6.5th value

= (6th + 7th observations) ÷ 2

= (24 + 28) ÷ 2

= 52 ÷ 2

= **26**

c) Lower quartile = value of middle of 1st half of data

$\mathbf{Q_1}$ = the median of 1, 11, 15, 19, 20, 24

= (3rd + 4th observations) ÷ 2

= (15+ 19) ÷ 2

= 17

Upper quartile = value of middle of 2nd half of data

$\mathbf{Q_3}$ = the median of 28, 34, 37, 47, 50, 57

= (3rd + 4th observations) ÷ 2

= (37+47) ÷ 2

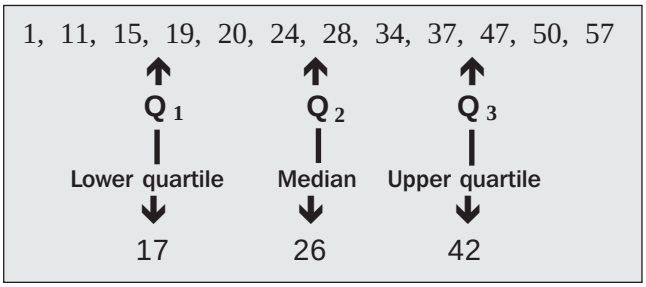
= 42

d) Interquartile range = $Q_3 - Q_1$

= 42-17

= **25**

These results can be summarised as follows:



Note: This example has an even number of observations. The median, Q_2 , lies between the centre two observations (24 and 28), so the calculation of Q_1 includes the observation 24 as it is below Q_2. Similarly, 28 is also included in the calculation of Q_3 as it is above Q_2 .

Consider an odd number of observations such as l, 2, 3, 4, 5, 6, 7. Here the value of Q_2 is 4. As the location of the median is right on the fourth observation, this value is *not* included in calculating Q_1 and Q_3 , as we are interested only in the data above and below Q_2 .

FIVE NUMBER SUMMARY

The *median* describes one location of a data set's centre. The *upper and lower quartiles* span the middle half of a data set, and hence provide one measure of spread. The *highest and lowest observations* provide additional information about how far the data actually spread.

These values, when presented together and ordered from lowest to highest, are called a *five number summary*. So, from the previous example, the five number summary would be:

1 ■ 17 ■ 26 ■ 42 ■ 57

BOX AND WHISKER PLOTS

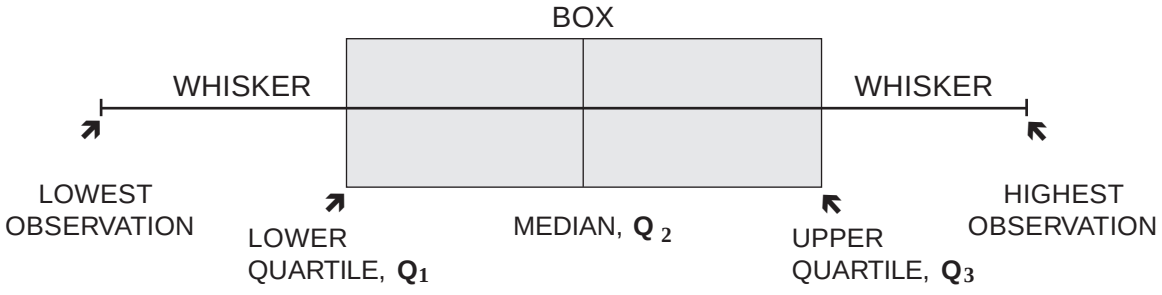
A box and whisker plot (sometimes called a boxplot) is a graph of a five number summary. It does not show a distribution in as much detail as a stem and leaf plot or histogram.

However, box and whisker plots are ideal for comparing similar distributions at a glance. The centre, spread and overall range are immediately apparent. They can also help detect symmetrical or skewed distributions.

In a box and whisker plot:

- the ends of the box are the upper and lower quartiles, so the box spans the interquartile range;
- the median is marked by a vertical line inside the box; and
- the whiskers are the two lines outside the box that extend to the highest and lowest observations.

It therefore looks like:



EXAMPLE

1. Another computer salesperson, Y, sells the following numbers of computers in 12 months: 51, 17, 25, 39, 7, 49, 62, 41, 20, 6, 43, 13.

a) Give a five number summary of Y's sales.

b) Make 2 boxplots, one for X's sales (page 160) and one for Y's.

c) Briefly describe the comparisons.

Answers.

a) First you must find the median. The data in order are:
6, 7, 13, 17, 20, 25, 39, 41, 43, 49, 51, 62.

Median = (12 + 1) = 6.5th value

= (6th + 7th observations) ÷ 2

= (25 + 39) ÷ 2

= **32**

There are 6 numbers below the median, namely:
6, 7, 13, 17, 20, 25.

Q_1 = the median of these 6 items

= (6 + 1) ÷ 2 = 3.5th value

= (3rd + 4th observations) ÷ 2

= (13+17) ÷ 2

= **15**

There are 6 numbers above the median, namely:
39, 41, 43, 49, 51, 62.

Q_3 = the median of these 6 items

= (6 + 1) ÷ 2 = 3.5th value

= (3rd + 4th observations) ÷ 2

= (43 + 49) ÷ 2

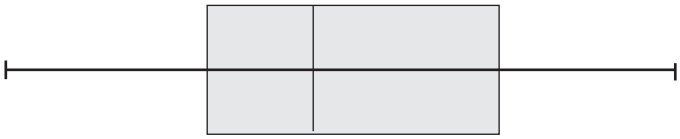
= **46**

Therefore, the five number summary is:

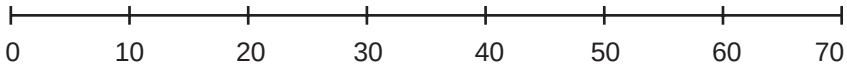
6 ■ 15 ■ 32 ■ 46 ■ 62

b) Box and whisker plots can be drawn either vertically or horizontally.

SALESPERSON X



SALESPERSON Y



c) Y's highest and lowest sales are higher than X's corresponding sales, and Y's median sales figure is higher than X's.

This suggests that Y is a consistently higher seller.

MEAN DEVIATION

Mean deviation is a better measure of spread than range because it compares all data with the mean and then averages the result. Thus:

DEFINITION

MEAN DEVIATION	=	AVERAGE OF SUM OF ABSOLUTE VALUES OF DEVIATIONS FROM MEAN

A step by step approach to finding the mean deviation is:

Calculate the mean.
▪
Subtract the mean from each observation.
▪
Change all the negative values to positive ones.
▪
Add these absolute values.
▪
Divide by the number of observations.

EXAMPLE

1. In a large school, the numbers of students absent on each day of a particular week were: 40, 30, 45, 50, 35.

Find the: a) mean

b) absolute deviation of each value from the mean

c) mean deviation

a) $\overline{X} = \frac{\sum x}{n}$

$= 200 \div 5$

$= \mathbf{40}$

b)

Day	Number absent (x)	Deviation ($x - \bar{x}$)	Absolute deviations
M	40	40-40 = 0	0
T	30	30-40 = -10	10
W	45	45-40 = 5	5
Th	50	50-40 = 10	10
F	35	35-40 = -5	5
			30

c) Mean deviation = sum of absolute deviations ÷ no. of observations

= 30÷5

= 6

Thus, the mean deviation is 6. In other words, the difference between the mean and each observation is, on average, 6.

A high mean deviation value indicates a wider spread of data values, while a low value indicates less variability in the data. If all data values are equal to the mean then it follows that the mean deviation is zero.

Mean deviation is easily interpreted and relatively simple to calculate. However, mean deviation and range are rarely used in practice to find the spread of a set of data.

Two other measures are preferred because of their wider mathematical uses in other areas of statistics. These measures are the *variance*, or more commonly, the square root of the variance: the *standard deviation*.

VARIANCE AND STANDARD DEVIATION

DEFINITION

Variance (symbolised by s^2) and standard deviation (symbolised by s) are similar in calculation to the mean deviation. However, instead of taking absolute values between the mean and each observation, the square of the values is used.

Variance involves squaring deviations, so it does not have the same unit of measurement as the original observations. For example, lengths measured in metres (m) have a variance measured in metres squared (m^2). Thus:

VARIANCE, S^2	= AVERAGE SQUARED DEVIATION OF VALUES FROM MEAN

Taking the square root gives us back the units used in the original scale. This is the standard deviation. Thus:

STANDARD DEVIATION, S	= SQUARE ROOT OF THE VARIANCE

Standard deviation is the measure of spread most commonly used in statistical practice when the mean is the measure of centre. Thus it measures spread about the mean. Because of its close links with the mean, standard deviation can be seriously affected if the mean is a poor measure of location. The standard deviation is also influenced by outliers; it is a good indicator of the presence of outliers because it is so sensitive to them. Therefore, the standard deviation is most useful for symmetric distributions with no outliers (normal distributions).

Standard deviation is useful when comparing the spread of two data sets. The data set with the smaller standard deviation has a narrower spread of measurements about the mean and, therefore, usually has comparatively fewer high or low values.

So, an item selected at random from a data set whose standard deviation is low has a better chance of being close to the mean than has an item from a data set whose standard deviation is high.

PROPERTIES OF STANDARD DEVIATION

When using standard deviation keep the following properties in mind.

- Standard deviation is only used to measure spread about the mean.
- Standard deviation is never negative.
- Standard deviation is sensitive to outliers. A single outlier can raise the standard deviation a great deal, distorting the picture of spread.
- The greater the spread, the greater the standard deviation.
- If all values of a data set are the same the standard deviation is zero.

When analysing normally distributed data, standard deviation can be used with the mean to calculate intervals within which data lie.

- about 68% of the data lie in the interval: $\bar{x} - s < x < \bar{x} + s$
- about 95% of the data lie in the interval: $\bar{x} - 2s < x < \bar{x} + 2s$
- about 99% of the data lie in the interval: $\bar{x} - 3s < x < \bar{x} + 3s$

where: $\bar{x}$ = mean; and
s = standard deviation

DISCRETE VARIABLES

The variance for a discrete variable made up of n observations is defined by:

$$s^2 = \frac{\sum (x - \bar{x})^2}{n}$$

The standard deviation for a discrete variable made up of n observations is the positive square root of the variance and is defined by:

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

A step by step approach to finding the standard deviation for a discrete variable is:

Calculate the mean.

Subtract the mean from each observation.

Square each result.

Add these squares.

Divide this sum by the number of observations.

Take the positive square root.

EXAMPLE

1. The weights (in grams) of 8 eggs are: 60, 56, 6 l, 68, 51, 53, 69, 54. Find the standard deviation.

First the mean must be calculated:

$$\bar{x} = \frac{\sum x}{n}$$

$$= 472 \div 8$$

$$= \mathbf{59}$$

Weight (x)	Deviation, $(x - \bar{x})$	$(x - \bar{x})^2$
60	1	1
56	-3	9
61	2	4
68	9	81
51	- 8	64
53	- 6	36
69	10	100
54	-5	25
472		**320**

From the above table:

$$\sum (x - x)^2 = 320$$

Thus, to calculate the standard deviation:

$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

$$= \sqrt{\frac{320}{8}}$$

$$= \mathbf{6.32} \text{ grams}$$

FREQUENCY TABLE (DISCRETE VARIABLES)

The formulas for variance and standard deviation change slightly if observations are grouped into a frequency table. Squared deviations are multiplied by each frequency's value, and then the sum of these results is calculated.

The variance for a discrete variable in a frequency table is defined by:

$$S^2 = \frac{\sum (x - \bar{x})^2 f}{n} \qquad \text{where: } n = \sum f$$

The standard deviation for a discrete variable in a frequency table is defined by:

$$S = \sqrt{\frac{\sum (x - \bar{x})^2 f}{n}}$$

A step by step approach to finding the standard deviation for a discrete variable in a frequency table is:

- Tally the x variables.
- Calculate and sum the frequencies.
- Multiply the frequencies with the x variables.
- From this, calculate the mean.
- Subtract the mean from each observation.
- Square each result.
- Multiply each square by the frequencies.
- Sum the results.
- Divide this sum by the sum of the frequencies.
- Take the positive square root.

1. Thirty graziers were asked how many shearers they hire during a shearing season. Their responses follow:

4, 5, 6, 5, 3, 2, 8, 0, 4, 6, 7, 8, 4, 5, 7, 9, 8, 6, 7, 5, 5, 4, 2, 1, 9, 3, 3, 4, 6, 4.

Shearers (x)	Tally	f	xf	$(x - \bar{x})$	$(x - \bar{x})^2$	$(x - \bar{x})^2 f$
0	I	1	0	-5	25	25
1	I	1	1	-4	16	16
2	II	2	4	-3	9	18
3	III	3	9	-2	4	12
4	𝍸 I	6	24	-1	1	6
5	𝍸	5	25	0	0	0
6	IIII	4	24	1	1	4
7	III	3	21	2	4	12
8	III	3	24	3	9	27
9	II	2	18	4	16	32
		30	**150**			**152**

To calculate the mean: $\bar{x} = \frac{\sum xf}{\sum f}$

$= 150 \div 30$

$= \mathbf{5}$

To calculate the standard deviation:

$$S = \sqrt{\frac{\sum (x - \bar{x})^2 f}{n}}$$

$$= \sqrt{\frac{152}{30}}$$

$= \mathbf{2.25}$

GROUPED VARIABLES (CONTINUOUS OR DISCRETE)

2. A group of 220 Year 10 students were asked how much time they spent watching television per week. The results are given below. Calculate the mean and standard deviation of hours spent watching television by the 220 students.

Hours	No. of students
10-14	2
15-19	12
20- 24	23
25- 29	60
30- 34	77
35-39	38
40- 44	8

First the mid-point of time intervals must be found. The number of students is the frequency. The mean can now be calculated.

$$\bar{X} = \frac{\sum xf}{\sum f}$$

$$= 6{,}670 \div 220$$

$$= \mathbf{30.32}$$

Then the calculations xf, $(x - \bar{x})$, $(x - \bar{x})^2$, and $(x - \bar{x})^2 f$ are made:

Hours	Mid-point (x)	f	xf	$(x - \bar{x})$	$(x - \bar{x})^2$	$(x - \bar{x})^2 f$
10 - 14	12.5	2	25.0	-17.82	318	636
15 - 19	17.5	12	210.0	-12.82	164	1 968
20 - 24	22.5	23	517.5	-7.82	61	1 403
25 - 29	27.5	60	1 650.0	-2.82	8	480
30 - 34	32.5	77	2 502.5	2.18	5	385
35 - 39	27.5	38	1 425.0	7.18	52	1 976
40 - 44	42.5	8	340.0	12.18	148	1 184
		220	6 670.0			8 032

Standard Deviation:

$$S = \sqrt{\frac{\sum (x - \bar{x})^2 f}{n}}$$

$$= \sqrt{\frac{8{,}032}{220}}$$

$$= \sqrt{36.51}$$

$$= \mathbf{6.04}$$

NOTE: When a variable is grouped by class intervals, it is assumed that all observations within each interval are equal to the mid-point of the interval. Thus, the spread of observations within each interval is ignored. Therefore, the standard deviation will *always* be less than the true value and should be regarded as an approximation.

3. Assuming the frequency distribution is approximately normal, calculate the interval within which 99% of the previous example's observations would be expected to occur.

$\bar{X} = 30.32$ $s = 6.04$

The interval is given by:

$$\bar{x} - 3s < x < \bar{x} + 3s$$

That is: $30.32 - (3 \times 6.04) < x < 30.32 + (3 \times 6.04)$

$30.32 - 18.12 < x < 30.32 + 18.12$

$\mathbf{12.20 < x < 48.44}$

This means that there is about a 99% certainty that an observation will lie between 12 hours and 48 hours. That is, a student in the sample will watch between 12 and 48 hours of television each week.

SUMMARY

There are several ways to describe the centre and spread of a distribution. One is to use a five number summary that uses the median as its centre and gives a brief picture of distribution.

Another method is to use the mean and standard deviation. This technique is best used with symmetric distributions with no outliers.

Despite this restriction, the mean and standard deviation are much more commonly used than the median and five number summary. The reason for this is that many natural phenomena can be approximately described by a normal distribution.

For normal distributions, the mean and standard deviation are the best measures of centre and spread.

EXERCISES

1. For the following sets of data find:

   i) the range ii) the mean deviation

   a) 6, 8, 11, 15, 24, 38

   b) 11, -6, -2, 16, 9, -8, 17, 19

   c) 6.4, 3.8, 5.9, 4.7, 5.3, 7.1, 3.2

2. The number of marriages registered in New South Wales from 1987 to 1996 were as follows:

Year	Number of marriages (x)
1987	40,650
1988	40,812
1989	41,300
1990	41,450
1991	39,594
1992	40,734
1993	39,993
1994	38,814
1995	37,828
1996	35,716

   Find the: a) range

   b) median

   c) upper and lower quartiles

   d) interquartile range

   e) five number summary

3. The maximum daily temperatures (in degrees Celsius) in Melbourne from April 21 to May 3 1993 were as follows:

   29.3, 29.1, 28.2, 19.1, 18.8, 22.4, 18.4, 17.0, 20.2, 25.0, 25.8, 24.1, 22.1.

   a) Find the range.

   b) Calculate the interquartile range.

   c) What is the five number summary?

   d) Draw a box and whisker plot for this data.

4. The number of industrial disputes in Queensland from 1982 to 1991 were as follows:

Year	Number of industrial disputes (x)
1982	266
1983	231
1984	223
1985	262
1986	260
1987	230
1988	191
1989	182
1990	165
1991	153

   a) Find the range.

   b) Calculate the interquartile range.

   c) What is the five number summary?

   d) Draw a box and whisker plot for this data.

   e) Calculate the mean deviation.

5. The number of basketball matches attended by 50 Perth Wildcat season ticket holders in 1997 were:

15, 10, 17, 11, 15, 12, 13, 16, 12, 14, 14, 16, 15, 18, 11, 16, 13, 17, 12, 16, 18, 15, 17, 15, 19, 13, 14, 17, 16, 15, 12, 11, 17, 16, 15, 10, 14, 15, 13, 16, 18, 15, 17, 11, 14, 17, 15, 14, 13, 16.

a) Tally the data.

b) Draw a column graph.

c) Calculate the mean, median and mode.

d) Calculate the variance and standard deviation.

e) Calculate the interval within which 95% of observations would be expected to occur.

f) Comment on the spread of the data.

SAMPLING METHODS

If you survey *every person* or a *whole set of units* in a population you are taking a *census*. However, this method is often impracticable; as it's often very costly in terms of time and money. For example, a survey that asks complicated questions may need to use trained interviewers to ensure questions are understood. This may be too expensive if every person in the population is to be included.

Sometimes taking a census can be impossible. For example, a car manufacturer might want to test the strength of cars being produced. Obviously, each car could not be crash tested to determine its strength!

To overcome these problems, *samples* are taken from populations, and estimates made about the total population based on information derived from the sample. A sample must be large enough to give a good *representation* of the population, but small enough to be manageable. In this section the two major types of sampling, random and non-random, will be examined.

RANDOM SAMPLING

In random sampling, all items have some chance of selection that can be calculated. Random sampling technique ensures that *bias* (see: page 179) is not introduced regarding who is included in the survey. Five common random sampling techniques are:

- simple random sampling,
- systematic sampling,
- stratified sampling,
- cluster sampling, and
- multi-stage sampling.

SIMPLE RANDOM SAMPLING

With simple random sampling, each item in a population has an equal chance of inclusion in the sample. For example, each name in a telephone book could be numbered sequentially. If the sample size was to include 2,000 people, then 2,000 numbers could be randomly generated by computer or numbers could be picked out of a hat. These numbers could then be matched to names in the telephone book, thereby providing a list of 2,000 people.

EXAMPLE

- A Tattslotto draw is a good example of simple random sampling. A sample of 6 numbers is randomly generated from a population of 45, with each number having an equal chance of being selected.

The advantage of simple random sampling is that it is simple and easy to apply when small populations are involved. However, because every person or item in a population has to be listed before the corresponding random numbers can be read, this method is very cumbersome to use for large populations.

SYSTEMATIC SAMPLING

Systematic sampling, sometimes called interval sampling, means that there is a gap, or interval, between each selection. This method is often used in industry, where an item is selected for testing from a production line (say, every fifteen minutes) to ensure that machines and equipment are working to specification.

Alternatively, the manufacturer might decide to select every 20th item on a production line to test for defects and quality. This technique requires the first item to be selected at random as a starting point for testing and, thereafter, every 20th item is chosen.

This technique could also be used when questioning people in a sample survey. A market researcher might select every 10th person who enters a particular store, after selecting a person at random as a starting point; or interview occupants of every 5th house in a street, after selecting a house at random as a starting point.

It may be that a researcher wants to select a fixed size sample. In this case, it is first necessary to know the whole population size from which the sample is being selected. The appropriate *sampling interval*, I, is then calculated by dividing population size, N, by required sample size, n, as follows:

$$I = N/n$$

EXAMPLE

- If a systematic sample of 500 students were to be carried out in a university with an enrolled population of 10,000, the sampling interval would be:

$$I = N/n \quad = 10{,}000/500 \quad = 20$$

Note: if I is not a whole number, then it is rounded to the nearest whole number.

All students would be assigned sequential numbers. The starting point would be chosen by selecting a random number between 1 and 20. If this number was 9, then the 9th student on the list of students would be selected along with every following 20th student. The sample of students would be those corresponding to student numbers 9, 29, 49, 69, ... 9929, 9949, 9969 and 9989.

The advantage of systematic sampling is that it is simpler to select one random number and then every 'Ith' (e.g. 20th) member on the list, than to select as many random numbers as sample size. It also gives a good spread right across the population. A disadvantage is that you may need a list to start with, if you wish to know your sample size and calculate your sampling interval.

STRATIFIED SAMPLING

A general problem with random sampling is that you could, by chance, miss out a particular group in the sample. However, if you form the population into groups, and sample from each group, you can make sure the sample is representative.

In stratified sampling, the population is divided into groups called strata. A sample is then drawn from within these strata. Some examples of strata commonly used by the ABS are States, Age and Sex. Other strata may be religion, academic ability or marital status.

EXAMPLE

- The committee of a school of 1,000 students wishes to assess any reaction to the re-introduction of Pastoral Care into the school timetable. To ensure a representative sample of students from all year levels, the committee uses the stratified sampling technique.

  In this case the strata are the year levels. Within each strata the committee selects a sample. So, in a sample of 100 students, all year levels would be included. The students in the sample would be selected using simple random sampling or systematic sampling within each strata.

Stratification is most useful when the stratifying variables are simple to work with, easy to observe and closely related to the topic of the survey.

An important aspect of stratification is that it can be used to select more of one group than another. You may do this if you feel that responses are more likely to vary in one group than another. So, if you know everyone in one group has much the same value, you only need a small sample to get information for that group; whereas in another group, the values may differ widely and a bigger sample is needed.

If you want to combine group level information to get an answer for the whole population, you have to take account of what proportion you selected from each group (see 'Bias in Estimation' on page 186).

CLUSTER SAMPLING

It is sometimes expensive to spread your sample across the population as a whole. For example, travel can become expensive if you are using interviewers to travel between people spread all over the country. To reduce costs you may choose a cluster sampling technique.

Cluster sampling divides the population into groups, or clusters. A number of clusters are selected randomly to represent the population, and then all units within selected clusters are included in the sample. No units from non-selected clusters are included in the sample. They are represented by those from selected clusters. This differs from stratified sampling, where some units are selected from each group.

Examples of clusters may be factories, schools and geographic areas such as electoral sub-divisions. The selected clusters are then used to represent the population.

EXAMPLE

- Suppose an organisation wishes to find out which sports Year 11 students are participating in across Australia. It would be too costly and take too long to survey every student, or even some students from every school. Instead, 100 schools are randomly selected from all over Australia.

  These schools are considered to be clusters. Then, every Year 11 student in these 100 schools is surveyed. In effect, students in the sample of 100 schools represent all Year 11 students in Australia.

Cluster sampling has several advantages: reduced costs, simplified field work and administration is more convenient. Instead of having a sample scattered over the entire coverage area, the sample is more localised in relatively few centres (clusters).

Cluster sampling's disadvantage is that less accurate results are often obtained due to higher sampling error (see page 62) than for simple random sampling with the same sample size. In the above example, you might expect to get more accurate estimates from randomly selecting students across all schools than from randomly selecting 100 schools and taking every student in those chosen.

M ULTI-STAGE SAMPLING

Multi-stage sampling is like cluster sampling, but involves selecting a sample within each chosen cluster, rather than including all units in the cluster. Thus, multi-stage sampling involves selecting a sample in at least two stages. In the first stage, large groups or clusters are selected. These clusters are designed to contain more population units than are required for the final sample.

In the second stage, population units are chosen from selected clusters to derive a final sample. If more than two stages are used, the process of choosing population units within clusters continues until the final sample is achieved.

EXAMPLE

- An example of multi-stage sampling is where, firstly, electoral sub-divisions (clusters) are sampled from a city or state. Secondly, blocks of houses are selected from within the electoral sub-divisions and, thirdly, individual houses are selected from within the selected blocks of houses.

The advantages of multi-stage sampling are convenience, economy and efficiency. Multi-stage sampling does not require a complete list of members in the target population, which greatly reduces sample preparation cost. The list of members is required only for those clusters used in the final stage. The main disadvantage of multi-stage sampling is the same as for cluster sampling: lower accuracy due to higher sampling error (see page 62).

NON-RANDOM SAMPLING

The types of methods described so far have all been random. That is, every item in a population has a known chance of being included in a sample.

In non-random sampling this is not the case. Indeed, one of the main criticisms of non-random sampling is: because it's non-random, bias is almost certainly introduced. A sample is said to be biased if:

NOT ALL OUTCOMES HAVE A KNOWN CHANCE OF OCCURRING OR IF SOME OUTCOMES HAVE A ZERO CHANCE OF OCCURRING.

Non-random sampling is useful when descriptive comments about the *sample itself* are desired.

However, it can be difficult to draw conclusions about the population based on information derived from a sample, as samples are often *unrepresentative* of the population.

Three common non-random sampling techniques are:

- quota sampling;
- convenience sampling; and
- volunteer sampling.

QUOTA SAMPLING

Quota sampling is a type of stratified sampling in which selection within the strata is non-random.

- Consider the previous case of wanting to survey 100 students (p. 177). It was established that the strata to be used were year levels. The table below gives the number of students in each year level in the school with a population of 1,000 students:

Year level	Number of students	Percentage of students (%)	Quota of students in sample of 100
7	150	15	15
8	220	22	22
9	160	16	16
10	150	15	15
11	200	20	20
12	120	12	12
	1,000	**100**	**100**

Calculation of the quota for Year 10 students is:

Percentage of Year 10's in the school = (150 ÷ 1,000) x 100 = 15%

As 15% of the school population is in Year 10, then you would expect 15% of the sample to contain Year 10 students. Therefore, to calculate the number of Year 10's to be included in the sample:

= 15% of 100 (sample size)

= (15 ÷ 100) x 100

= **15 students**

The main difference between stratified sampling and quota sampling is in the selection of 15 Year 10 students to be included in the sample. Recall that stratified sampling would select these students using a simple random sampling or a systematic sampling method.

In quota sampling, no such technique is used. The 15 students might be selected by choosing the first 15 Year 10 students to enter school or choosing 15 students in the first two rows in a particular classroom.

However, these samples may be biased because not everyone gets a chance of selection. For example, those who come late or sit at the back of the class may be different in some way to those who do have a chance of selection. They may have different views on issues you want to survey.

Market and opinion researchers often use quota sampling. Its main advantages are that it is less costly and easier to administer than many other methods.

Quota sampling ensures that there will be a representative sample of the population *for specified criteria or strata*, in this case year level. However, the actual sample may not necessarily be selected in a random manner, and therefore, the sample may not be representative for some other important criteria.

The main argument against quota sampling, as already explained, is that it does not meet the basic requirement of randomness. Some units may have no chance of selection, or the chance of selection may be unknown. Therefore, the sample may be *biased*.

CONVENIENCE SAMPLING

Convenience sampling does not produce a representative sample of the population because people or items are only selected for a sample if they can be accessed easily and conveniently.

EXAMPLE

Examples of convenient samples include selecting:

- the first ten cars to enter a car park,
- the first ten people to walk through a turnstile at a sporting event, or
- females in the first row of a concert.

The obvious advantage of this type of sampling is its ease of use, but this is greatly offset by the sample being biased.

VOLUNTEER SAMPLING

A common method of volunteer sampling is phone-in sampling, used mainly by television and radio stations to gauge public opinion on current affairs issues such as preferred political party, capital punishment, etc. People are asked to telephone their vote on a particular issue within a certain time, with no limit to the number of people who can call in.

Unfortunately, there is also no limit to the number of times a person may phone through their vote. This is a major reason why it is highly unlikely that this type of sampling will produce a representative sample. As well, people who tend to call in for these surveys may have quite different views from those who do not to call in.

EXAMPLE

- A television station may ask viewers to phone in to give their preferred opinion on whether Australia should become a Republic. The station would give two numbers to ring: one for 'Yes' voters and the other for 'No' voters.

  They would possibly give voters three hours to call after which the lines would be closed and a conclusion formed. If 200 people called in, and 114 voted 'Yes' and 86 voted 'No', then the television station would report that 57% of callers voted 'Yes' and 43% voted 'No'. However, this may or may not represent the opinion of the whole population.

The main advantages of phone-in sampling are that it is cheap in terms of time and money, and very easy to monitor and control.

However, the chance that the sample will be *biased* is very high because only those with a telephone can vote, and only those watching television or listening to radio at the time would be aware of the survey. As mentioned above, each person can make any number of calls registering their vote, and those not interested in calling will not be included.

ESTIMATION

Estimation is a mathematical technique for producing information about a population based on a sample of units from that population. Different sampling techniques require different estimation techniques.

Estimation allows you to derive measures of location, spread, and totals for the whole population. This and following pages will outline the estimation techniques for the mean and total of a population from a simple random sample only.

ESTIMATE OF POPULATION MEAN

For a simple random sample, the estimate of the population mean is the *same* as the mean of the sample:

$$\hat{\bar{X}} = \frac{\sum x}{n}$$

where: x stands for an observed value,

$\hat{\bar{x}}$ stands for the estimate of population mean,

$\sum x$ stands for the sum of all observed x values in the sample, and

n stands for the number of observations in the sample.

NOTE: Lower case x and n should be used if you are referring to a sample survey, and upper case X and N if referring to a population.

If the sample results have been summarised in a frequency table then the estimate for the population mean is again the same as the sample:

$$\hat{\bar{X}} = \frac{\sum xf}{\sum f}$$

where: x stands for an observed value,
$\hat{\bar{x}}$ stands for the estimate of the population mean,
$\sum xf$ stands for the sum of all xf values in the sample, and
$\sum f$ stands for the sum of the frequencies in the sample.

EXAMPLE

1. 10 eggs were selected randomly from a set of 200 eggs. The weights were recorded as:

   0.75, 0.70, 0.55, 0.50, 0.60, 0.65, 0.75, 0.65, 0.75 and 0.50 grams?

   What is the mean weight of the 200 eggs?

Using the formula on the previous page:

$$\hat{\bar{X}} = \frac{\sum x}{n}$$

$= 6.4 \div 10$

$=$ **0.64 grams**

ESTIMATE OF POPULATION TOTAL

For a simple random sample the estimate of *population total* is given by:

$$\hat{X} = N \frac{\sum x}{n}$$

where: x stands for an observed value,

$\hat{x}$ stands for estimated population total,

$\sum x$ stands for sum of all observed x values in the sample,

n stands for number of observations in the sample, and

N stands for total number of observations in the population.

If sample results have been summarised in a frequency table then the estimate for *population total* is given by:

$$\hat{X} = N \frac{\sum xf}{\sum f}$$

where: x stands for an observed value,

$\hat{x}$ stands for estimated population total,

$\sum xf$ stands for sum of all observed xf values in the sample,

$\sum f$ stands for sum of frequencies in the sample, and

N stands for total number of observations in the population.

BIAS IN ESTIMATION

There are a number of sources that can introduce bias into survey results: response errors, incorrect procedures and processing were discussed on pages 63-65. Bias can also be introduced if estimation is not appropriate to the sampling method used.

For example, in Exercise 3 on the next page, a stratified random sample has been drawn from all capital cities. If the proportion of Labor supporters over all capital cities is estimated as:

> total Labor supporters/total sample (531/1,220 — **43.5%**)

the estimate would be biased.

The reason is that *all units in the sample did **not** have the same chance of being selected*. For example:

> the chances of a person from Sydney being selected were about:
> 300/3,740,000 = **0.0000802**
>
> the chances of a person from Canberra being selected were about:
> 60/300,000 = **0.0002**

Thus, the estimate would be biased toward Canberra preferences.

(Note: total population figures have been taken from the 1996 Census.)

EXERCISES

1. Give an example of a simple random sample and briefly describe why it is classed in this category.

2. a) If a company has a workforce of 2,700 people, and a sample of 300 people were to be systematically surveyed, what would the sampling interval be?

   b) Choose a number at random as a starting point for the above sample. What would be the first 5 numbers in the sample? What would be the last 5 numbers in the sample?

3. The response from a stratified sample of people (18 years and over) in capital cities in Australia to the question 'Which political party would you prefer to be in power?' follows:

	MELB.	ADEL.	PERTH	SYD.	BRIS.	HOB.	DAR.	CANB.
LABOR	85	65	81	127	74	40	22	37
LIBERAL	80	70	60	135	50	40	26	13
OTHER	10	31	6	22	13	8	18	6
UNDECIDED	25	14	13	16	13	12	4	4
TOTAL	**200**	**180**	**160**	**300**	**150**	**100**	**70**	**60**

   a) In which city was the greatest percentage of people:

      i) in favour of Labor?

      ii) in favour of Liberal?

      iii) in favour of another political party?

      iv) undecided?

   b) In which city was the least percentage of people:

      i) in favour of Labor?

      ii) in favour of Liberal?

      iii) in favour of another political party?

      iv) undecided?

   c) Is it possible to estimate overall percentages for capital cities from the above table?

4. In a school, the number of students in each year level from kindergarten to Year 12 is as follows:

	K	P	1	2	3	4	5	6	7	8	9	10	11	12
Males	9	8	9	9	13	20	23	28	78	74	69	71	60	48
Females	6	8	11	10	13	18	35	34	63	62	61	88	70	56
Total	**15**	**16**	**20**	**19**	**26**	**38**	**58**	**62**	**141**	**136**	**130**	**159**	**130**	**104**

K= Kindergarten, P= Pre-school

The school has been granted a sum of money to build a new library or gym. The Principal wishes to take into consideration the opinion of students as to whether they would prefer a library or a gym.

The Principal wants to ensure that a sample survey contains students from different year levels and sexes. To determine student numbers for each year level and sex, the Principal will assume each value is to be represented proportionally.

For example, to calculate male kindergarten student numbers in the sample the Principal would use this formula:

$$\frac{\text{number of male kindergarten students}}{\text{number of total students}} \times \text{size of sample survey}$$

Once the number of students in each category has been determined, the students will be selected randomly.

a) What type of sampling technique is this called?

b) If the Principal wishes to survey 200 students, how many students of each sex and in each year level should be surveyed?

(Results should be rounded to the nearest whole number.)

CLASS ACTIVITIES

1. Use one of the random sampling methods described to obtain a random sample from your class or year level. Use this sample to find out one or more of the following:

   a) average number of children in a family,

   b) type of transport used to get to school,

   c) number of students in favour of capital punishment,

   d) amount of pocket money received,

   e) type of pets kept,

   f) number of people in a family who have had tertiary education.

2. Obtain a list showing the name and gender of each student by year level in your school. Using the stratified sampling technique, survey 20% of the school's population to find the students' favourite subject. Use the strata of year level and gender.

APPENDIX

STATISTICAL RESOURCES FOR STUDENTS

STATISTICAL RESOURCES FOR STUDENTS

A major role of the Australian Bureau of Statistics (ABS) is to encourage informed decision-making in the Australian community. This role extends to secondary schools where Australia's future decision-makers reside. The ABS offers a wide range of products and services for schools. This appendix lists a number of these products and services, and suggests other information sources that might be useful to students.

ABS PUBLICATIONS

AUSTRALIA — WORKING IT OUT!

(ABS Cat. No. 1332.2, 1990, 240pp, $19.50)

This publication has been produced for upper secondary courses in Australian Studies. It is framed around the study design of the Year 11 Victorian Certificate of Education Australian Studies course.

The theme of work and society is explored in detail, with much historical information provided. Four chapters examine the meaning and measurement of work, profile Australia's labour force, distribution of reward for work, and technology and work. There is also an introductory chapter on how to use statistics. First year economics students may find chapter 2 useful.

MEASURING AUSTRALIA'S ECONOMY

(ABS Cat. No. 1360.0, 1997, 175pp, $30.00)

This publication is a general reference and information resource for students wishing to understand the major economic indicators used to measure the economy's performance. It contains economic indicators, with additional information to assist student understanding and interpretation of presented statistics.

The economic indicators include: international accounts and trade, domestic consumption and investment, prices, incomes, and the labour force. There is a chapter devoted to international comparisons, and one that explains statistical methods and concepts used to collect, compile and present the data. *Measuring Australia's Economy* is an annual publication updated for the beginning of each academic year.

SURVIVING STATISTICS

(ABS Cat. No. 1332.0, 1991, 92pp, $12.50)

This publication is a basic guide to understanding and using statistics. Interesting examples are used to explain the collection, organisation and interpretation of statistical data. This publication will benefit anyone who wishes to have a better general understanding of statistics.

ABS LIBRARY EXTENSION PROGRAM (LEP)

School students can access ABS information through all major public libraries in Australia. Rather than travelling to your nearest ABS office, which may be hundreds of kilometres away, you may be able to go to your local library and ask for ABS information. Many universities and TAFEs now have campuses outside capital cities, and these also receive a range of ABS publications.

The ABS Library Extension Program entitles participating libraries to receive free copies of a core set of ABS publications, which can meet many of your statistical enquiries. Please contact your local library to see if they have ABS publications, or refer to the list of current LEP libraries provided on the ABS website (http:\\www.abs.gov.au). Otherwise, you can contact your state/territory's ABS Information Services section for the location of your nearest LEP library. A list of ABS offices occurs at the back of this publication.

USING THE LIBRARY EXTENSION PROGRAM

An ideal strategy is to investigate the ABS Catalogue of Publications and Products (held in most school libraries). Use the Subject Index to find the publication name and catalogue number that serves your information needs. Then ring your local LEP library to see if they have the publication. Note that smaller libraries are affiliated to larger regional libraries and can get ABS information, but have to do so through their regional library.

EXAMPLE

1. An Australian Studies, Social Studies or Economics student needs information on discouraged jobseekers. Looking under 'discouraged jobseekers' in the subject index of the ABS's Catalogue of Publications and Products, the student is referred to the publication *Persons Not in the Labour Force, Australia* (ABS Cat. No. 6220.0).

The student can then ring their nearest major public library and see if they have the publication available by quoting the above catalogue number. If not, the student can ask the librarian to order a copy from the ABS.

2. Environmental Studies students may want statistical information on pollution for a class project. Again, they can first refer to the ABS Catalogue (available in most school libraries) and locate 'pollution' in the subject index. They will be referred to more than one publication. By examining the information provided about each publication, they can decide which will satisfy their needs and order it as in Example 1.

ABS CLASS PRESENTATIONS

The ABS may be able to provide officers to speak to classes of students in your school. General presentations are available for upper secondary students. The content may include why Australia takes a census, information and modern society, and common mistakes in the use of statistical information.

In Victoria, more subject specific presentations can be given to Economics students. Presentations may include statistical information and issues associated with unemployment, inflation, economic growth, and 'women and the economy'. The talks can be presented in a double period and may combine some form of general presentation as mentioned above.

To inquire about possible class presentations contact the Education Servicing Officer in your State (see list of ABS offices at the end of this publication).

ABS WEBSITE (http:\\www.abs.gov.au)

The ABS Website, known as ABS Statsite, is constantly expanding as a source of information about what the ABS has available. Although most of the actual publications and other sources of data produced by the ABS are not contained on the website, it does contain:

- News about latest release ABS products and services.
- Copies of the main findings from a range of recently released ABS publications and copies of Information papers.
- Details about the ABS, such as copies of annual reports, contact details for all ABS offices, and list of all LEP libraries.

ABS Statsite also enables you to easily access information about the education services available from the ABS. By clicking on the PRODUCTS button on the ABS home page, the Education Services pages will show you:

- What's new in education at the ABS.
- The Statspak Catalogue (lists statistical resources available for school education).
- Resources for curriculum learning areas, and
- Keydata Education Toolkit.

For further information about Education Services products and services, please ring free call 1800 623 273 or (03) 9615 7041, or email: client.services@abs.gov.au .

OTHER INFORMATION RESOURCES

Apart from the ABS, there are many other sources of information that students may find useful. Three of these are:

Australian Government Publishing Service (AGPS) Bookshops: The AGPS has bookshops in all State and Territory capital cities and runs an extensive mail order service for people in rural Australia. The bookshops contain a great deal of information published by Australia's Commonwealth departments and agencies.

Industry and Trade Union Organisations: These organisations publish information about their members and issues associated with them. For industry organisations such as the Australian Mining Industry Council refer to the 'Business Organisations' section of Telstra's Yellow Pages, and for Trade Union organisations refer to the 'Trade Union' section.

Public and Community Organisations: There are literally hundreds of non-government public organisations involved in a wide range of social, economic and environmental issues. The *Directory of Australian Associations*, published by Information Australia, is a good reference guide to locate the names and addresses of public and community organisations. The directory can be found in the reference section of most public libraries.

INDEX

GLOSSARY OF STATISTICAL TERMS

Confidence interval: a specified interval, with the sample statistic at the centre, within which the corresponding population value is said to lie with a given level of confidence.

Data Item: the smallest piece of information that can be obtained from a survey or census.

Data Set: data collected for a particular study. A data set represents a collection of elements; and for each element, information on one or more characteristics is included.

Distribution: pattern of observation values in a data set.

Estimate: information about a population derived from a sample of units from that population.

Frequency: number of times an observation occurs in a data set.

Non-sampling error: inaccuracies that occur due to reporting imperfections by respondents and interviewers, and errors made in coding and processing data. These errors can occur whether information is derived from a sample or census.

Observation: a single piece of data about a variable.

Sampling error: difference between an estimate derived from a sample survey and the true value that would result if a census of the whole population was taken.

Unit: an entity about which information is being collected.

Variables: mutually exclusive characteristics such as sex, age, and employment status. Surveys often aim to describe the distribution of characteristics comprising a variable in a population.

ANSWERS TO EXERCISES

INTRODUCTION:

2. DATA - INFORMATION - KNOWLEDGE

4. Fewest = Example 4. Most = Example 2.

5. Nurse. No, the value 0 can be regarded as information.

6. Age-group = 35-39 (for both males and females)

7. A. Keating, C. Sampson, R. Jameson

8. 11 (Mark Philippoussis' fastest serve can be regarded as illustrated information)

9. Historians for research, history students for inclusion in assignments.

10. Governments, for planning health and social policies.

11. No. Statistics on the number of possessions and disposals do not necessarily accurately measure a player's overall contribution.

12. Example 4

13. None show all the individual observations collected. (In Example 4, a number of players' service speeds would have been collected but only 11 are shown.)

INFORMATION STUDIES:

DATA COLLECTION

2. A sample survey is less expensive and quicker to undertake.

4. The size of population to be surveyed, speed with which you want results, need for small area information, money and personnel you have to conduct data collection, and degree of accuracy you want from the results.

DATA PROCESSING

1. DATA - COLLECTION - PROCESSING - INFORMATION

4. Information without editing of data is almost certainly less accurate.

5. a) Instead of 'yes' the *number* of ewes mated should be shown.

   b) You cannot be *both* 'never married' and 'divorced'!

   c) You cannot *go to work* on a motorbike and claim you also 'did not go to work'. However, someone could legitimately put a mark in both the 'Motorbike' and 'Worked at home' fields. Can you say why?

INFORMATION- PROBLEMS WITH USING

1. a) Inappropriate estimation based on an unrepresentative sample.

   b) Various problems associated with volunteer sampling (see page 182).

   c) This will always be the case!

   d) Misunderstands definition of unemployment for ABS sample survey.

   e) Possible difference in the respective definition of forest cover.

STATS MATHS: ORGANISING DATA

1. a) c
b) d
c) d
d) c
e) c
f) d
g) c
h) c
i) d
j) d
k) c
I) d

2. Various answers

3. a)

(x)	Tally	Frequency (f)
1	I	1
2	I I	2
3	~~IIII~~	5
4	I I I	3
5	I	1
		12

b) 3 occurs the most

4. a) Discrete

b)

Number of customers (x)	Tally	Frequency (f)
20	I I	2
21	~~IIII~~ I I	7
22	I I I I	4
23	I I I	3
24	~~IIII~~	5
25	I I	2
26	I I	2
		25

c) 21

d)

(x)	(f)	Relative frequency	Percentage frequency
20	2	0.08	8
21	7	0.28	28
22	4	0.16	16
23	3	0.12	12
24	5	0.20	20
25	2	0.08	8
26	2	0.08	8
	25	1.00	100

5. a) Continuous b)

Windspeed(x)	Tally	Frequency(f)
0 - <5	| | |	3
5 - <10	| | | |	4
10 - <15	𝍸 𝍸 | | | |	14
15 - <20	𝍸 𝍸 |	11
20 - <25	𝍸 | |	7
25 - <30		0
30 - <35	|	1
		40

c) 10 - <15 d)

Windspeed (x)	(f)	Relative frequency	Percentage frequency
0 - <5	3	0.075	7.5
5 - <10	4	0.100	10.0
10 - <15	14	0.350	35.0
15 - <20	11	0.275	27.5
20 - <25	7	0.175	17.5
25 - <30	0	0.000	0.0
30 - <35	1	0.025	2.5
	40	**1.000**	**100.0**

e) The most common occurring windspeed is from 10 to less than 15 knots, and has a 35% chance of occurring on any one day based on this sample of 40.

6. a)

Stem	Leaf
0	2 4 9 1 0
1	8 0 4 5 9 2 6
2	1 7 5 9 4 6 8 2 6 3
3	5 1 8 7 3
4	3 1 0

b)

Stem	Leaf
0	0 1 2 4 9
1	0 2 4 5 6 8 9
2	1 2 3 4 5 6 6 7 8 9
3	1 3 5 7 8
4	0 1 3

7. a)

Stem	Leaf
$0^{(0)}$	0 3 4
$0^{(5)}$	5 7 8 9
$1^{(0)}$	0 0 1 1 2 2 2 2 3 3 4 4 4 4
$1^{(5)}$	5 5 6 6 7 7 7 8 9 9 9
$2^{(0)}$	0 1 2 3 3 4 4
$2^{(5)}$	
$3^{(0)}$	4

b) 34 is an outlier. This was either because of a particularly windy or stormy day during 40 days of recording wind speed, or it might have been a measurement error.

c) i) The distribution has only one main peak.

ii) The distribution is very roughly symmetrical or could even be roughly skewed to the left if the outlier is removed. It is probably best to call it an irregular shape.

iii) The centre is 24 knots.

8. a) Discrete b)

Stem	Leaf
0	7 8 8
1	0 4 5 7 7 7
2	0 0 3 4 6 6 6 8 9 9 9
3	0 0 1 1 2 2 2 2 2 3 6 7 8

c)

Stem	Leaf
$0^{(5)}$	7 8 8
$1^{(0)}$	0 4
$1^{(5)}$	5 7 7 7
$2^{(0)}$	0 0 3 4
$2^{(5)}$	6 6 6 8 9 9 9
$3^{(0)}$	0 0 1 1 2 2 2 2 2 3
$3^{(5)}$	6 7 8

d) No

e) i) One main peak

ii) Skewed to the left

iii) 28 road fatalities

9. a) Continuous

b)

Stem	Leaf
5	7
6	1 2 2 4 4 8 8 8 9
7	0 2 3 6 8 8 9
8	1 1 8 9

c) No, the stems are not overcrowded.

d) 8.8 and 8.9 are possible outliers. They are due to particularly warm years where high minimum daily temperatures gave a high mean minimum temperature for the year.

e) The distribution has one peak, and its general shape is roughly symmetric (although this is difficult to observe with a small amount of data). The distribution's centre is 7.0°C.

10. a) Discrete

b)

Weekly salary (x)	Tally	Frequency (f)
420 - <440	I	1
440 - <460	I I I I	4
460 - <480	~~I I I I~~ ~~I I I I~~ I I I I	14
480 - <500	~~I I I I~~ I I I I	9
500 - <520	~~I I I I~~ I I I	8
520 - <540	~~I I I I~~ I I I	8
540 - <560	~~I I I I~~	5
560 - <580	I	1
		50

c) $460 - <$480

d)

Relative frequency	Percentage frequency
0.02	2
0.08	8
0.28	28
0.18	18
0.16	16
0.16	16
0.10	10
0.02	2
1.00	**100**

e) Most people in the company earn between $460 and $480 a week based on this sample of 50 people. Only 1 staff member earned over $560 a week.

f)

Stem	Leaf
43	7
44	0 1 3
45	9
46	1 1 3 3 6 6
47	0 0 0 1 3 3 6 8
48	1 4 4 6 6 7
49	0 7 9
50	2
51	1 3 4 4 7 9 9
52	1 2 3 3 5 7 8
53	9
54	2 3 6 8
55	5
56	4

g) $555 and $564 could be outliers. They exist because 2 of the 50 people surveyed may have been managers or directors and thus on higher salaries, or 2 people may have deliberately provided misleading responses.

h) i) The distribution has a number of peaks, possibly bimodal.
ii) The distribution has no symmetry nor is it skewed.
iii) The centre is between $487 and $490.

CUMULATIVE FREQUENCY AND PERCENTAGE

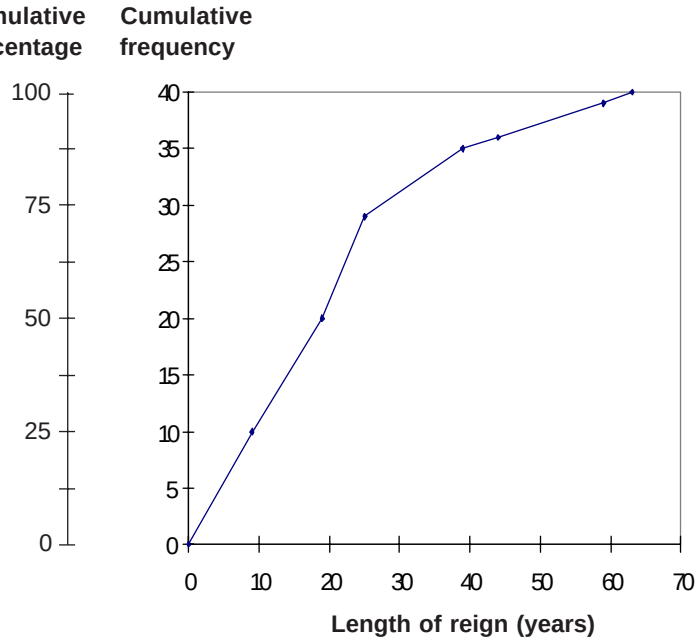
1. a & d)

Stem	Leaf	Frequency (f)	Actual upper value	Cumulative frequency	Cumulative percentage
0	0 1 2 3 5 6 6 7 9 9	10	9	10	25.0
1	0 0 2 3 3 3 3 5 7 9	10	19	20	50.0
2	0 1 2 2 2 4 4 5 5	9	25	29	72.5
3	3 5 5 5 8 9	6	39	35	87.5
4	4	1	44	36	90.0
5	0 6 9	3	59	39	97.5
6	3	1	63	40	100.0

b) Possible outliers are 56, 59 and 63. (Find out which monarch reigned for 63 years.) However, as this is factual data, they exist simply because the monarchs who reigned for this time lived longest after coming to the throne early in their lives.

c) i) Two peaks appear at the beginning of the distribution.
ii) The distribution could be said to be skewed to the right.
iii) The centre is approximately 19 years.

e)



f) 10 g) 4

h) At the time of writing, Queen Elizabeth II has reigned for 44 years. This was well above the centre of distribution and only 4 other monarchs have reigned longer.

2. a)

Stem	Leaf
$2^{(0)}$	1
$2^{(5)}$	
$3^{(0)}$	1 2 4
$3^{(5)}$	5 7 7 8 8
$4^{(0)}$	0 0 2 3 3 3 4
$4^{(5)}$	5 5 6 6 7 7 8
$5^{(0)}$	0 1 4 4
$5^{(5)}$	5 9

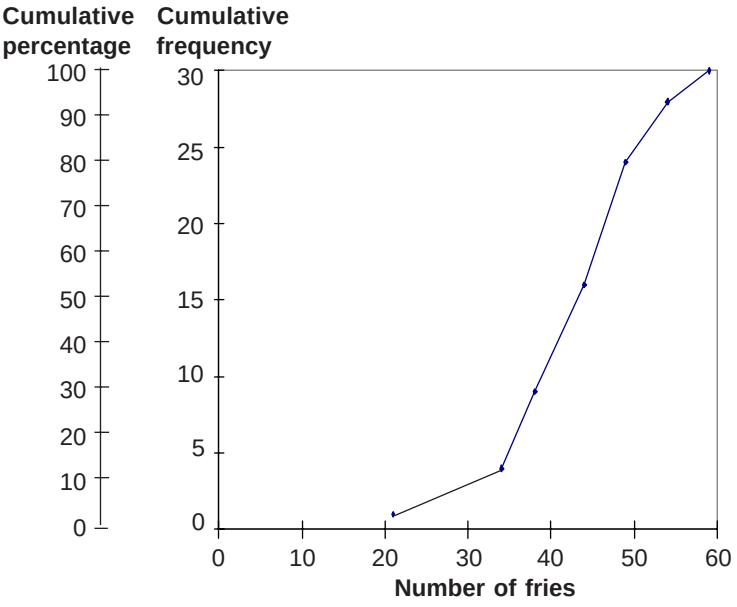
d)

Frequency (f)	Actual upper value	Cumulative frequency	Cumulative percentage
1	1	1	3.3
3	34	4	13.3
5	38	9	30.0
7	44	16	53.3
8	49	24	80.0
4	54	28	93.3
2	59	30	100.0
30			

b) 21 is an outlier. Perhaps only 21 fries were left in a batch when the student ordered fries that particular day.

c) i) Unimodal
ii) Roughly symmetric if the outlier is removed.
iii) The centre is approximately 43 fries.

e)



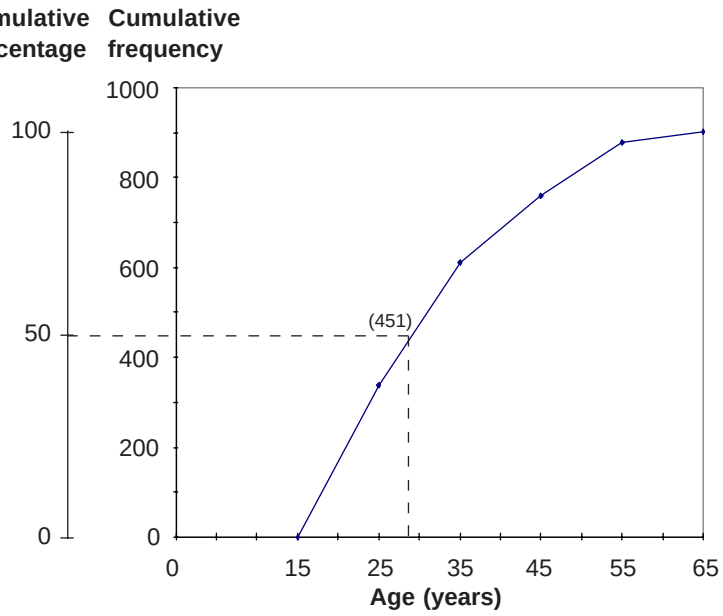
f) 9 g) 46.7% h) 44

3. a) Continuous

b)

Age group	Number of females	End-point	Cumulative frequency	Cumulative percentage
		15	0	0.0
15-24	339	25	339	37.6
25-34	273	35	612	67.8
35-44	147	45	759	84.1
44-54	121	55	880	97.6
55-64	22	65	902	100.0

c)



d) No-one under 15 years of age can be classified as unemployed.

e) 25-34, (approximately 29 years old).

f) 37.6%.

g) 2.4%.

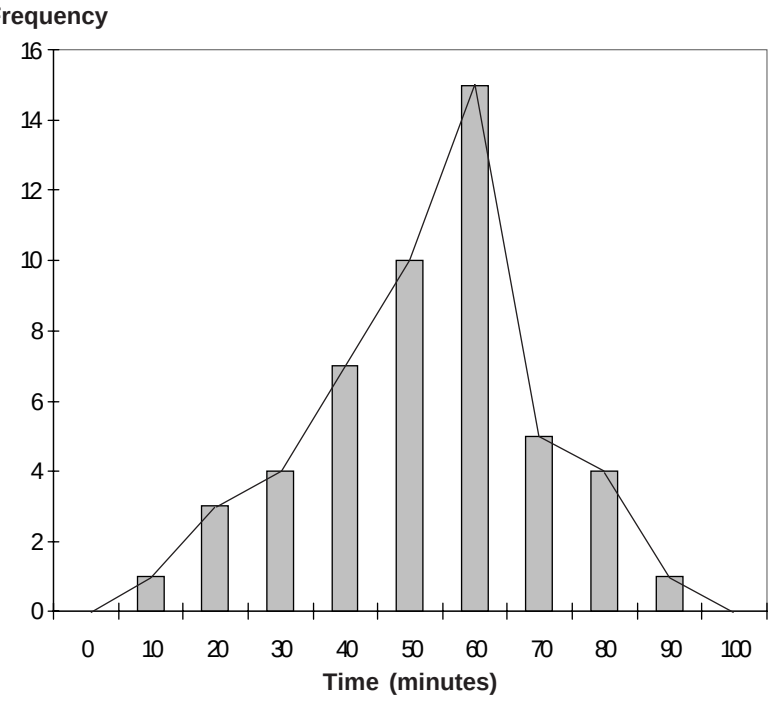
h) Governments can establish job creation schemes directed at particular age groups (in this case, the most likely would be for those under 25 years of age).

4. a) Continuous

b)

Time (x)	Tally	Frequency	Relative frequency	Percentage frequency
0 - <10		0	0.00	0
10 - <20	I	1	0.02	2
20 - <30	III	3	0.06	6
30 - <40	IIII	4	0.08	8
40 - <50	𝍸 II	7	0.14	14
50 - <60	𝍸 𝍸	10	0.20	20
60 - <70	𝍸 𝍸 𝍸	15	0.30	30
70 - <80	𝍸	5	0.10	10
80 - <90	IIII	4	0.08	8
90 - <100	I	1	0.02	2
Total		**50**	**1.00**	**100**

c)



d)

Stem	Leaf
0	
1	2
2	259
3	1378
4	0134559
5	0122566889
6	001233445567899
7	13567
8	0379
9	8

f)

Frequency (f)	End-point	Cumulative frequency	Cumulative percentage
0	10	0	0
1	20	1	2
3	30	4	8
4	40	8	16
7	50	15	30
10	60	25	50
15	70	40	80
5	80	45	90
4	90	49	98
1	100	50	100

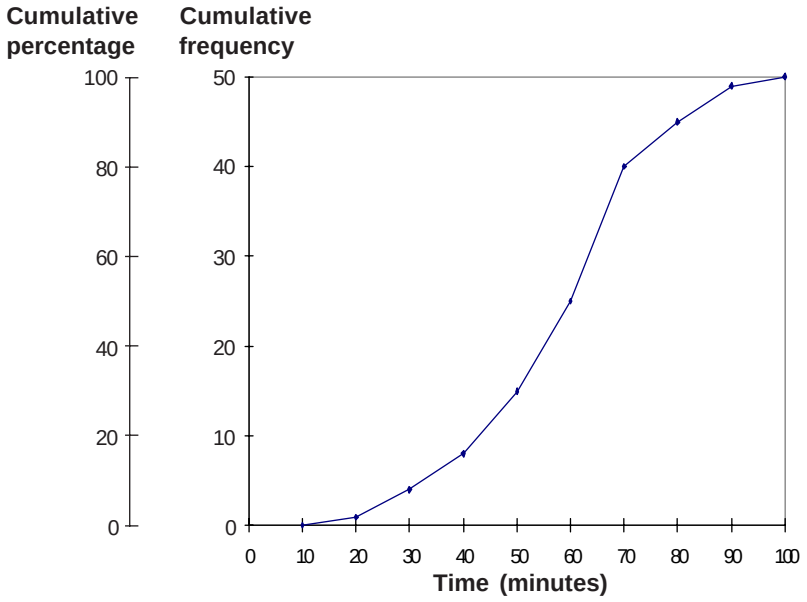
98 is a possible outlier. This person may have had difficulty in getting to work, or simply lives quite a distance from work.

e) i) Unimodal

ii) The distribution is quite symmetric.

iii) The approximate centre is 59 minutes.

g)



h) 60 - <70 minutes i) 2% j) 8

MEASURES OF LOCATION

1. a) i) 0.1 ii) 0 iii) 0

b) i) 2 ii) 2 iii) 2

c) i) 2.78 ii) 2.5 iii) 3.9

d) i) 154.3 ii) 154.3 iii) 152.3

2. a) i) 0 ii) 0 iii) 0

iv) The mean, median and mode are equal. This distribution is almost symmetrical.

b) i) 6.6 ii) 6.7 iii) 6.7

iv) Distribution is skewed left, so the mean is less than the median and therefore closer to centre. The mode and median are the same.

c) i) 1.85 ii) l iii) 1

iv) The median and mode are the same. The distribution is skewed right, so the mean is more than the median and therefore closer to centre. In b) and c) the mean has been influenced by a few low and high values respectively.

3. a) i) 48 ii) 40-49

b) i) 23 ii) 20-24

4. a) 72,186.5 b) 68,953.5

c) The measures are quite close together, given the size of each observation, hence the difference is not significant. The median probably gives the best indication of the data's centre, as there is a large diversity of observation values. The median would not be affected by the very large or very small values.

d) A government could use these measures to plan for building schools, hospitals, roads etc. It could also use them to help predict revenue intake from taxation.

5. a)

Score (X)	Tally	Frequency
0		
1	II	2
2	III	3
3	IIII	4
4	IIII	4
5	IIII	4
6	II	2
7	~~IIII~~ ~~IIII~~	10
8	III	3
9	~~IIII~~ I	6
10	II	2
		40

b) mean = 5.9, median = 7, mode = 7

c) The median is higher than the mean because most of the observations have high values. The mean is influenced by the lower scores. The mode is equal to the median.

6. a) 33.6

b) 25-34 (Note: interval sizes are not the same. If they were, the 15-24 interval would be the modal-class interval.)

c) 25-34

d) All three results lie within the same interval, but distribution is skewed to the right.

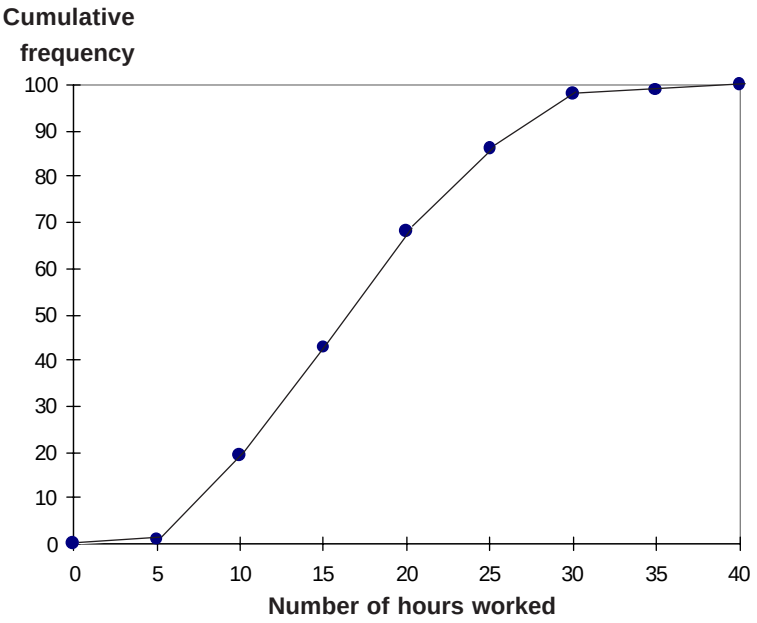
e) The younger age groups, 15-19 and 20-24, are filled with school leavers who have not yet been able to get a job, and are too young to have acquired the experience necessary to qualify for many jobs. The age groups after 25-34 contain a larger proportion of people who have left the workforce temporarily or simply retired.

f) To plan employment schemes that cater for younger people; to try to create work for a younger workforce.

7. a)

Hours	Number of men (x)	End-point	Cumulative frequency	Cumulative percentage
		0	0	0
0 - <5	1	5	1	1
5 - <10	18	10	19	19
10 - <15	24	15	43	43
15 - <20	25	20	68	68
20 - <25	18	25	86	86
25 - <30	12	30	98	98
30 - <35	1	35	99	99
35 - <40	1	40	100	100

b)



c) Median = 17 hours. The middle of the distribution is 17 hours.

d) 15 - <20 hours

e) 16.8 hours. The mean number of hours that a married man spends doing unpaid household work is 16.8 hours.

f) The mean and median are very similar, and all measures lie in the modal-class interval. The distribution is close to symmetrical.

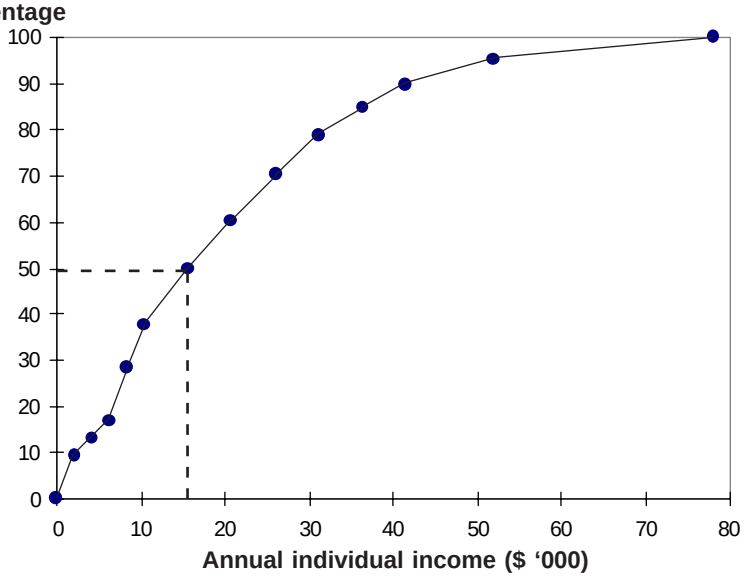
g) A similar survey could be done (possibly even surveying the wives of men who participated in this survey!), analysing the results in a similar fashion, and comparing the results.

8. a) $10,400 - $15,599. (Note that interval sizes are not the same.)

b)

Income ($)	Persons	End-point	Cumulative frequency	Cumulative percentage
		0	0	0.0
0 - 2,079	114,195	2,079	114,195	9.4
2,080 - 4,159	44,817	4,159	159,012	13.1
4,160 - 6,239	45,862	6,239	204,874	16.9
6,240 - 8,319	139,611	8,319	344,485	28.4
8,320 - 10,399	114,192	10,399	458,677	37.8
10,400 - 15,599	148,276	15,599	606,953	50.0
15,600 - 20,799	123,638	20,799	730,591	60.2
20,800 - 25,999	121,623	25,999	852,214	70.2
26,000 - 31,199	103,402	31,199	955,616	78.7
31,200 - 36,399	73,463	36,399	1,029,079	84.8
36,400 - 41,599	59,126	41,599	1,088,205	89.7
41,600 - 51,999	68,747	51,999	1,156,952	95.3
52,000 - 77,999	56,710	77,999	1,213,662	100.0

c) **Cumulative percentage**



d) The median is approximately $15,500.

e) The mean is $20,691.

f) It is difficult to compare the mode with the mean and median because of the difference between the sizes of the intervals. The mean is higher than the median because it is affected by the higher incomes. This means that the distribution is skewed to the right.

g) The median, as it is not influenced by extreme values.

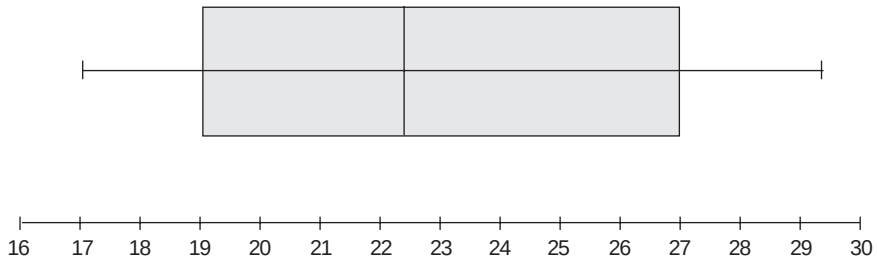
h) Some possible answers include: social welfare organisations interested in the number of low income earners; businesses interested in the number of high income earners; and governments and other service providers would use such data, especially when broken down by such characteristics as age, sex and geographic area, to locate services appropriately.

MEASURES OF SPREAD

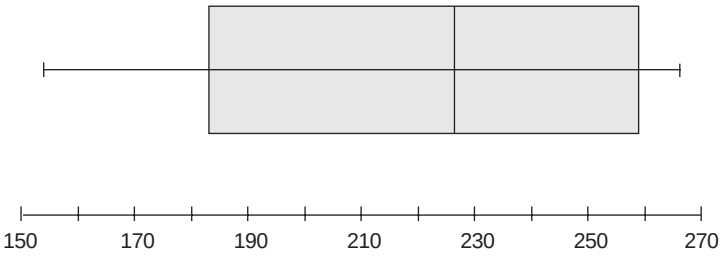
1. a) i) 32 ii) 9.3
b) i) 27 ii) 9.25
c) i) 3.9 ii) 1.11

2. a) 5,734 b) 40,321.5 c) Q_1= 38,814 Q_2= 40,812 d) 1,998
e) 35,716 38,814 40,321.5 40,812 41,450

3. a) 12.3 b) 8.05 c) 17.0, 18.95, 22.4, 27.0, 29.3
d)



4. a) 113 b) 78 c) 153, 182, 226.5, 260, 266
d)

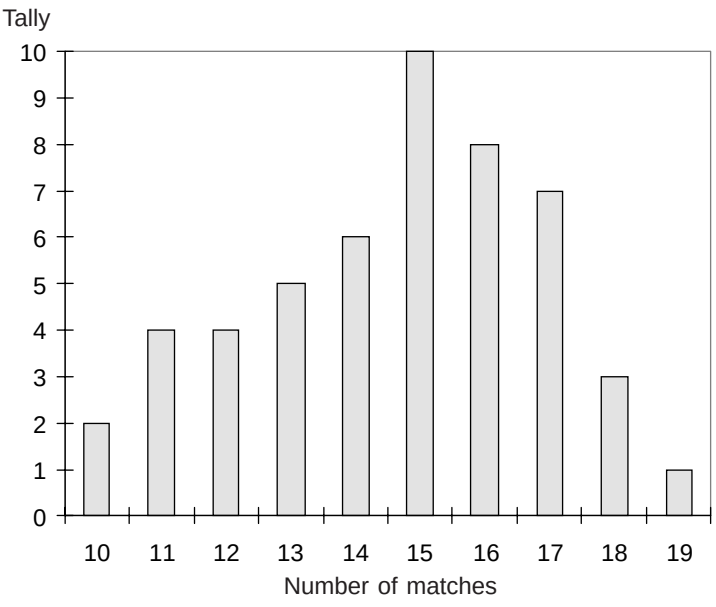


e) 34.84

5. a)

Number of matches (x)	Tally	Frequency (f)
10	II	2
11	IIII	4
12	IIII	4
13	𝍸	5
14	𝍸 I	6
15	𝍸 𝍸	10
16	𝍸 III	8
17	𝍸 II	7
18	III	3
19	I	1
		50

b)



c) mean = 14.62, median = 15, mode = 15

d) $S^2 = 4.96$, $S = 2.23$ e) $10.17 < x < 19.07$

f) The standard deviation is quite low, which indicates that the data is not widely spread about the mean. The mean and median are very close together, which indicates that the data is roughly symmetrical.

SAMPLING METHODS

1. Various answers

2. a) 9 b) Various answers

3. a) i) Canberra b) i) Darwin
ii) Sydney ii) Canberra
iii) Darwin iii) Perth
iv) Melbourne iv) Sydney

c) No, the table does not give total population figures (see page 186).

4. a) Stratified sampling

b)

	K	P	1	2	3	4	5	6	7	8	9	10	11	12
Males	2	2	2	2	2	4	4	5	15	14	13	13	11	9
Females	1	2	2	2	2	3	7	6	12	12	12	17	13	11

2133100001962
ISBN 0 642 25744 2

RRP $21.00

Produced by the Australian Bureau of Statistics