Diploma
in
Business Administration

Study Manual

Quantitative Methods

The Association of Business Executives

William House • 14 Worple Road • Wimbledon • London • SW19 4DD • United Kingdom
Tel: + 44(0)20 8879 1973 • Fax: + 44(0)20 8946 7153
E-mail: info@abeuk.com • www.abeuk.com

ABE Diploma in Business Administration

Study Manual

Quantitative Methods

Contents

Diploma in Business Administration – Part 2

Quantitative Methods

Syllabus

Aims

1. Achieve an overall understanding of how and why statistics and mathematics are used in economic and business decisions.
2. Demonstrate the ability to collect, present, analyse and interpret quantitative data using standard statistical techniques.

Programme Content and Learning Objectives

After completing the programme, the student should be able to:

1. Demonstrate an overall understanding of the data collection process.

This includes sources of data, sampling methods, problems associated with surveys, questionnaire design, measurement scales (nominal, ordinal, interval and ratio scales) and sampling error.

2. Use a range of descriptive statistics to present data effectively.

This includes the presentation of data in tables and charts, frequency and cumulative frequency distributions and their graphical representations, measures of location, dispersion and skewness, index numbers and their applications.

3. Understand the basic concepts of probability and probability distributions.

This includes the basic 'rules' of probability, expected values and the use of probability and decision trees, the binomial and Poisson distributions and their applications, and the characteristics and use of the normal distribution.

4. Apply the normal distribution and the t distribution in estimation and hypothesis testing.

This includes sampling theory and the Central Limit Theorem. The construction of confidence intervals for population means and proportions, using the standard normal distribution or the t distribution, as appropriate, and hypothesis tests of a single mean, a single proportion, the difference between two means and the difference between two proportions.

5. Use correlation and regression analysis to identify the strength and form of relationships between variables.

In correlation analysis, this includes the use of scatter diagrams to illustrate linear association between two variables, Pearson's coefficient of correlation and Spearman's 'rank' correlation coefficient and the distinction between correlation and causality. In regression analysis, students are expected to be able to estimate the 'least squares' regression line for a two-variable model and interpret basic results from simple and multiple regression models.

6. Demonstrate how time-series analysis can be used in business forecasting.

This includes the use of the additive and multiplicative models to 'decompose' time-series data, the calculation of trends and cyclical and seasonal patterns, and simple forecasting.

7. Distinguish between parametric and non-parametric methods and use the chi-squared statistic in hypothesis testing.

This includes using the chi-squared statistic as a test of independence between two categorical variables and as a test of goodness-of-fit.

8. Show how mathematical relationships can be applied to economic and business problems.

This includes the algebraic and graphical representation of demand and supply functions and the determination of equilibrium price and quantity in a competitive market. It also includes the algebraic and graphical representation of cost, revenue and profit functions, with applications of pricing and output determination (including break-even analysis).

Throughput, students will be expected to be able to define relevant terms and to interpret all results.

Method of Assessment

By written examination. The pass mark is 40%. Time allowed 3 hours.

The question paper will contain:

Eight questions of which four must be answered.

Probability tables for the binomial distribution, the normal distribution, the t distribution and the chi-squared distribution will be provided. Students may use electronic calculators, but are reminded of the need to show explicit workings.

Reading List:

Essential Reading

- Curwin, J. and Slater, R. (1996), *Quantitative Methods for Business Decisions*; Thomson Business Press

Additional Reading

- Kazmier, L. and Pohl, N. (1987), *Basic Statistics for Business and Economics*, 2nd Edition; McGraw-Hill
- Silver, M. (1997), *Business Statistics*; McGraw-Hill

Study Unit 1

Methods of Collecting Data

Contents *Page*

A. INTRODUCTION

We will start the course by seeing how we collect data. This study unit looks at the various sources of data and the numerous methods available to collect it.

Units of Measurement

The figures used in any analysis requiring measurement must be expressed in units such as metres, litres, etc. These units must be **suitable** for the substance or object being measured, e.g. the amount of coal produced at various pits should be measured in tonnes (or tons) not kilograms.

It is always necessary to deal with units of a **constant size**. Mistakes often occur, for instance, by making detailed comparisons between various months when the months contain a varying number of days.

Categorisation of Data

Any characteristic on which observations can be made is called a **variable or variate**. For example, height is a variable because observations taken are of the heights of a number of people. Variables, and therefore the data which observations of them produce, can be categorised in two ways:

(a) Quantitative/Qualitative Categorisation

Variables may be either quantitative or qualitative. Quantitative variables, to which we shall restrict discussion here, are those for which observations are numerical in nature. Qualitative variables have non-numeric observations, such as colour of hair, although, of course, each possible non-numeric value may be associated with a numeric frequency.

(b) Continuous/Discrete Categorisation

Variables may be either continuous or discrete. A **continuous variable** may take **any value** between two stated limits (which may possibly be minus and plus infinity). Height, for example, is a continuous variable, because a person's height may (with appropriately accurate equipment) be measured to any minute fraction of a millimetre. A **discrete variable**, however, can take only **certain values** occurring at intervals between stated limits. For most (but not all) discrete variables, these intervals are the set of integers (whole numbers).

For example, if the variable is the number of children per family, then the only possible values are 0, 1, 2, ... etc., because it is impossible to have other than a whole number of children. However, in Britain, shoe sizes are stated in half-units, and so here we have an example of a discrete variable which can take the values 1, 1½, 2, 2½, etc.

You may possibly see the difference between continuous and discrete variables stated as "continuous variables are measured, whereas discrete variables are counted". While this is possibly true in the vast majority of cases, you should not simply state this if asked to give a definition of the two types of variables.

Types of Data

(a) Primary Data

If data is collected for a **specific** purpose then it is known as primary data. For example, the information collected direct from householders' television sets via a microcomputer link-up to a mainframe computer owned by a television company is used to decide the most popular television programmes and is thus primary data. The Census of Population, which is taken

every ten years, is another good example of primary data because it is collected specifically to calculate facts and figures in relation to the people living in the UK**.**

(b) Secondary Data

Secondary data is data which has been collected for some purpose **other** than that for which it is being used. For example, if a company has to keep records of when employees are sick and you use this information to tabulate the number of days employees had flu in a given month, then this information would be classified as secondary data.

Most of the data used in compiling business statistics is secondary data because the source is the accounting, costing, sales and other records compiled by companies for administration purposes. Secondary data must be used with **great care**; as the data was collected for another purpose, you must make sure that it provides the information that you require. To do this you must look at the sources of the information, find out how it was collected and the exact definition and method of compilation of any tables produced.

B. PRELIMINARY CONSIDERATIONS

It is extremely important that, before you start collecting data, you consider the following points.

Aim

Many people fail to think clearly about the problem that is being investigated. You must write down the objective of your survey, stating exactly what information you want to get out of the data that you are planning to collect. Then you will collect only relevant information.

Units

It is essential that you decide what units to use before you start collecting information. Your choice of units should be influenced by the possible need to compare sets of data collected from different sources. Frequently, as you will see later, the data will be collected by a number of people all using different units. Some conversion factors would therefore be needed.

Accuracy

If the level of accuracy is not defined beforehand then you will not know the amount of detail to be collected. For example, if you wish to compare the rainfall in various towns you may find that some records are given to the nearest inch whilst others are correct to three decimal places. Also, if the level of accuracy is stated beforehand, it will be easier to estimate the cost of the data collection.

Methods of Collection

The method of collecting the data must be decided. It will usually be one of the following methods:

- Use of published statistics
- Interviews
- Postal questionnaires
- Personal observations

We will be discussing these methods in the next sections.

C. USE OF PUBLISHED STATISTICS

You must begin any investigation by consulting published sources to see if all or part of the information you require is already available. (Sources of internal and external data will be discussed later in the study unit.) This step is best taken at the planning stage as soon as you have defined the information that you require. You may find that similar information has been collected before, and so you may have to modify your plans in order to avoid duplication of effort.

The information you require may not be found in one source but parts may appear in several different sources. Although this search may be time-consuming it can lead to data being obtained relatively **cheaply** and this is one of the advantages of this type of data collection. Of course the disadvantage is that you could spend a considerable amount of time looking for information which may not be available.

Another disadvantage of using data from published sources is that the definition used for variables and units may not be the same as those you wish to use. It is sometimes difficult to establish the definitions from published information but, before using the data, you **must** establish exactly what it represents.

D. INTERVIEWS

Interviewing is the most common of all the methods of collecting information. It involves employing specially **trained** interviewers to question people on the subject of the survey. This type of interviewing technique is often called **face to face**. Suitable people are chosen as interviewers and then trained in the necessary interviewing techniques. As part of their training they will be shown how to use a **questionnaire**. Some form of questionnaire is always used to obtain the information from the person being interviewed. The design and content of the questionnaire is very important and has a direct effect on the value of the data collected.

Questionnaires

A questionnaire is a list of questions used in a survey or census. Questionnaires may be used by an interviewer or be free-standing, as when they are sent to the respondent through the post.

The design of a questionnaire will reflect the way in which it is to be used. Many problems can be avoided by careful design, particularly where the information on the questionnaire has to be transferred to analysis sheets or entered into a computer. For example, if all the responses are aligned down one side of the sheet it is a great deal easier to read them off than if they are scattered around the sheet.

A particular problem is the attitude of most people to forms which are too obviously designed with transfer of information to a computer in mind. For example, questionnaires sometimes have little squares into which we are asked to fill our names and addresses. Few people like this as they look too mechanical and offend our wishes to be individual! Although they are often necessary, they should be made as unobtrusive as possible. If answers are to be entered by hand, then the space given must be adequate – a line spacing of at least $\frac{1}{3}$rd of an inch is best. Most typewriters and computer printers have a standard line spacing of $\frac{1}{6}$th inch – so avoid ¼ inch spacing as it will be difficult to align typewritten entries.

Overall a questionnaire form should not look too overpowering; good layout can improve response considerably. Equally, questionnaires should be kept as short as possible, unless there is a legal

compulsion to fill it in; as with many government surveys, a several-page questionnaire will probably be put on one side and either forgotten or returned late.

The above discussion only touches on a few of the considerations in designing a questionnaire; hopefully it will make you think about what is involved. Professional help is a good idea when designing a questionnaire.

(a) The Questions

The general principle to keep in mind when designing a set of questions is that if a question can be misread, it will be. Questions must always be tested on someone who was not involved in setting them, and preferably on a small sample of the people they will be sent to. The troubles that can arise were well illustrated by the difficulty experienced in setting a question to establish ethnic origin in the population census of 1981. A trial showed that:

- The question did not give sufficient responses, i.e. people from the Indian sub-continent wished to record their religion as a part of their ethnic description.
- It was not answered in the way the designers expected, i.e. in families of West Indian origin the parents often entered themselves correctly as "West Indian" but not their children if they were born in the UK; they regarded the nationality "British" as more important.
- It was not acceptable – some leaders of ethnic groups regard the question as likely to lead to disadvantages for their groups.

The last opposition was so strong and seemed likely to lead to such problems with the whole questionnaire that it was eventually dropped. A question was designed for the 1991 census and successfully used, the problems described above being avoided by careful consultation and explanation of the purpose of the question before the census was taken. The question was far less detailed than the first attempt in 1981:

Ethnic group	White	0
	Black-Caribbean	1
	Black-African	2
	Black-Other *please describe*	
	Indian	3
	Pakistani	4
	Bangladeshi	5
	Chinese	6
	Any other ethnic group *please describe*	

For cases of mixed ancestry, respondents were asked either to choose the ethnic group they regarded themselves as in, or to tick the "Any other" response and describe their ancestry.

(b) Design Principles

The principles to observe when designing a questionnaire are:

(i) Keep it as short as possible, consistent with getting the right results.

(ii) Explain the purpose of the investigation so as to encourage people to give answers.

(iii) Individual questions should be as short and simple as possible.

(iv) If possible, only short and definite answers like "Yes", "No" or a number of some sort should be called for.

(v) Questions should be capable of only one interpretation, and leading questions should be avoided.

(vi) Where possible, use the "alternative answer" system in which the respondent has to choose between several specified answers.

(vii) The questions should be asked in a logical sequence.

(viii) The respondent should be assured that the answers will be treated confidentially and not be used to his detriment.

(ix) No calculations should be required of the respondent.

You should always apply the above principles when designing a questionnaire, and you should understand them well enough to be able to remember them all if you are asked for them in an examination question. They are **principles** and not rigid rules – often you have to break some of them in order to get the right information. Governments can often break these principles because they can make the completion of the questionnaire compulsory by law, but other investigators must follow the rules as far as practicable in order to make the questionnaire as easy and simple to complete as possible – otherwise they will receive no replies.

When a survey has been planned and a suitable questionnaire has been designed, the task of collecting the information is entrusted to a team of interviewers (unless postal questionnaires are to be used – see later). These interviewers have been trained in the use of questionnaire and advised how to present it so that **maximum** co-operation is obtained from the respondent. This training is very important and must be carefully thought out. The interviewers must be carefully selected so that they will be suitable for the type of interview envisaged. The type of interviewer and the method of approach must be varied according to the type of respondent selected, e.g. the same technique should not be used for interviewing housewives and bank managers.

Here is an example of a simple questionnaire:

1. Please tick your sex.	Male	☐
	Female	☐
2. Which age bracket do you fall in?	Under 25 yrs	☐
	25 yrs – under 45 yrs	☐
	45 yrs – under 65 yrs	☐
	Over 65 yrs	☐
3. Which subjects do you enjoy studying most? *You may tick more than one box.*	Maths	☐
	Languages	☐
	Arts	☐
	Sciences	☐
	Don't enjoy studying	☐
4. Which style of education do you prefer?	Full-time	☐
	Part-time/Day release	☐
	Evening classes	☐
	Correspondence courses	☐
	Self-tuition	☐
	Other	☐
	No preference	☐
5. How do you feel at this stage of the course?	Very confident	☐
	Confident	☐
	Not sure	☐
	Unconfident	☐
	Very unconfident	☐

Your assistance in this matter will help our researchers a great deal. Thank you for your co-operation.

Methods of Interviewing

There are two main methods of interviewing:

(a) The form is left for the respondent to complete at leisure. In this approach the questionnaire is collected at a second visit. The interviewer will be prepared to help the respondent to complete the form at either or both visits.

(b) The questionnaires are completed by the interviewer on the spot. This is the **face-to-face** interview, and is the most common. The interviewer talks directly to the respondent and records the answers to the questions on the form.

There are several variations of these techniques which can be used for special investigations.

The respondents for the interviews are pre-selected and listed for the interviewers. The various methods used in this selection process are described in a later study unit.

Advantages of Interviewing

There are many advantages of using interviewers in order to collect information:

(a) The major one is that a large amount of data can be collected relatively **quickly and cheaply**. If you have selected the respondents properly and trained the interviewers thoroughly, then there should be few problems with the collection of the data.

(b) This method has the added advantage of being very **versatile** since a good interviewer can adapt the interview to the needs of the respondent. If, for example, an aggressive person is being interviewed, then the interviewer can adopt a conciliatory attitude to the respondent; if the respondent is nervous or hesitant, the interviewer can be encouraging and persuasive.

The interviewer is also in a position to explain any question, although the amount of explanation should be defined during training. Similarly, if the answers given to the question are not clear, then the interviewer can ask the respondent to elaborate on them. When this is necessary the interviewer must be **very careful** not to lead the respondent into altering rather than clarifying the original answers. The technique for dealing with this problem must be tackled at the training stage.

(c) This face-to-face technique will usually produce a high response rate. The response rate is determined by the proportion of interviews that are successful. A successful interview is one which produces a questionnaire with every question answered clearly. If most respondents interviewed have answered the questions in this way, then a high response rate has been achieved. A low response rate is when a large number of questionnaires are incomplete or contain useless answers.

(d) Another advantage of this method of collecting data is that with a well-designed questionnaire it is possible to ask a large number of short questions in one interview. This naturally means that the cost per question is lower than in any other method.

Disadvantages of Interviewing

Probably the biggest disadvantage of this method of collecting data is that the use of a large number of interviewers leads to a **loss of direct control** by the planners of the survey. Mistakes in selecting interviewers and any inadequacy of the training program may not be recognised until the interpretative stage of the survey is reached. This highlights the need to train interviewers correctly.

It is particularly important to ensure that all interviewers ask questions in a similar way. It is possible that an inexperienced interviewer, just by changing the tone of voice used, may give a different emphasis to a question than was originally intended. This problem will sometimes become evident if unusual results occur when the information collected is interpreted.

In spite of these difficulties, this method of data collection is widely used as questions can be answered cheaply and quickly and, given the correct approach, this technique can achieve high response rates.

E. POSTAL QUESTIONNAIRES

In this method of data collection the postal service is generally used to distribute the questionnaire to the selected respondents, who can be single persons, a household, a firm or a football team. The points made about questionnaires in the previous section apply equally well when they are sent by post. However, two other guidelines must be considered:

(a) Size

This is extremely important for two main reasons. Firstly, when posting anything, you must remember the physical limitations of post and letter boxes. Secondly, the size of the document presented to the respondent will affect the response rate. If the respondents are presented with a large and bulky questionnaire, they are less likely to answer it than if it is small.

(b) Presentation

The way in which the questionnaire is presented is vital for a good response rate. The purpose of the questionnaire, as it is not presented by an interviewer, must be contained in a clear and concise way either as a covering letter or as a note at the **top** of the questionnaire.

Advantages of Postal Questionnaires

This technique has a number of advantages, the major one being its **cheapness**. As there are no interviewers, the only direct cost is that of the postage. This means that the questionnaires can be distributed to a **wider range** of respondents at a cheaper rate, and this may increase the response rate.

This type of data collection allows the respondents **plenty of time** to consider their answers. Compare this with the interviewing technique where the interviewer requires an immediate response.

The final advantage is the **elimination of interviewer bias**, as even some of the best-trained interviewers will tend to put their own slant on any interview. In some cases, if the interviewer is biased or inadequately trained, this can lead to serious distortion of the results.

Disadvantages of Postal Questionnaires

The major disadvantage of this method of data collection is the inability of the planners to control the number of responses: some respondents will not bother to reply, and others will feel that they are not qualified to reply. For example, if questionnaires about fast motor cars were sent to a cross-section of the population, then only those people who owned a fast motor car might return the questionnaire. People without fast cars might think the questionnaire did not apply to them and consequently would not send it back. Therefore, as the percentage of people returning the questionnaire is very low, the **response rate is low**.

This situation can be improved either by sending out a very large number of questionnaires, so that even though the actual response rate is low, the number responding is high enough for the purpose of the survey; or by offering some form of incentive, such as a lottery prize, for the return of the form. Both of these methods would involve an increase in cost which would counteract the greatest advantage of this method, that of cheapness.

The problem introduced by the first method above is that even though the number of responses is sufficient, they do not represent the views of a typical cross-section of the number of people first approached. For example, very few replies would be received from those not owning fast motor cars, so that any deductions drawn from the data would be biased. So, you can see that you have very little control over the response rate with this method of collection. As there are no interviewers, you have to rely on the **quality** of the questionnaire to encourage the respondents to co-operate.

This means that great care has to be taken with the design of the questionnaire. In particular it is extremely important that the wording of the questions is very simple, and any question that could be interpreted in more than one way should be left out. The required answers should be a simple yes/no or at the most a figure or date. You should not ask questions that require answers expressing an attitude or opinion, while using this particular technique.

Finally, it is important to remember that this type of data collection **takes much longer** to complete than the other methods described. Experience shows that about 15% of the questionnaires sent out will be returned within a week, but the next 10% (bringing the response rate up to a typical 25%), may take anything up to a month before they come back.

F. PERSONAL OBSERVATION

This method is used when it is possible to **observe directly** the information that you wish to collect. For example, data for traffic surveys is collected in this way: observers stand by the roadside and count and classify the vehicles passing in a given time. Increasingly, computers are replacing human observers in this method of data collection as they are considerably cheaper and often more reliable. There are numerous examples of this, and most traffic information is now collected by rubber tubes laid across the road and linked to a small computer placed alongside the road.

The main advantage of this method of data collection is that the data is observed directly instead of being obtained from other sources. However, when observers are used, you must allow for human error and personal bias in the results. Although this type of bias or error is easy to define, it is sometimes extremely difficult to recognise and even harder to measure the degree to which it affects the results. Personal bias can be more of a problem when only part of the data available is being observed.

This problem will be covered in greater detail in a later study unit which deals with sampling.

Provided proper and accurate instructions are given to the observers in their training, this type of bias can be reduced to a minimum.

Personal observation means that the data collected is usually limited by the resources available. It can often be expensive, especially where large amounts of data are required. For these reasons this method is not widely used. However, the increasing use of the computer is reducing both the amount of bias and the cost.

G. CHOICE OF METHOD

The type of information required will often determine the method of collection. If the data is easily obtained by automatic methods or can be observed by the human eye without a great deal of trouble, then the choice is easy. The problem comes when it is necessary to obtain information by questioning respondents. The best guide is to ask whether the information you want requires an attitude or opinion or whether it can be acquired from short yes/no type or similar simple answers. If it is the former, then it is best to use an interviewer to get the information; if the latter type of data is required, then a postal questionnaire would be more useful.

Do not forget to check published sources first to see if the information can be found from data collected for another survey.

Another yardstick worth using is time. If data must be collected quickly, then use an interviewer and a short simple questionnaire. However, if time is less important than cost, then a postal questionnaire, since this method may take a long time to collect relatively limited data but is cheap.

Sometimes a question in the examination paper is devoted to this subject. The tendency is for the question to state the type of information required and ask you to describe the appropriate method of data collection, giving reasons for your choice. More commonly, specific definitions and explanations of various terms such as interviewer bias are contained in multi-part questions.

H. INTERNAL AND EXTERNAL SOURCES OF DATA

We have emphasised the need for you to consult published sources before deciding to go out and collect your own data. We will now describe where to look for business data. You will often find useful information from several sources, both within an organisation and outside.

Scanning Published Data

When you examine published data from whatever source, it is helpful to adopt the following procedure:

(a) Overview the Whole Publication

Flip through the pages so that you get a feel for the document. See if it contains tables only, or if it uses graphs and tables to describe the various statistics.

(b) Look at the Contents Pages

A study of the contents pages will show you in detail exactly what the document contains. This will give you a good idea of the amount of detail contained in the document. It will also show you which variables are described in the tables and charts.

(c) Read the Introduction

This will give a general indication of the origin of the statistics shown in the document. It may also describe how the survey which collected the information was carried out.

(d) Look at Part of the Document in Detail

Take a small section and study unit that in depth. This will give you an appreciation of just what information is contained and in what format. It will also get you used to studying documents and make you appreciate that most tables, graphs or diagrams include some form of notes to help explain the data.

Internal Data Sources

All types of organisation will collect and keep data which is therefore internal to the organisation. More often than not it applies to the organisation where you work, but you should not think of it meaning just this type of organisation. It is important, when looking for some particular type of data, to look internally because:

- It will be **cheaper** if the data can be obtained from an internal source as it will save the expense of some form of survey.
- Readily available information can be used much more **quickly** especially if it has been computerised and can be easily accessed
- When the information is available from within your organisation, it can be **understood** much more easily as documentation is likely to be readily available.

Overall there are several advantages from using internal data, although there is a tendency when using this type of data to make do with something that is **nearly** right.

Published or External Sources

The sources of statistical information can be conveniently classified as:

- Central and local government sources together with EU publications

- Private sources

The data produced by these sources can be distinguished as:

- Data collected **specifically** for statistical purposes – e.g. the population census.
- Data arising as a by-product of other functions – e.g. unemployment figures.

This latter distinction is well worth noting because it sometimes helps to indicate the degree of reliability of the data. Do not forget, of course, that very often the statistician has to be his own source of information; then he must use the techniques of data collection which we have already discussed.

The main producer of statistics in this country is central government, and for this purpose an organisation has been set up called the Government Statistical Service (GSS). The GSS exists primarily to service the needs of central government. However, much of the information it produces is eminently suitable for use by the business community as well, and indeed central government is increasingly becoming aware of the need to gear its publications so that they can be used by the business sector.

Local government also produces a wealth of information, but because of its localised nature it is not often found on the shelves of all libraries or made available outside the area to which it applies. One source which is increasingly becoming available is documents produced by the European Union (EU). Similarly, the United Nations publications are available, which cover world-wide statistics in subjects such as population and trade.

Government Publications

The principal statistics provided by the government can be found in various publications. They include the weekly British Business (formally Trade and Industry, and before that the Board of Trade Journal) and the monthly publication Employment Gazette (formerly the Department of Employment and Productivity Gazette and earlier the Ministry of Labour Gazette). Statistics found in these two journals are also included in various other publications such as the Monthly Digest of Statistics, Financial Statistics (monthly) and Economic Trends (monthly). Annual publications include the Annual Abstract of Statistics, National Income and Expenditure (the Blue Book) and Regional Trends.

A summary of the major publications and their original sources follows.

(a) General

Publication	*Description*	*Source*
Annual Abstract of Statistics	Main economic and social statistics for the UK	Central Statistical Office (CSO)
Monthly Digest of Statistics	Main economic and social statistics for the UK	CSO
Regional Trends (annual)	Main economic and social statistics for regions of the UK	CSO
Scottish Abstract of Statistics (annual)	Main Scottish statistics	Scottish Office
Digest of Welsh Statistics (annual)	Main Welsh statistics	Welsh Office

(b) National Income and Expenditure

Publication	*Description*	*Source*
UK National Accounts (Blue Book) (annual)	National account statistics	CSO
Family Expenditure Survey Reports (annual)		Dept. for Education and Employment (DfEE)
Employment Gazette	Employment, labour, retail prices and wage statistics	DfEE
Economic Trends (monthly)	Primary statistics on the current economic situation	CSO
Inland Revenue Statistics (annual)	Taxation, incomes, capital and valuation statistics	Inland Revenue
Report of the Commissioners of HM Customs and Excise (annual)	Customs and excise duties collected	HM Customs and Excise
Household Food Consumption and Expenditure (annual)		Ministry of Agriculture, Fisheries and Food (MAFF)

(c) Business Monitor Series

This series consists of over 100 titles and is prepared by the Department of Trade and Industry. Detailed statistical information on a wide range of economic activities is given. Some of the publications are monthly, others quarterly or annual.

- The **Production** series consists of more than 100 publications, mostly quarterly, and covers individual industries under the following general group headings: mining, food, drink and tobacco; chemicals and allied industries; mechanical engineering; shipbuilding and marine engineering; vehicles; metal goods; textiles; leather goods and fur; clothing and footwear; bricks, pottery, glass, cement, etc.; timber, furniture, etc.; paper, printing and publishing; other industries.
- The **Distributive and Services** series contains the following groups, predominantly monthly publications: food shops; clothing and footwear shops; durable goods shops; miscellaneous non-food shops; catering trades; instalment credit business of financial houses; instalment credit business of retailers.
- The **Miscellaneous** series covers motor vehicle registrations; cinemas; company finance; overseas transactions; insurance companies' and private pension funds' investment; overseas travel and tourism; acquisitions and mergers of companies; nationality of vessels in seaborne trade.

(d) Trade

Publication	*Description*	*Source*
British Business (weekly)	Production, prices, trade, industrial materials and commodities	Department of Trade and Industry
Annual Statement of the Overseas Trade of the UK	Imports and exports analysed by commodity and country; trade at ports	HM Customs and Excise
Overseas Trade Statistics of the UK (annual)	UK import and export by commodity	HM Customs and Excise
UK Balance of Payments (Pink Book) (annual)	Balance of payments over past years	CSO

(e) Other

Publication	*Description*	*Source*
Financial Statistics (monthly)	UK monetary and financial statistics	CSO
Monthly Bulletin of Construction Statistics	Statistics on building and civil engineering, local authority design work and building materials	Department of the Environment, Transport and the Regions (DETR)
Housing and Construction Statistics (quarterly)		DETR, Scottish Office, Welsh Office
Health and Personal Social Services Statistics	Statistics for health and related welfare services	Department of Health and Social Security
Agricultural Statistics: UK		MAFF
Social Trends	Social conditions statistics	CSO
Energy Trends	Statistics of energy, fuel and power	Department of Trade and Industry
CAA Monthly Statistics	All aviation activities	Civil Aviation Authority
Transport Statistics (annual)	Statistics on vehicles, traffic and road transport	DETR
Passenger Transport in Great Britain (annual)		DETR
Population Monitors	Demographic statistics	Office of Population Censuses and Surveys
Statistical News (quarterly)	Articles and notes on all new developments in official statistics	CSO

Census of Production

A census of production is the collection of information about the productive activity of a country. In order to understand the interest of governments in production statistics, it is only necessary to remember that production is the key to national prosperity.

The census of production covers production in its narrowest sense, and relates to the mining, quarrying, building, manufacturing, and gas, electricity and water-supplying industries, including the activities of public and local authorities where they fall within those headings. The census does not include agriculture, commerce of transport.

The census is conducted by sending an enquiry form to all firms engaged in productive activity, except those employing less than ten persons. The information required relates to a number of areas, primarily:

- Details about employees, wages etc.
- Sales and work produced
- Production costs (i.e. raw materials, transport costs, stocks etc.)

Study Unit 2

Sampling Procedures

Contents *Page*

A. SAMPLES

A considerable portion of modern statistical theory revolves around the use of samples, and many of the practical applications of this theory are possible only if samples are collected.

Illustrative Example

Preceding a general election, the public is told by the media, quoting the results of opinion polls, that the various political parties enjoy the support of certain percentages of the electorate. These results cannot be obtained by asking every voter in the country for his or her political views, as this is clearly impracticable because of cost and time. Instead, some of the voters are asked for their views, and these, after a certain amount of statistical analysis, are published as the probable views of the whole electorate. In other words, the results of a survey of a minority have been **extended** to apply to the majority.

This example illustrates the principles of sampling, and we must now define some of the terms involved in sampling.

Definitions

- **Population**

  A population is the set of all the individuals or objects which have a given characteristic, e.g. the set of all persons eligible to vote in a given country.

- **Sample**

  A sample is a sub-set of a population, e.g. the voters selected for questioning about their views.

- **Sampling**

  Sampling is the process of taking a sample.

- **Sample Survey**

  The process of collecting the data from a sample is called a sample survey, e.g. asking the selected voters their political views is a sample survey.

- **Census**

  The process of collecting data from a **whole population** is called a census, e.g. a population census in which data about the entire population of a country is collected. (Note that the ten-yearly population census taken in the UK is one of the few questionnaires that the head of a household is **compelled by law** to complete.)

Reasons for Sampling

The advantages of using a sample rather than the whole population are varied:

(a) Cost

Surveying a sample will cost much less than surveying a whole population. Remember that the size of the sample will affect the accuracy with which its results represent the population from which it has been drawn. So, you must balance the size of the sample against the level of accuracy you require. This level of accuracy must be determined before you start the survey (the larger the sample, the greater the reliance that you can put on the result).

(b) Control

A sample survey is easier to control than a complete census. This greater control will lead to a higher response rate because it will be possible to interview every member of the sample under similar conditions. A comparatively small number of interviewers will be needed, so standardisation of the interviews will be easier.

(c) Speed

Apart from the lower cost involved in the use of a sample, the time taken to collect the data is much shorter. Indeed, when a census is taken, a sample of the data is often analysed at an early stage in order to get a general indication of the results likely to arise when the census information is fully analysed.

(d) Quality

When only a few interviews are needed, it is easier to devote a greater degree of effort and control per interview than with a larger number of interviews. This will lead to better-quality interviews and to a greater proportion of the questions being answered correctly without the necessity of a call-back. (A call-back is when an interviewer has to return to the respondent, if that is possible, in order to clarify the answer to a question.)

(e) Accuracy

The level of accuracy of a survey is assessed from the size of the sample taken. Since the quality of the data obtained from a sample is likely to be good, you can have confidence in this assessment.

(f) Industrial Application

Sampling is not confined to surveys such as opinion polls which involve interviews and postal questionnaires; it is also important in controlling industrial production. On an assembly line in a factory producing manufactured goods, it is necessary to check the standard of the product continuously. This is carried out by selecting a sample of the same product every day and testing that each item in the sample meets the manufacturer's specifications.

Sometimes this testing involves destroying the product. For example, in a tyre factory each tyre will be required to have a minimum safe-life in terms of distance driven and to withstand a minimum pressure without a blow-out. Obviously the whole population of tyres cannot be tested for these qualities. Even when the testing involves nothing more than measuring the length of a bolt or the pitch of a screw, a sample is used because of the saving in time and expense.

B. STATISTICAL INFERENCE

Among the reasons for taking a sample is that the data collected from a sample can be used to infer information about the population from which the sample is taken. This process is known as **statistical inference**. The theory of sampling makes it possible not only to draw statistical inferences and conclusions from sample data, but also to make precise probability statements about the reliability of such inferences and conclusions. Future study units will enlarge on this subject.

Before we continue we must define some terms which are generally used in statistical inference:

- **parameter –** a constant measure used to describe a characteristic of a population.
- **statistic –** a measure calculated from the data set of a sample.

- **estimate –** the value of a statistic which, according to sampling theory, is considered to be close to the value of the corresponding parameter.
- **sampling unit –** an item from which information is obtained. It may be a person, an organisation or an inanimate object such as a tyre.
- **sampling frame –** a list of all the items in a population.

The sampling theory which is necessary in order to make statistical inferences is based on the mathematical theory of probability. We will discuss probability later in the course.

C. SAMPLING

Once you have decided to carry out a sample survey, there are various decisions which must be made before you can start collecting the information. They are:

- Procedure for selecting the sample
- Size of the sample
- Elimination of bias
- Method of taking the sample

We will discuss these in some detail.

Procedure for Selecting the Sample

In selecting a sample you must first define the sampling frame from which the sample is to be drawn. Let's consider a particular survey and discuss how the stages, defined above, may be carried out.

Example

Suppose you are the chairman of Bank A, which is in competition with Banks B, C and D, and you want to find out what people think of your bank compared with the other three banks. It is clearly a case for a sample survey, as cost alone would prohibit you from approaching everyone in the country to find out their views. The information required for the survey would involve questions of **opinion**, so an **interviewing** technique is the best method to use.

If you want a cross-section of views throughout the country, then the sampling frame could be all the adults in the country. However, if you are interested only in existing customers' views, then the sampling frame would be all the customers of Bank A. In this case a list of all the customers at the various branches can be obtained. In the former case a list of all the adults in the country can be found in the electoral roll, which is a record of all those people eligible to vote.

You must be careful to make sure that the sampling frame represents the population exactly as, if it does not, the sample drawn will not represent a true cross-section of the population. For example if the electoral roll is used as the sampling frame but the population you want is all present and prospective customers, then customers under the age of 18 would not be represented, since only those persons of voting age, 18 and over, are included in the electoral roll. So, if you decide that the population should include persons old enough to have bank accounts but under 18, the sampling frame must include school rolls (say) as well. Thus you can see that there are often several sampling frames available, and you have to take great care in matching the sample frame with the scope of the survey. You have to decide whether the effort and cost involved in extending the sampling frame justifies the benefits gained.

Sampling Size

Having chosen the sampling frame, you now have to decide on the size of the sample, and this is a very complex problem. The cost of a survey is directly proportional to the sample size, so you need to keep the sample as small as possible. However, the level of accuracy (and hence the degree of confidence that you can place on your deductions) also depends on the sample size and is improved as this size increases. You have to strike a delicate balance between these conflicting requirements.

In addition, the method of analysis depends, to some extent, on the sample size. The relationship between the size of the sample and the size of the population from which it is taken does **not** affect the accuracy of the deductions. This problem will be discussed again later, but the theory on which the decision is based is outside the scope of this course. You only need to be aware of the problem and to know the formulae (given later) used to calculate the degree of confidence associated with deductions.

Elimination of Bias

Three common sources of bias are:

(a) Inadequacy of Sampling Frame

The sampling frame chosen may not cover the whole population, so that some items will not be represented at all in the sample and some will be over-represented or duplicated. This bias can be avoided by a careful statement of the aim of the survey and a check that none of the sampling units has been ignored.

For example, if a survey of unemployment is undertaken by randomly speaking to people in South-East England, a biased result will be obtained. This is because the survey population does not contain people in the rest of England. Thus, although the selection process may have been fair and totally random, it will be very biased and non-representative of the whole of England.

(b) Items of Selected Sample Not All Available

It is possible that, when a sample has been selected, some of the items chosen cannot be located, e.g. some voters on the electoral roll may not have notified a change of address. If the missing items are not replaced or are incorrectly replaced, a bias will be introduced. This bias can be reduced to a minimum by returning to the sampling frame and using the same method to select the replacements as was used to select the original sample.

For example, a survey on sickness at a large industrial company could be done by randomly drawing a sample of 500 personal files. However, having randomly selected 500 employees it may transpire that some personal files are missing (they may be in transit from other departments). This could be easily rectified by returning to the frame and randomly selecting some replacements.

Care must obviously be taken to ensure that the reason why the files are missing is not related to the survey – e.g. if they are out for updating because the person has just resumed work after yet another period of sickness!

(c) Interviewer or Observer Bias

This is often the commonest of all types of bias. All interviewers and observers are given a list of their sampling units. Sometimes to save time and effort they may substitute missing units on the spot without considering how the other units have been chosen. Other sources of bias arise when the interviewers do not follow the questionnaires exactly, allow their own ideas to

become evident, or are careless in recording the responses; observers may measure or record their results inaccurately.

This type of bias is difficult to recognise and correct. It can be reduced by careful choice and training of the team, and by close supervision when the survey is taking place. For example, during a high street survey an interviewer is eager to speed up responses. In order to do so she prompts people who hesitate with replies. Although a question reads, "What type of mineral water do you prefer?", she goes on to add, "Most people have said 'Lemonade', which seems quite sensible". This would inevitably lead the respondent either to agree or appear non-sensible.

Bias can rarely be eliminated completely, but the results of the survey may still be useful provided that the final report states any assumptions made, even if they are not fully justified, e.g. if the sampling frame is not complete.

Method of Taking the Sample

The final decision you have to make is about the method to use to select the sample. The choice will depend on the aim of the survey, the type of population involved, and the time and funds at your disposal. The methods from which the choice is usually made are:

- Simple random sampling
- Systematic sampling
- Stratified sampling
- Multi-stage sampling
- Cluster sampling
- Quota sampling

In the next section we will define, explain and discuss the major advantages and disadvantages of these methods.

D. SAMPLING METHODS

Simple Random Sampling

The word **random** has a definite and specific meaning in the statistical theory of sampling. The dictionary definition of random is "haphazard" or "without aim or purpose", but the statistical definition is **a process by which every available item has an equal chance of being chosen.**

For example, looking at the bank survey again and given that the sampling frame is everybody over 18 shown on any electoral roll throughout the UK, everyone on the roll is given a unique number from 1 to n, (n being the total number of people in the sampling frame). Each number is now written on a slip of paper and put in a box. If you want a sample of a thousand people you mix up these slips thoroughly and draw out a thousand slips. The numbers on these slips then represent the people to be interviewed. In theory each slip would stand an equal chance of being drawn out and so would have been chosen in a **random** manner. It is fundamental to simple random sampling that every element of the sampling frame stands an **equal** chance of being included in the sample.

This method sounds almost foolproof but there are some practical difficulties: if, for instance, there are 52 million people in the sampling frame, another method of drawing a sample in a random fashion has been devised – using a computer, for example.

The most convenient method for drawing a sample for a survey is to use a table of random numbers. Such a table is included in your copy of Mathematical Tables for Students. These tables are compiled with the use of a computer, so that each of the digits from 0 to 9 stands an equal chance of appearing in any position in the table. If a sample of a thousand is required, for example, then the first thousand numbers falling within the range 1 to n that are found in the table, **form the sample** (where n is the total number in the sampling frame). Many pocket calculators have a built-in program for selecting random numbers.

- ***Advantages***

  The advantage of this method of selection is that it always produces an **unbiased sample**.

- ***Disadvantages***

  Its disadvantage is that the sampling units may be difficult or expensive to contact, e.g. in the bank survey sampling units could be drawn in any area from John O'Groats to Lands End.

Systematic Sampling

This method is sometimes called **quasi-random sampling** and involves the selection of a certain **proportion of the total population**. Drawing a simple random sample as described above can be very time-consuming. The systematic sampling method simplifies the process.

First you decide the size of the sample and then divide it into the population to calculate the proportion of the population you require. For example, in the bank survey you may have decided that a tenth of the population would provide an adequate sample. Then it would be necessary to select every tenth person from the sampling frame. As before, each member of the population will be given a number from 1 to n, the starting number is selected from a table of random numbers by taking the first number in the table between 1 and 9. Say a 2 was chosen, then the 2nd, 12th, 22nd, 32nd . . . . etc. person would be selected from the sampling frame. This method of sampling is often used as it reduces the amount of time that the sample takes to draw. However, it is not a purely random method of selecting a sample, since once the starting point has been determined, then the items selected for the sample have also been set.

- ***Advantages***

  The main advantage of this method is the **speed** with which it can be selected. Also it is sufficiently close to simple random sampling, in most cases, to justify its widespread use.

- ***Disadvantages***

  It is important to **check**. A major disadvantage occurs if the sampling frame is arranged so that sampling units with a particular characteristic occur at regular intervals, causing over- or under-representation of this characteristic in the sample. For example, if you are choosing every tenth house in a street and the first randomly chosen number is 8, the sample consists of nos. 8, 18, 28, 38 and so on. These are all even numbers and therefore are likely to be on the same side of the street. It is possible that the houses on this side may be better, more expensive houses than those on the other side. This would probably mean that the sample was biased towards those households with a high income. A sample chosen by systematic sampling must always be examined for this type of bias.

Stratified Sampling

Before we discuss this method of sampling, we have to define two different types of population:

- **Homogenous population**: sampling units are all of the same kind and can reasonably be dealt with in one group.
- **Heterogeneous population**: sampling units are different from one another and should be placed in several separate groups.

In the sampling methods already discussed we have assumed that the populations are homogeneous, so that the items chosen in the sample are typical of the whole population. However, in business and social surveys the populations concerned are very often heterogeneous. For example, in the bank survey the bank customers may have interests in different areas of banking activities, or in a social survey the members of the population may come from different social classes and so will hold different opinions on many subjects. If this feature of the population is ignored, the sample chosen will not give a true cross-section of the population.

This problem is overcome by using **stratified sampling**. The population is divided into groups or **strata according to the characteristics** of the different sections of the population, and a simple random sample is taken from each stratum. The sum of these samples is equal to the size of the sample required, and the individual sizes are proportional to the sizes of the strata in the population. An example of this would be the division of the population of London into various social-economic strata.

- ***Advantages***

  The advantage of this method is that the results from such a sample will not be distorted or biased by undue emphasis on extreme observations.

- ***Disadvantages***

  The main disadvantage is the difficulty of defining the strata. This method can also be time-consuming, expensive and complicated to analyse.

Multi-stage Sampling

This method consists of a number of stages and is designed to retain the advantage of simple random sampling and at the same time cut down the cost of the sample. The method is best explained by taking the bank survey already discussed as an example and working through the various stages.

Suppose you have decided that you need a sample of 5000 adults selected from all the adults in the UK but that the expense of running the survey with a simple random sample is too high, then you could proceed as follows:

Stage 1: Use all the administrative counties of the UK as the sampling units and select a simple random sample of size 5 from this sampling frame.

Stage 2: Each county will be divided into local authority areas, so use these as the sampling units for this stage and select a simple random sample of size 10 from each of the 5 counties chosen in stage 1. You now have 50 local authority areas altogether.

Stage 3: Divide each of the selected local authority areas into postal districts and select one of these districts randomly from each area. So you now have 50 randomly selected small regions scattered throughout the country.

Stage 4: Use the electoral rolls or any other appropriate list of all the adults in these districts as the sampling frame and select a simple random sample of 100 adults from each district.

If you check back over the stages you will find that you have a multi-stage sample of total size 5000 which is divided equally between 50 centres. The 100 persons at each centre will be easy to locate and can probably be interviewed by one or two interviewers. The subdivisions at each stage can be chosen to fit in conveniently with the particular survey that you are running. For instance, a survey on the health of school children could begin with local education authorities in the first stage and finish with individual schools.

- ***Advantages***

  The advantages of this method are that at each stage the samples selected are small and that interviews are carried out in 50 small areas instead of in 5000 scattered locations, thus economising on time and cost. There is no need to have a sampling frame to cover the whole country. The sample is effectively a simple random sample.

- ***Disadvantages***

  The main disadvantages are the danger of introducing interviewer bias and of obtaining different levels of accuracy from different areas. The interviewers must be well chosen and thoroughly trained if these dangers are to be avoided.

Cluster Sampling

We have already considered the cost and time problems associated with simple random sampling, and cluster sampling is another method of overcoming these problems. It is also a useful means of sampling when there is an **inadequate sampling frame** or when it is too expensive to construct the frame. The method consists of dividing the sampling area into a number of small concentrations or **clusters** of sampling units. Some of these clusters are chosen at random, and every unit in the cluster is sampled.

For example, suppose you decided to carry out the bank survey using the list of all the customers as the sampling frame but wished to avoid the cost of simple random sampling, you could take each branch of the bank as a cluster of customers. Then you select a number of these clusters randomly and interview every customer on the books of the branches chosen. As you interview all the customers at the randomly selected branches, the sum of all interviews forms a sample which is representative of the sampling frame, thus fulfilling your major objective of a random sample of the entire population.

A variation of this method is often used in the United States, because of the vast distances involved in that country (often referred to as **area sampling**). With the use of map references, the entire area to be sampled is broken down into smaller areas, and a number of these areas are selected at random. The sample consists of all the sampling units to be found in these selected areas.

- ***Advantages***

  The major advantages of this method are the reduction in cost and increase of speed in carrying out the survey. The method is especially useful where the size or constitution of the sampling frame is unknown. Nothing needs to be known in advance about the area selected for sampling, as all the units within it are sampled; this is very convenient in countries where electoral registers or similar lists do not exist.

- ***Disadvantages***

  One disadvantage is that often the units within the sample are homogeneous, i.e. clusters tend to consist of people with the same characteristics. For example, a branch of a bank chosen in a wealthy suburb of a town is likely to consist of customers with high incomes. If all bank branches chosen were in similar suburbs, then the sample would consist of people from one

social group and thus the survey results would be biased. This can be overcome to some extent by taking a large number of small clusters rather than a small number of large clusters. Another disadvantage of taking units such as a bank branch for a cluster is that the variation in size of the cluster may be very large, i.e. a very busy branch may distort the results of the survey.

Quota Sampling

In all the methods discussed so far, the result of the sampling process is a list of all those to be interviewed. The interviewers must then contact these sampling units, and this may take a considerable amount of time. It is possible that, in spite of every effort, they may have to record "no contact" on their questionnaire. This may lead to a low response rate and hence the survey result would be biased and a great deal of effort, time and money would have been wasted.

To overcome these problems, the method of quota sampling has been developed, in which a sampling frame and a list of sampling units is not necessary; it is sometimes referred to as a **non-probability sampling method**. The basic difference between this method and those we have already discussed is that the final choice of the sampling units is left to the sampler (interviewer).

The organisers of the survey supply the sampler, usually an interviewer, with the area allocated to him/her and the number and type of sampling units needed. This number, called a **quota**, is usually broken down by social class, age or sex. The interviewers then take to the street and select the units necessary to make up their quota. This sounds simple but in reality selecting the quota can be difficult, especially when it comes to determining certain characteristics like the social class of the chosen person. It requires experience and well-trained interviewers who can establish a good relationship quickly with those people being interviewed.

- ***Advantages***

  The advantages of this method are that it is probably the **cheapest way** of collecting data; there is no need for the interviewers to call back on any respondent; they just replace any respondent with another more convenient to locate; it has been found to be very successful in skilled hands.

- ***Disadvantages***

  The disadvantages are that as the sample is **not random**, statistically speaking, it is difficult to assess a degree of confidence in the deductions; there is too much reliance on the judgement and integrity of the interviewers and too little control by the organisers.

E. PILOT SURVEY

After all the preliminary steps for a survey have been taken, you may feel the need for a trial run before committing your organisation to the expense of a full survey. This trial run is called a **pilot survey** and will be carried out by sampling only a small proportion of the sample which will be used in the final survey. The analysed results of this pilot survey will enable you to pick out the weaknesses in the questionnaire design, the training of the interviewers, the sampling frame and the method of sampling. The expense of a pilot survey is worth incurring if you can correct any planning faults before the full survey begins.

F. CHOICE OF SAMPLING METHOD

The sampling method is probably the factor which has most effect on the quality of survey results so it needs very **careful thought**. You have to balance the advantages and disadvantages of each method for each survey. When you have defined the aim of the survey, you have to consider the type of population involved, the sampling frame available and the area covered by the population.

If you are to avoid bias there should be some element of randomness in the method you choose. You have to recognise the constraints imposed by the level of accuracy required, the time available and the cost.

If you are asked in an examination to justify the choice of a method, you should list its advantages and disadvantages and explain why the advantages outweigh the disadvantages for the particular survey you are required to carry out.

Study Unit 3

Tabulating and Graphing Frequency Distributions

Contents ***Page***

Continued over

A. RAW DATA

Collection of Raw Data

Suppose you were a manager of company and wished to obtain some information about the heights of the company's employees. It may be that the heights of the employees are currently already on record, or it may be necessary to measure all the employees. Whichever the case, how will the information be recorded? If it is already on record, it will presumably be stored in the files of the personnel department, and these files are most likely to be kept in alphabetical order, work-number order, or some order associated with place or type of work. The point is that it certainly will **not** be stored in height order, either ascending or descending.

Form of Raw Data

It is therefore most likely that, when all the data has been collected, it is available for use, but not in such a form as to be instantly usable. This is what usually happens when data is collected; it is noted down as and when it is measured or becomes available. If, for example, you were standing by a petrol pump, noting down how many litres of petrol each motorist who used the pump put into his car, you would record the data in the order in which it occurred, and not, for example, by alphabetical order of cars' registration plates.

Suppose your company has obtained the measurements of 80 of its employees' heights and that they are recorded as follows:

Table 3.1: Heights of Company Employees in cm

173	177	168	173	182	176	179	173
179	163	180	168	188	167	183	187
160	173	174	184	163	188	176	169
175	178	177	162	176	181	188	183
181	170	179	173	170	169	164	176
164	175	164	180	174	165	174	179
183	181	170	177	185	173	171	165
189	181	175	186	166	177	179	169
179	183	182	165	180	171	173	174
172	166	182	186	181	178	178	187

Table 3.1 is simply showing the data in the form in which it was collected; this is known as **raw data**. What does it tell us? The truthful answer must be, not much. A quick glance at the table will confirm that there are no values above 200, and it appears that there are none below 150, but within those limits we do not have much idea about any pattern or spread in the figures. (In fact, all the values are between 160 and 190.) We therefore need to start our analysis by rearranging the data into some sort of order.

B. ORDERED DATA

Arrays

There is a choice between the two obvious orders for numerical data, ascending and descending, and it is customary to put data in **ascending order**. A presentation of data in this form is called an **array**, and is shown in Table 3.2.

It becomes immediately obvious from this table that all the values are between 160 and 190, and also that approximately one half of the observations occur within the **middle third**, between 170 and 180. Thus we have information not only on the **lower and upper limits** of the set of values, but also on their **spread** within those limits.

Table 3.2: Array of Heights of Company Employees in cm

160	166	170	173	176	179	181	184
162	166	171	174	176	179	181	185
163	167	171	174	177	179	181	186
163	168	172	174	177	179	182	186
164	168	173	174	177	179	182	187
164	169	173	175	177	180	182	187
164	169	173	175	178	180	183	188
165	169	173	175	178	180	183	188
165	170	173	176	178	181	183	188
165	170	173	176	179	181	183	189

Ungrouped Frequency Distribution

However, writing out data in this form is a time-consuming task, and so we need to look for some way of presenting it in a more concise form. The name given to the number of times a value occurs is its **frequency**. In our array, some values occur only one, i.e. their frequency is 1, while others occur more than once, and so have a frequency greater than 1.

In an array, we write a value once for every time it occurs. We could therefore shorten the array by writing each value only once, and noting by the side of the value the frequency with which it occurs. This form of presentation is known as an **ungrouped frequency distribution**, because all frequency values are listed and not grouped together in any form. (See Table 3.3.) By frequency distribution we mean the way in which the frequencies or occurrences are distributed throughout the range of values.

Note that there is no need to include in a frequency distribution those values (for example, 161) which have a frequency of zero.

Table 3.3: Ungrouped Frequency Distribution of Heights of Company Employees in cm

Height	Frequency	Height	Frequency	Height	Frequency
160	1	171	2	181	5
162	1	172	1	182	3
163	2	173	7	183	4
164	3	174	4	184	1
165	3	175	3	185	1
166	2	176	4	186	2
167	1	177	4	187	2
168	2	178	3	188	3
169	3	179	6	189	1
170	3	180	3		

Total frequency = 80

Grouped Frequency Distribution

The ungrouped frequency distribution, however, does not enable us to draw any further conclusions about the data, mainly because it is still rather lengthy. What we need is some means of being able to represent the data in summary form. We are able to achieve this by expressing the data as a **grouped frequency distribution**.

In grouped frequency distribution, certain values are grouped together. The groups are usually referred to as **classes**. We will group together all those heights of 160 cm and upwards but less than 165 cm into the first class; from 165 cm and upwards but less than 170 cm into the second; and so on. Adding together the frequencies of all values in each class gives the following grouped frequency distribution – Table 3.4.

(Always total up the frequencies, as it gives you a good check on the grouping you have carried out.)

Table 3.4: Grouped Frequency Distribution of Heights of Company Employees in cm.

Heights *(cm)*	**Frequency**
160 - under 165	7
165 - under 170	11
170 - under 175	17
175 - under 180	20
180 - under 185	16
185 - under 190	9
Total	80

This table is of a more manageable size and the clustering of a majority of the observations around the middle of the distribution is quite clear. However, as a result of the grouping we no longer know exactly how many employees were of one particular height. In the first class, for example, we know only that seven employees were of a height of 160 cm or more but less than 165 cm. We have no way of telling, just on the information given by this table, exactly where the seven heights come within the class. As a result of our grouping, therefore, we have lost some accuracy and some information. This type of trade-off will always result.

C. CLASS LIMITS

Choosing Class Limits

If we divide a set of data into classes, there are clearly going to be values which form dividing lines between the classes. These values are called **class limits**. Class limits must be chosen with **considerable care**, paying attention both to the form of the data and the use to which it is to be put.

Consider our grouped distribution of heights. Why could we not simply state the first two classes as 160-165 cm, and 165-170 cm, rather than 160 to under 165 cm, etc.? The reason is that it is not clear into which class a measurement of **exactly 165 cm** would be put. We could not put it into both, as this would produce double-counting, which must be avoided at all costs. Is one possible solution to state the classes in such terms as 160-164 cm, 165-169 cm? It would appear to solve this problem as far as our data is concerned. But what would we do with a value of 164.5 cm? This immediately raises a query regarding the recording of the raw data.

How to Record Observations

The raw data consisted of the heights of 80 people in centimetres, and all were exact whole numbers. Could we honestly expect that 80 people would all have heights that were an exact number of centimetres? Quite obviously not. Therefore, some operation must have been performed on the originally measured heights before they were noted down, and the question is What? There are two strong possibilities. One is that each height was rounded to the nearest cm; the other that only the whole number in centimetres of height were recorded, with any additional fraction being ignored (this procedure is often referred to as **cutting**).

Let's consider what would produce a recorded value of 164 cm under both these procedures.

(a) Rounding

A value of 163.5 cm would be recorded as 164 cm (working on the principle that decimals of 0.5 and above are always rounded up), and so would all values up to and including 164.49999....cm.

(b) Cutting

A value of 164 cm would be recorded as 164 cm, and so would all values up to and including 164.99999....cm.

Applying the same principles to other values, we can see that if the data had been cut, the class stated as 160 - under 165 cm would represent exactly that, i.e. all values from exactly 160 cm up to but not including 165 cm. If the data had been rounded, however, the class would represent all values from exactly 159.9 cm up to, but not including, 164.5 cm. In other words, a measured value of, say, 164.7 cm would be recorded as a rounded value of 165 cm and would appear in the grouped frequency distribution in the 165 to under 170 cm class, **not** the 160 to under 165 cm class. From this

you can see that it is advisable always to discover how data has been recorded before you do anything with it.

Thus we can see that both the form in which raw data has been recorded, and whether the variable in question is discrete or continuous, play an important part in determining class limits.

D. CLASS INTERVALS

Definition

The width of a class is the difference between its two class limits, and is known as the **class interval**. It is essential that the class interval should be able to be used in calculations, and for this reason we need to make a slight approximation in the case of continuous variables.

The true class limits of the first class in our distribution of heights (if the data has been rounded) are 159.5 cm and 164.4999...cm. Therefore the class interval is 4.999...cm. However, for calculation purposes we approximate slightly and say that because the lower limit of the first class is 159.5 cm and that of the next class is 164.5 cm, the class interval of the first class is the difference between the two, i.e. 5 cm.

Unequal Class Intervals

You will see that, using this definition, all the classes in our group frequency distribution of heights have the same class interval, 5 cm. While this will almost certainly make calculations based on the distribution simpler than might otherwise be the case, it is **not** absolutely necessary to have equal class intervals for all the classes in a distribution. If having equal intervals meant that the majority of observations fell into just a few classes, while other classes were virtually empty, there would be a good reason for using unequal class intervals. Often it is a case of trial and error.

Open-ended Classes

Sometimes it may happen that one or both of the end limits of the distribution (the lower limit of the first class and the upper limit of the last class) are not stated exactly. This technique is used if, for example, there are a few observations which are quite spread out some distance from the main body of the distribution, or (as can happen) the exact values of some extreme observations are not known.

If this had occurred with our distribution of heights, the first and last class could have been stated as under 165 cm and 185 cm and over, respectively. (Note the last class would **not** be stated as over 185 cm, because the value 185 cm would then not be included in any class.) Classes such as this are said to be **open-ended** and, for calculation purposes, are assumed to have class intervals equal to those of the class next to them. This does introduce an approximation but, provided the frequencies in these classes are small, the error involved is small enough to be ignored.

E. CHOOSING CLASS LIMITS AND INTERVALS

Which Choices Need to be Made?

In choosing classes into which to group the set of heights, you have two decisions to make: one related to the position of class limits, the other to the size of class intervals. We chose limits of 160 cm, etc., and a class interval of 5 cm (although, as we have seen, we need not have kept the class interval constant throughout the distribution). These are not the correct values, because there are no

such things as correct values in this context. There are, however, some values which are better than others in any particular example.

Reasons for Choice

Firstly, we noted that the observations taken as a set were quite compact; there were no extreme values widely dispersed from the main body of the distribution. Consequently we did not need to use open-ended classes, or classes with a wider than normal interval, to accommodate such values. We could thus make all the class intervals equal.

Purely for the sake of ease of calculation and tidiness of presentation, we chose a class interval of 5 cm. Why did we not use 10 cm as the class interval? Surely, you may ask, that would make calculation even easier? Yes, it would; but it would also mean that we would have only three or four classes (depending on where we fixed the class limits), and that is not enough.

We have seen that grouping data simplifies it, but it also introduces a considerable amount of approximation. The smaller the number of classes, the wider will be the class intervals, and so the greater the approximation.

Guidelines on Which to Base Choice

The guidelines which you must consider when choosing class limits and intervals are as follows:

(a) As far as practicable, have **equal** class intervals, but if the spread of the observations implies that you need to use unequal class intervals and/or open-ended classes, then do so.

(b) For ease of calculation, try to work with values which are multiples of 5 to 10 but if this would impose unwarranted restriction on your choice in other ways, then ease of calculation should be sacrificed. (Remember, the main consideration is that your grouped distribution should bear a reasonable resemblance to the original data.)

(c) Try to keep the number of classes between 5 and 15. This will make your distribution simple enough to interpret and work with, but also accurate enough for you to have confidence in the results of your calculations.

F. DIRECT CONSTRUCTION OF A GROUPED FREQUENCY DISTRIBUTION

We obtained the grouped frequency distribution of employees' heights from the raw data by constructing an array from all of the data, then an ungrouped distribution, and finally a grouped distribution. It is not necessary to go through all these stages – a grouped frequency distribution may be obtained directly from a set of raw data.

Determining Frequencies by Using Tally Marks

A short study of a set of raw data will enable you to determine (not necessarily exactly – approximate values are sufficient at this stage) the highest and lowest observations, and the spread of the data, e.g. are the observations closely packed together; are there a few extreme observations? etc. On this basis you will be able to set up initial classes.

Then go through the raw data item by item, allocating each observation to its appropriate class interval. This is easily done by writing out a list of classes and then using "tally marks" – putting a mark against a particular class each time an observation falls within that class; every fifth mark is put diagonally through the previous four. Thus the marks appear in groups of five.

This makes the final summation simpler, and less liable to error. Thus, for the "160 to under 165" cm class in the distribution of heights, the tally marks would appear as |||| || giving a frequency of 5 + 2 = 7. Similarly, the "165 to under 170 cm" class would appear as giving a frequency of ||. Having obtained the frequencies for each class, first check that they do sum to the known total frequency. It is essential that errors are eliminated at this stage.

Possible Revision of Initial Distribution

By looking at the grouped frequency distribution you have constructed, you will be able to see if it can be improved. You may decide that groups at either end of the distribution have such low frequencies that they need to be combined with a neighbouring class; and a look at exactly where the extreme observations lie will help you to make the decision as to whether or not the first and last classes should be open-ended. You may decide that, although your class intervals are correct, the class limits ought to be altered.

If some classes (particularly those near the middle of the distribution) have very high frequencies compared with the others, you may decide to split them, thus producing a larger number of classes, some of which have a smaller interval than they did originally. With practice you will acquire the ability to make such decisions.

G. CUMULATIVE FREQUENCY DISTRIBUTIONS

So far we have discovered how to tabulate a frequency distribution. There is a further way of presenting frequencies and that is by forming **cumulative frequencies**. This technique conveys a considerable degree of information and involves adding up the number of times (frequencies) values less than or equal to a particular value occur.

You will find this easier to understand by working through our example on employees' heights in Table 3.4. We start with the value 0 as there are no employees less than 160 cm in height. There were seven employees with a height between 160 and less than 165 cm. Therefore the total number of employees less than 165 cm in height is seven. Adding the number in the class "165 but under 170 cm", you find that the total number of employees less than 170 cm in height is 18. There are 35 employees who are not as tall as 175 cm, and so on. The cumulative frequencies are shown in Table 3.5.

Table 3.5: Less Than Cumulative Frequencies Table of Employees' Heights

Height *(cm)*	**Frequency**	**Cumulative Frequencies**
Under 165	7	0 + 7 = 7
Under 170	11	7 + 11 = 18
Under 175	17	18 + 17 = 35
Under 180	20	35 + 20 = 55
Under 185	16	55 + 16 = 71
Under 190	9	71 + 9 = 80
	80	

You can see that the simplest way to calculate cumulative frequencies is by adding together the actual frequency in the class to the cumulative frequencies of the previous classes. It can also work in reverse if you want to obtain class frequencies from cumulative frequencies. Work it out for the above example, and you will see how easy it is. You will also notice in the table that the class descriptions have changed slightly, to read "Under 165 cm", etc. This is a true description of what the cumulative frequencies actually represent.

It is possible to switch the descriptions round, so that they read: "More than 160", "More than 165", etc., as shown in the following table. This is known as the more than cumulative frequency distribution, as set out in Table 3.6.

Table 3.6: More Than Cumulative Frequency Table of Employees' Heights

Heights *(cm)*	**Cumulative Frequencies**
More than 160	80
More than 165	73
More than 170	62
More than 175	45
More than 180	25
More than 185	9

However, distributions are not usually presented in this way. In future examples we shall deal solely with the less than cumulative frequency distribution.

H. RELATIVE FREQUENCY DISTRIBUTIONS

Relative Frequency

Relative frequencies are the actual number of frequencies in a class divided by the total number of observations, i.e.:

$$\text{Relative frequency} = \frac{\text{Actual frequency}}{\text{Total number of observations}}$$

Let's go back to our example of employees' heights. There are 7/80 or 0.0875 employees who are less than 165 cm tall, and 20/80 or 0.25 (one quarter) who are between 175 cm and under 180 cm tall. Table 3.7 shows the relative frequencies.

Table 3.7: Relative Frequencies of Employees' Heights

Heights *(cm)*	**Frequency**	**Relative Frequency**				
160 - under 165	7	7/80	=	0.0875	or	8.75%
165 - under 170	11	11/80	=	0.1375	or	13.75%
170 - under 175	17	17/80	=	0.2125	or	21.25%
175 - under 180	20	20/80	=	0.25	or	25.00%
180 - under 185	16	16/80	=	0.2	or	20.00%
185 - under 190	9	9/80	=	0.1125	or	11.25%
	80					

In Table 3.7 we have expressed the fractions also as percentages, something that is extremely useful and that improves a table. You can see at a glance that 20% of all employees measured were more than 180 cm, but less than 185 cm tall. The main advantage of relative frequencies is their ability to describe data better.

Cumulative Relative Frequency

We have seen how to calculate cumulative frequencies. Using the same logic, you can obtain cumulative relative frequencies by adding the relative frequencies in a particular class to that already arrived at for previous classes. See Table 3.8:

Table 3.8: Cumulative Relative Frequencies of Employees' Heights

Heights *(cm)*	**Cumulative Relative Frequency**	**Cumulative Percentage**
Under 165	0.0875	8.75
Under 170	0.225	22.5
Under 175	0.4375	43.75
Under 180	0.6875	68.75
Under 185	0.8875	88.75
Under 190	1.0000	100.00

You will notice that, in the above table, an extra column has been added which is labelled Cumulative Percentage. This column is the cumulative relative frequency converted to a percentage. This makes it easier for conclusions to be drawn from this table. For example, 88.75% of all employees measured were less than 185 cm tall.

J. WAYS OF PRESENTING FREQUENCY DISTRIBUTIONS

We have seen how to tabulate frequency distributions, and we now have to consider ways of bringing these distributions to life by presenting them in such a way that, even though some of the detail may

be lost, the main points contained in the data come across to the reader. We shall look at various types of diagram that are commonly used to represent frequency distributions.

Histograms

A histogram can be used to present discrete data, although it is more commonly used to illustrate continuous data. However, first we will look at its use for discrete variables; this will make it easier to follow its use in describing a frequency distribution of a continuous variable.

(a) Discrete Variables

Our data is in Table 3.9:

Table 3.9: Hand-built Cars Produced in a Month

Cars produced	Frequency
1	4
2	5
3	11
4	6
5	3
6	1
7	0
	30

We shall plot vertical rectangles on the x-axis. The width of these rectangles on the horizontal axis will depend on the class each represents. In our case, as the values in each class are discrete, i.e. 0, 1, 2 .. 7, the width is +0.5, with the discrete values being the midpoints.

Each measurement or value of the variable falls within a class (as described earlier). Every class has a midpoint and a pair of limits, referred to as **class limits**. Sometimes, as in our example, there will be only one value in a class and the midpoint of that class is that value. A class encompasses all values between the class limits. In the example, the class limits were +0.5. Therefore, in the class 0.5 to 1.5 all values between these limits would be added into the class. A **midpoint** is the centre of the class and is the value you usually consider that class to represent.

Obviously, in an example which includes only discrete variables, this would apply only to the discrete values. However, it is more relevant to continuous variables, as you will see later.

Each class is plotted on the x-axis. The y-axis measures the frequency with which the observations occurred in each of the x-axis classes. In a histogram, each observation is represented by a **finite amount of space**. The space assigned to each observation is the same in every case. Since each space represents an observation and the dimensions of that finite space **do not vary** within each class, the sum of these spaces constitutes an area analogous to the sum of the frequencies.

Figure 3.1 represents the data on hand-built cars. As you can see, all classes are one car in size, although the class limits range from ±0.5. The midpoints are 0, 1, 2, 3 .. 7.

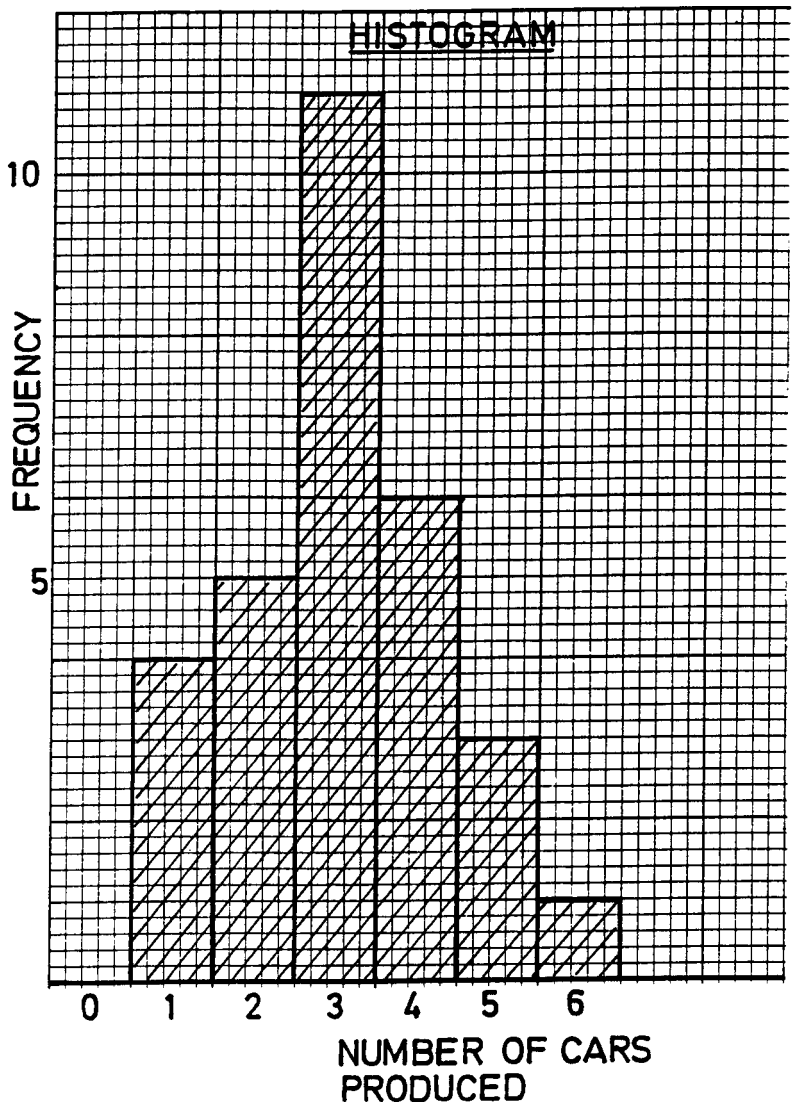


Figure 3.1

In Figure 3.1 the frequency in each class is represented by the rectangle. However, it is important to realise that it is **not the height of the rectangle** that represents the frequency, but **the area within the rectangle**. If we now move on to using histograms to plot continuous variables, this point will become clearer.

(b) Continuous Variables

Consider the following data on employees' heights:

Table 3.10

Height *(cm)*	**Frequency**
160 but under 165	7
165 but under 170	11
170 but under 175	17
175 but under 180	20
180 but under 185	16
185 but under 190	9
	80

A histogram would show this information much more clearly and interestingly. If the frequencies involved are very large, it is unrealistic to plot every observation separately, so a scale can be introduced on the y-axis. See Figure 3.2(a).

The class limits are 160-164.9, 165-169.9, 170-174.9.....185-189.9, all having a range of 5 cm. Therefore the midpoints are 162.5, 167.5, 172.5 .. 187.5.

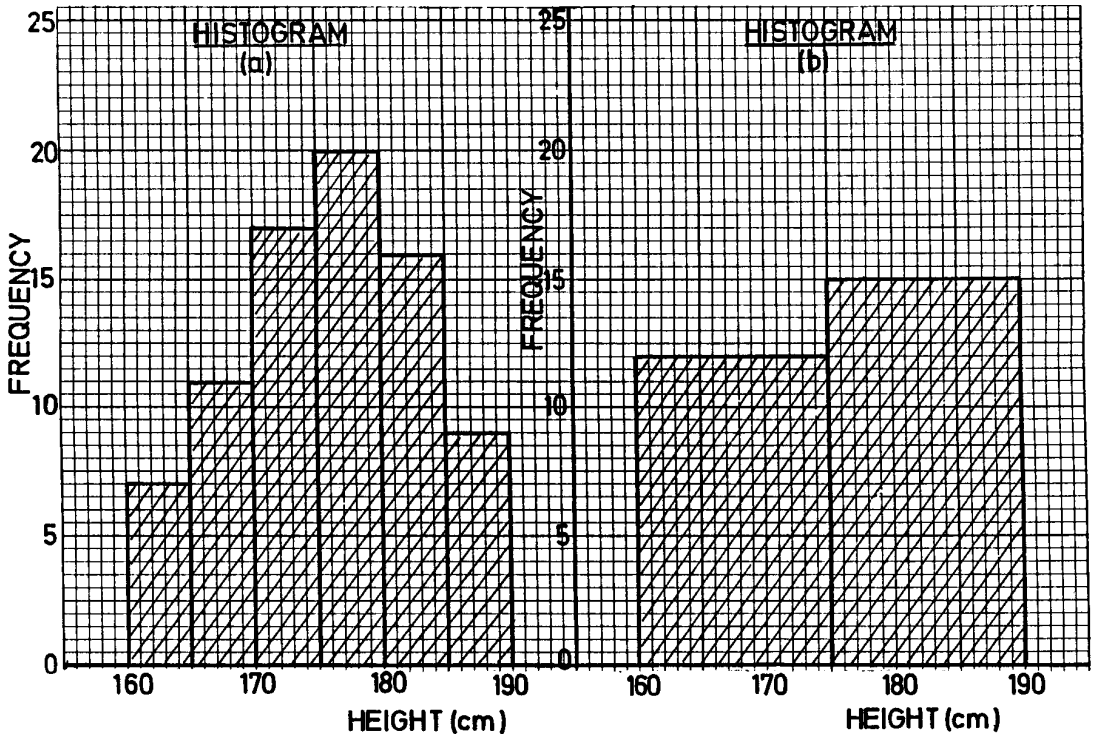


Figure 3.2

You can see from the histogram that the most frequent class is 175 to 180 cm. If the classes were grouped differently, another picture would emerge. Assume that only two classes were

used – “160 but under 175” and “175 but under 190”; this would give frequencies of 35 and 45 respectively, with the resultant histogram, shown in Figure 3.2(b), being meaningless. This is an extreme example but it shows the importance of **proper construction of classes**.

Something strange appears in this example. Why is it that the frequencies between 175 and 190 are not drawn at 45? This is because frequencies are represented by area, **not** height, of the rectangle. How to calculate them is shown below.

Let’s look at the problem involved in drawing a histogram where the class limits are **unequal**. We will keep the example as before but change the class limits and rework the frequencies, as follows:

Table 3.11: Revised Frequency Distribution

Height *(cm)*	**Frequency**
160 but under 164	4
164 but under 168	9
168 but under 176	25
176 but under 184	32
184 but under 190	10
	80

First look at the class widths of our new frequency distribution:

Table 3.12

Height *(cm)*	**Class Widths** *(cm)*
160 but under 164	4
164 but under 168	4
168 but under 176	8
176 but under 184	8
184 but under 190	6

In drawing the histogram, this time one unit in the horizontal scale will represent a class width of 4 cm in height. Therefore the first two classes are one unit wide, the next two classes are two units (i.e. 8 cm ÷ 4 cm), with the final class being 1.5 units (i.e. 6 cm ÷ 4 cm).

Using the frequencies set out above, the **heights of the rectangles** in each class are the **frequencies in each class divided by the number of units applicable to that class**. These heights work out as follows:

Table 3.13

Heights *(cm)*	**Height of Rectangle**
160 but under 164	$\frac{4}{1} = 4$
164 but under 168	$\frac{9}{1} = 9$
168 but under 176	$\frac{25}{2} = 12.5$
176 but under 184	$\frac{32}{2} = 16$
184 but under 190	$\frac{10}{1.5} = 6.66$

The histogram is shown in Figure 3.3.

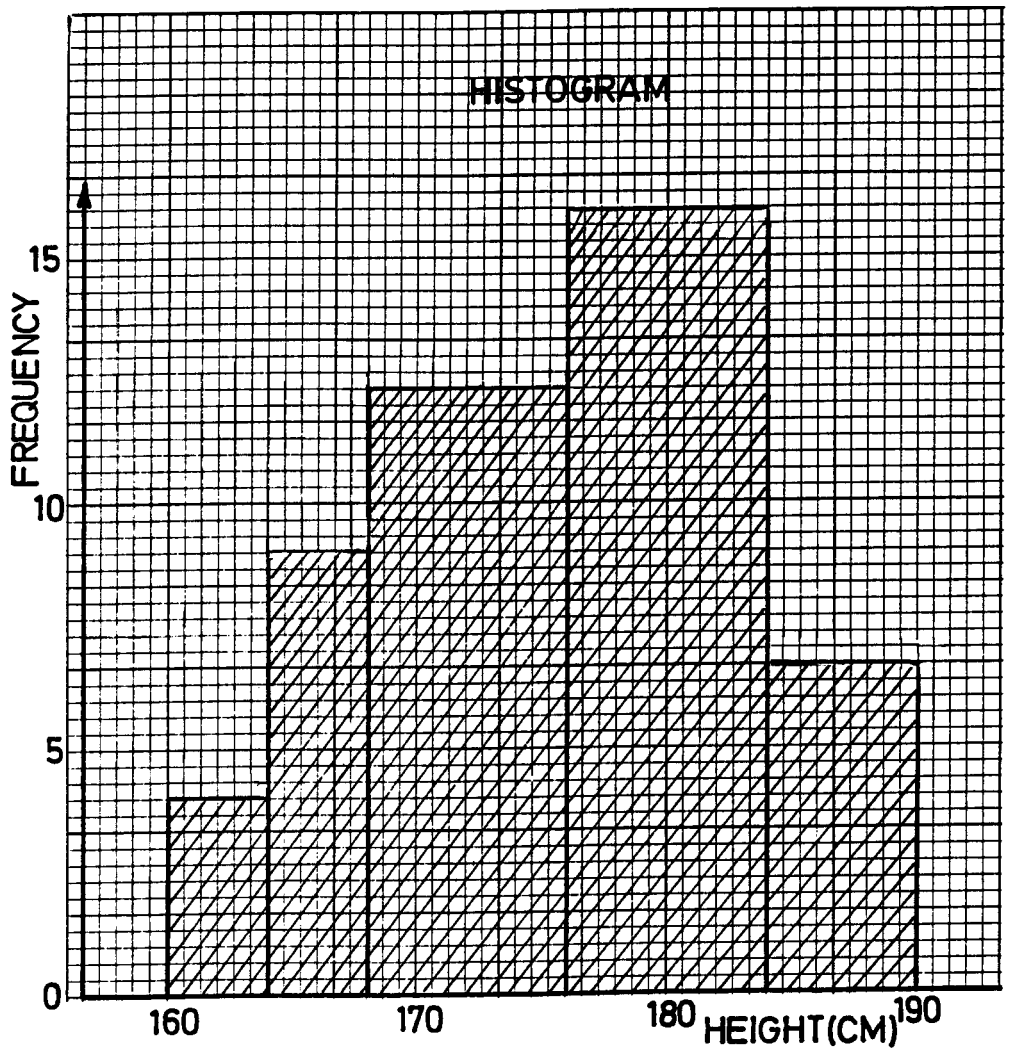


Figure 3.3

The reduction in the number of classes from six to five has meant some loss of detail in the histogram, compared with that shown in Figure 3.2(a).

This last example has shown you how to handle grouped frequency data with classes of unequal size. Although this practice is not recommended, it is often used and this type of question may well come up in the examination.

Another practice sometimes used, although again not recommended, is open-ended classes. In our original example, the last class could be referred to as "185 and above". It would thus be necessary to make an intelligent assessment of the likely class width. This guess could be based on an inspection of the known class limits and the shape of the distribution drawn so far.

In our example, after the peak is reached at the class "175 but under 180", frequencies diminish. It would therefore be foolish for you to group all observations under the value 185 and end up by drawing a tall, thin rectangle nearly as high as the previous class. It would be realistic to use the same class widths as all previous classes – which is what happened in our frequency distribution. Thus, with open-ended classes, **use your common sense**, tempered with an inspection of the shape of the distribution.

There is one further diagram that can be used to represent frequency distributions of both discrete and continuous variables. This is called a **frequency polygon**.

Frequency Polygon

This is a very quick method of drawing the shape of a frequency distribution. Refer back to Figure 3.2 (a), which shows the histogram drawn for our original data on heights of individuals. By joining together the midpoints at the top of each rectangle, we form a frequency polygon (see Figure 3.4).

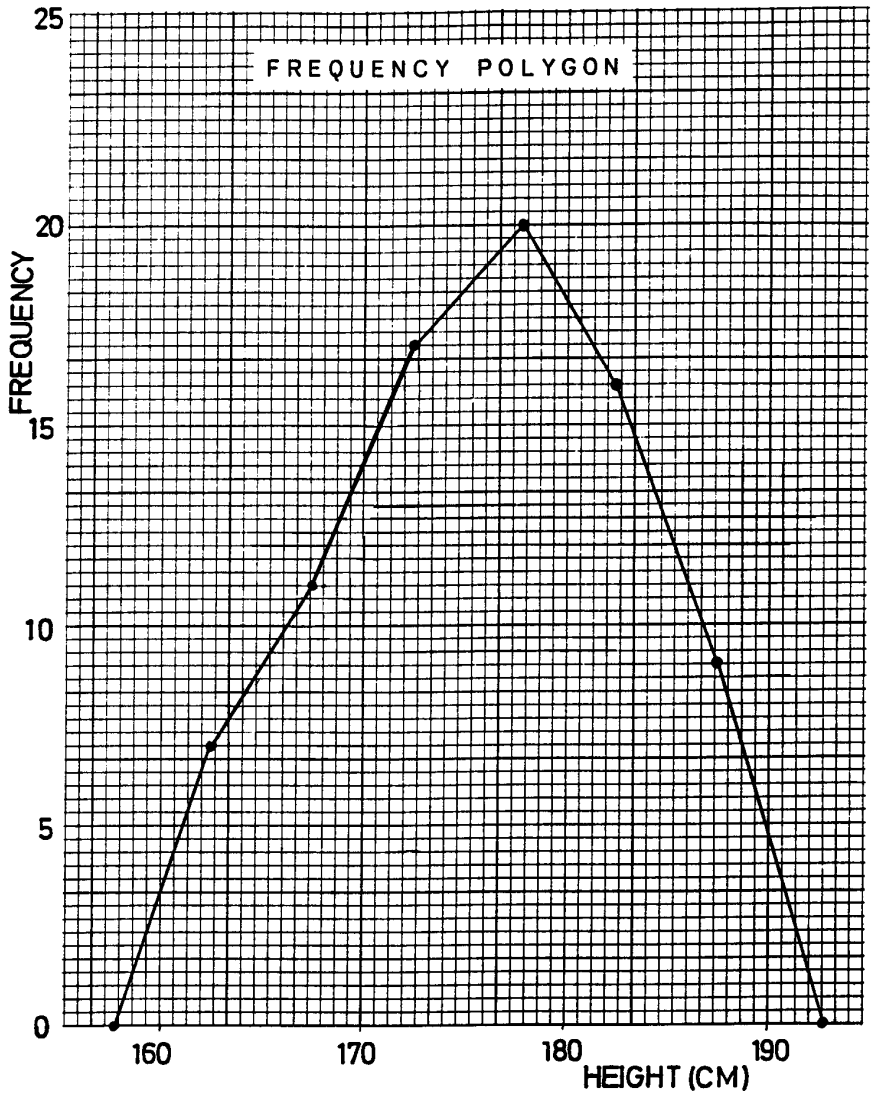


Figure 3.4

As you can see in Figure 3.4, it is necessary to extend the lines at each end of the histogram to the midpoints of the next highest and lowest classes, which will have a frequency of zero. The lines are extended to the x-axis, so that the area of the polygon will equal that of the histogram it represents. This is **vital**; this principle is extremely important in a number of branches of statistics. It is not necessary, however, to draw the histogram first. The polygon can be drawn just by plotting the frequencies of the classes at their midpoints.

You will often find that an examination question will ask you to describe a frequency polygon and its uses. Be sure you understand exactly what it is and how to draw one.

K. PRESENTING CUMULATIVE FREQUENCY DISTRIBUTIONS

Cumulative Frequency Polygon

This type of diagram is used to represent **cumulative frequencies**. As in all diagrams that plot frequency distributions, the frequencies are plotted on the y-axis, with the values of the variables on the x-axis. The difference this time is that the y-axis contains the cumulative frequencies, starting from zero and finishing at the **total** number of frequencies. The cumulative frequencies are plotted not on the midpoints of their respective classes, as for histograms, but at the upper limits of the class. This is called a **less than** cumulative frequency distribution, and is the one commonly plotted.

Height	Cumulative Frequency
Less than 160	0
Less than 165	7
Less than 170	18
Less than 175	35
Less than 180	55
Less than 185	71
Less than 190	80

Height	Cumulative Frequency
More than 160	80
More than 165	73
More than 170	62
More than 175	45
More than 180	25
More than 185	9
More than 190	0

If the more than type of cumulative frequency distribution is required, then the upper limits are still used but the diagram starts with the total frequency and finishes with zero. The less than and more than cumulative frequency polygons for our example on employees' heights are shown in Figure 3.5(a) and (b) respectively.

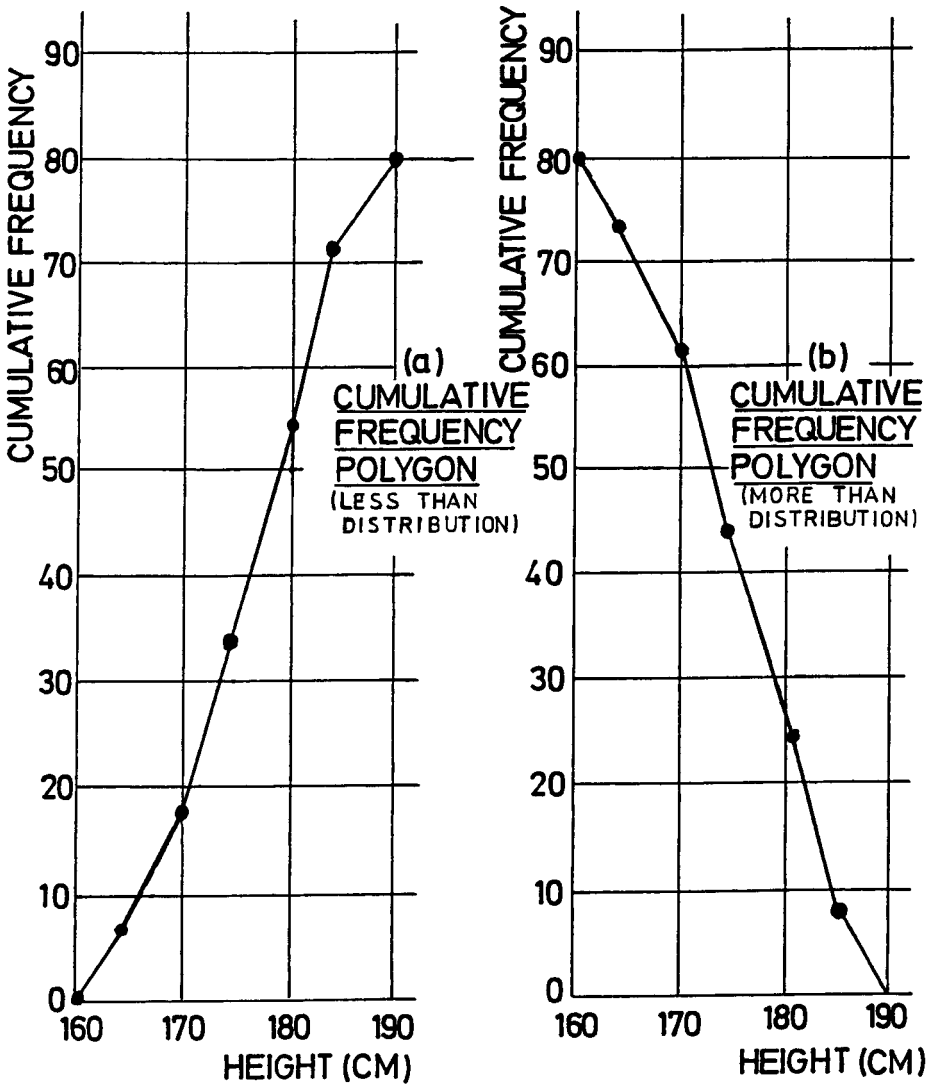


Figure 3.5

In a cumulative frequency polygon, the cumulative frequencies are joined together by straight lines. In a cumulative frequency curve, a smooth curve joins the points.

This type of diagram is often referred to as an **ogive**. An ogive is basically the name given to a graph or diagram of a cumulative frequency distribution.

Percentage Ogive

If you wish to present the **percentage relative cumulative frequency**, which is often referred to as a percentage cumulative frequency, then a percentage ogive is used. In this diagram the percentage cumulative frequencies are again plotted on the y-axis and the points joined together by a **smooth** curve. Figure 3.6 shows the percentage ogive of the information on employees' heights.

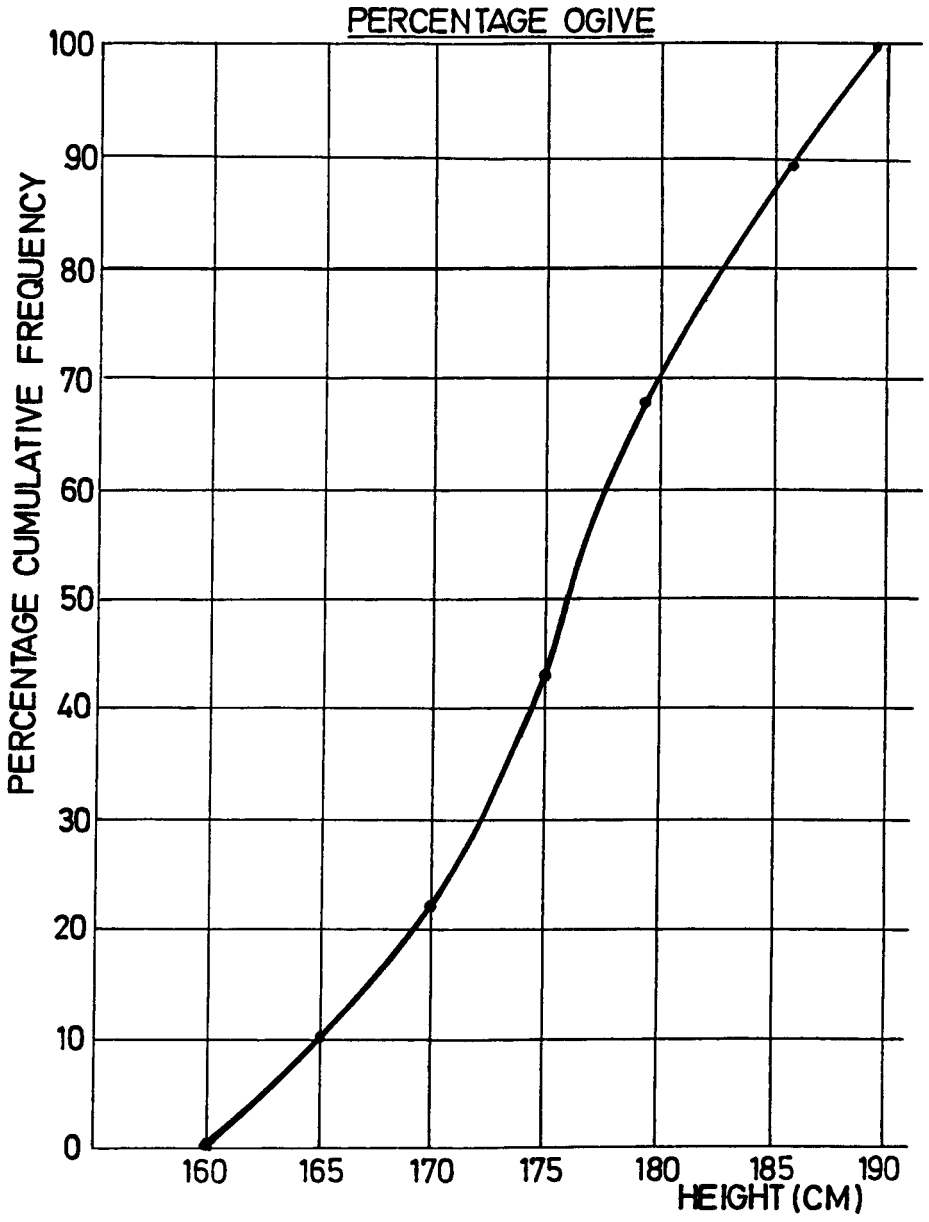


Figure 3.6

The diagram plots an upward curve starting at zero % and ending at 100%. This is extremely useful when it comes to comparing several distributions. Say, for example, you have collected two sets of observations on employees' heights – one on women's heights and the other on men's heights. It would be possible to plot the percentage cumulative frequencies for each distribution on the same diagram. Thus the differences, if any, could be detected at a glance.

A percentage ogive could also be used to make statements concerning the characteristics of the observed distribution in percentage terms. For example, 50% of those employees measured were taller than 176 cm, etc.

L. FREQUENCY CURVE

Suppose your raw data consisted of a large number of observations of a continuous variable, which were subsequently formed into a large number of classes. It would, of course, be possible to construct a histogram using these classes. The larger the number of classes, the narrower would become the rectangles in the histogram. If a line was drawn through the tops of the rectangles as if constructing a frequency polygon, eventually, as the numbers of classes increased, so the straight lines joining together the rectangles would become a smooth curve. It is this curve that is known as a **frequency curve** and is illustrated in Figure 3.7:

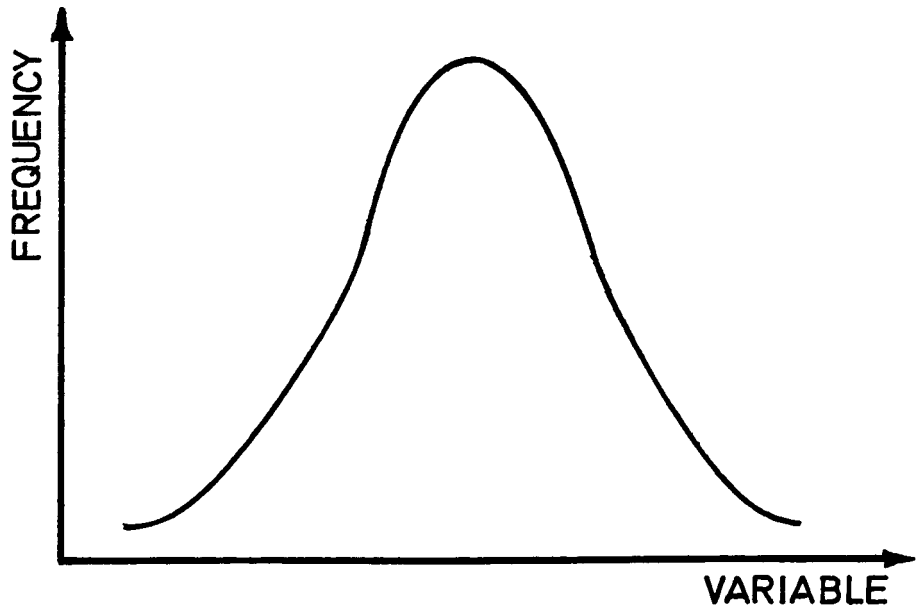


Figure 3.7

The concept behind a frequency curve is fundamental to all statistics.

All distributions that we obtain by one means or another are incomplete. They contain only a **proportion** of what may really be available. Take our example of heights of employees. We were only able to draw a distribution containing only 80 measurements. This may have been 1/10th or 1/100th of the total number of employees available. Therefore, the resulting frequency curve which we drew was only an approximation of the true curve which would be obtained from all the employees. As we shall see later, it is a good approximation but nevertheless it is still only part of the whole distribution.

In later study units, notably on normal distribution, we shall use the idea of a frequency curve to illustrate distributions of variables, even though actual figures may not be presented.

Study Unit 4

Statistical Charts and Diagrams

Contents *Page*

A. PURPOSE OF GRAPHICAL METHODS

Graphs and diagrams are used mainly for efficient and convenient **presentation** of statistical data and results. They are **not** generally used for the actual *analysis* of data. They may, however, be of use in indicating what kind of analysis is feasible.

B. PICTOGRAMS

One of the common ways of presenting statistical data to the general public is by means of diagrams in which the information is represented by small drawings. For example, the imports of oil in a particular year may be represented by a number of drawings of barrels, and the imports for another year by a different number of barrels, as in Figure 4.1.

To give another example, the strengths of the armies of several different nations may be represented by drawings of a number of soldiers. You will be able to note other examples in newspapers, business magazines and government pamphlets.

An alternative sometimes used is to employ one diagram for each year (or nation, etc.) but to vary the size of it. This is not a very good idea because there is always doubt as to whether it is **dimensions** or **area** of the diagram which is relevant. Always use the former method for preference.

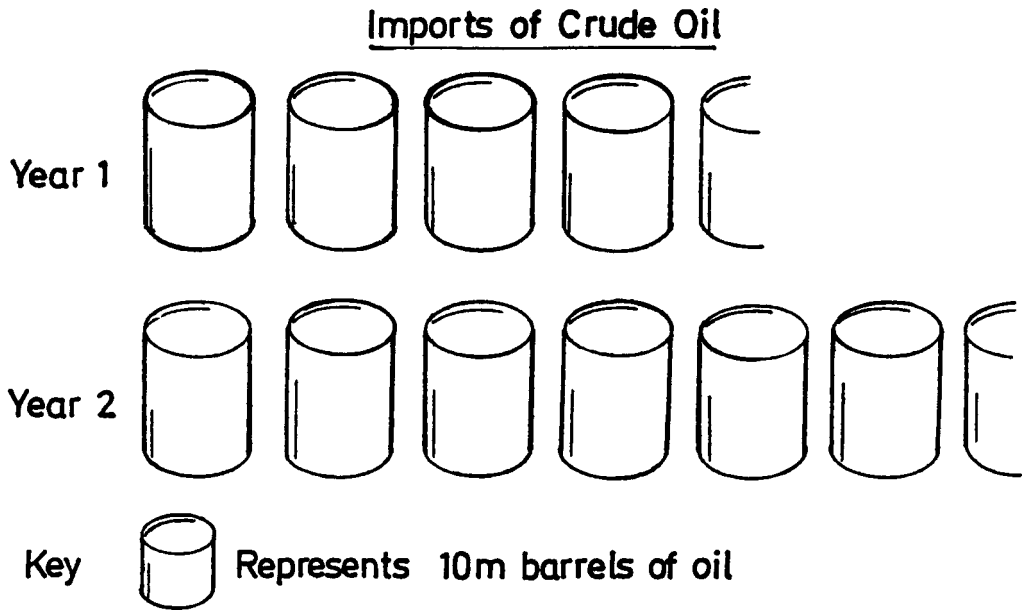


Figure 4.1

These diagrams are variously called pictograms, ideograms, picturegrams or isotypes – the words all refer to the same thing. Their use is confined to the simplified presentation of statistical data for the general public.

They are not precise enough for other purposes; in Figure 4.1 it is difficult to represent a quantity less than 10m barrels accurately.

C. CIRCULAR DIAGRAMS

These diagrams, known also as **pie charts**, are used to show how various components add up to a total. Like pictograms, they are used to display only very simple information.

Suppose we wish to illustrate the sales of gas in Great Britain in a certain year. The figures are taken from the Annual Abstract of Statistics as follows:

Gas Sales in Great Britain in ....

Uses	**Million Therms**	**%**
Domestic	1,383	51
Industrial	843	31
Commercial	437	16
Public*	60	2
Total	2,723	100

* *Central and local government uses, including public lighting*

The figures are illustrated in the pie or circle diagram in Figure 4.2.

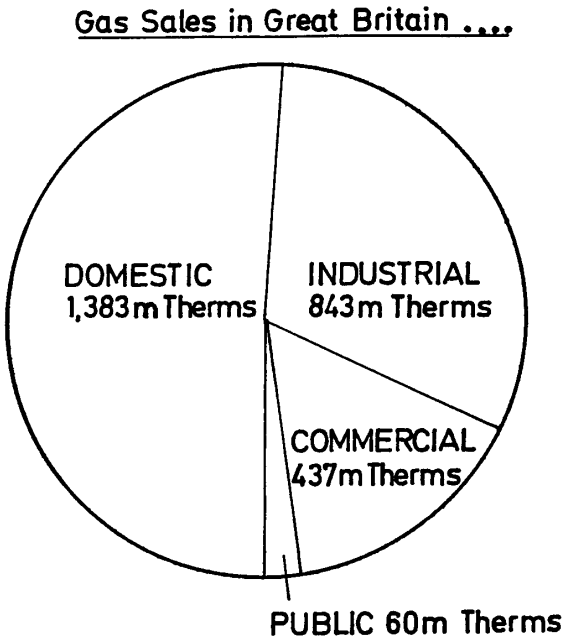


Figure 4.2

To construct the pie chart, the rules to follow are:

(a) Tabulate the data and calculate the percentages.

(b) Convert the percentages into degrees, e.g.

$$51\% \text{ of } 360^\circ = \frac{51}{100} \times 360^\circ = 183.6^\circ, \text{ etc.}$$

(c) Construct the diagram by means of a pair of compasses and a protractor. Don't overlook this point, because examiners dislike inaccurate and roughly drawn diagrams.

(d) Label the diagram clearly, using a separate **legend** or **key** if necessary.

(e) It is best not to use a diagram of this kind with more than four or five component parts.

Note: The actual number of therms can be inserted on each sector as it is not possible to read this exactly from the diagram itself.

The main use of a pie chart is to show the relationship each component part bears to the whole. They are sometimes used side by side to provide comparisons, but this is not really to be recommended unless the whole diagram in each case represents exactly the same total amount, as other diagrams (such as bar charts) are much clearer.

D. BAR CHARTS

A bar is simply another name for a thick line. In a frequency bar chart the bars represent, by their length, the frequencies of different values of the variable. The idea of a bar chart can, however, be extended beyond the field of frequency distributions, and we will now illustrate different types of bar chart in common use.

Component Bar Chart

This serves the same purpose as a circular diagram and, for that reason, is sometimes called a component bar diagram – see Figure 4.3. The lengths of the components represent the amounts, and the components are drawn in the same order so as to facilitate comparison. These bar charts are preferable to circular diagrams because:

(a) They are easily read, even where there are many components.

(b) They are more easily drawn.

(c) It is easier to compare several bars side by side than several circles.

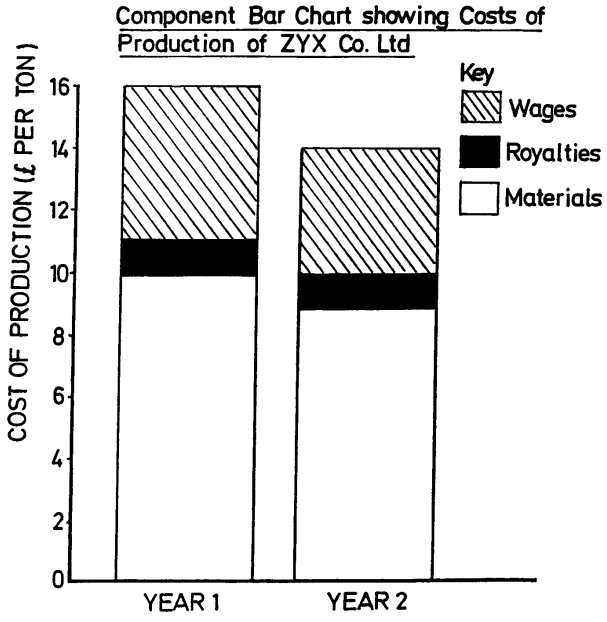


Figure 4.3

Bar charts with vertical bars are sometimes called column charts to distinguish them from those in which the bars are horizontal (see Figure 4.4).

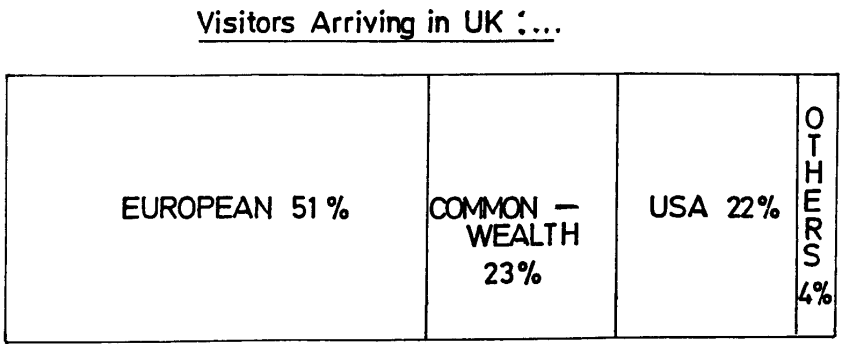


Figure 4.4

In **percentage component bar charts** the information is expressed in percentages rather than in actual numbers of visitors. If you compare several percentage component bar charts, you must be careful. Each bar chart will be the same length, as they each represent 100%, but they will not necessarily represent the same actual quantities, e.g. 50% of the visitors arriving in the UK in 1960 might have been 1 million, whereas in 1970 it was probably nearer 4 million and as many as 8 million in 1980.

Horizontal Bar Charts

A typical case of representation by a horizontal bar chart is shown in Figure 4.5. Note how a loss is shown by drawing the bar on the other side of the zero line.

Pie charts and bar charts are especially useful for categorical variables as well as for numerical variables. The example in Figure 4.5 shows a categorical variable, i.e. the different branches form the different categories, whereas in Figure 4.3 we have a **numerical** variable, namely time.
Figure 4.5 is also an example of a **multiple or compound bar chart** as there is more than one bar for each category. Here we have two bars at each branch corresponding to the profits in the two years.

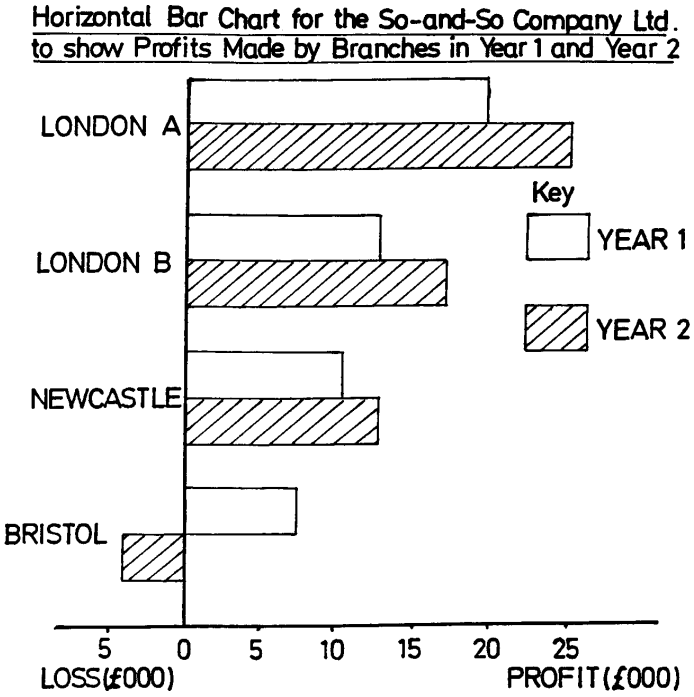


Figure 4.5

Gantt Chart

This is a special type of bar chart developed to show how actual performance and planned performance in, for example, sales or output, compare over a period of time. It is thus often referred to as a **progress** chart. For each period of time over which performance is being monitored, two bar charts are drawn, one giving the planned performance and the other the actual performance.

Example

Sales of Company XYZ

Week	**Sales Quota** *(units)*	**Actual Sales** *(units)*
1	1,500	1,205
2	1,200	1,316
3	1,400	1,452
4	1,500	1,530
5	1,500	1,481

The Gantt chart is shown in Figure 4.6. The thin line denotes the sales quota, and the thick bar represents the actual sales achieved. Any discrepancy between the two can easily be recognised and investigated.

The basic idea can be refined to include the cumulative performance over a longer period of time, and coding can be introduced to indicate on the chart specific reasons for a poor performance in one

particular time period, perhaps arising from shortage of supplies. The chart can also be drawn using actual sales figures as percentages of the quota figures.

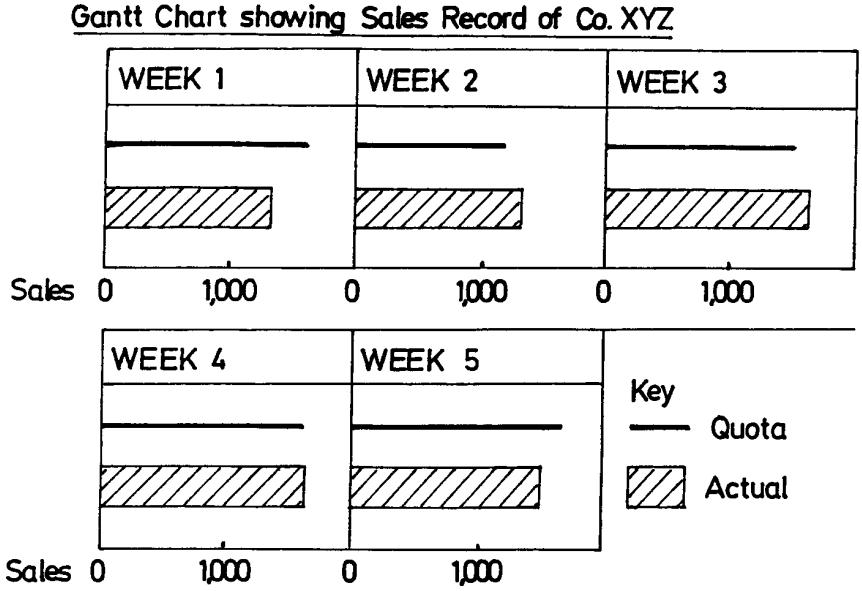


Figure 4.6

E. GENERAL RULES FOR GRAPHICAL PRESENTATION

There are some general rules to remember when planning and using graphical methods:

(a) Graphs and charts must be given clear but brief titles.

(b) The axes of graphs must be clearly labelled, and the scales of the values clearly marked.

(c) Diagrams should be accompanied by the original data, or at least by a reference to the source of the data.

(d) Avoid excessive detail, as this defeats the object of using diagrams.

(e) Wherever necessary, guide lines should be inserted to facilitate reading.

(f) Try to include the origins of scales. (An exception to this rule occurs with logarithmic graphs, which we shall meet shortly.)

F. Z CHART (ZEE CHART)

This is a very useful device for presenting to management on one chart such business information as sales, turnovers, profits, etc. For example, we present data relating to the following questions:

- How are things doing from month to month (or week to week, etc.)?
- How does the current year's performance to date compare with the target or programme?
- How does the present performance compare with that of the same period last year?

The first of these we do by drawing a graph of the **time series** of the data under discussion; the second by drawing a **cumulative target line** and a **cumulative actual line**; the third by plotting a

line showing the **moving annual total**. Here are some figures which we will use to compile the Z chart in Figure 4.7:

The ZYZ Company – Sales For This Year And Last

Month	**Last year's sales** *£000*	**This year's sales** *£000*	**Cumulative for this year** *£000*	**Moving annual total** *£000*
Jan	430	450	450	6,830
Feb	365	340	790	6,805
Mar	365	400	1,190	6,840
Apr	680	680	1,870	6,840
May	560	610	2,480	6,890
Jun	800	760	3,240	6,850
Jul	630	700	3,940	6,920
Aug	760	800	4,740	6,960
Sep	540	570	5,310	6,990
Oct	635	590	5,900	6,945
Nov	630	620	6,520	6,935
Dec	415	430	6,950	6,950

Notes

(a) Last year's figures are needed to enable us to calculate the moving annual totals (MATs).

(b) Each MAT is the sum of the twelve monthly figures up to and including the "present" month, i.e. MAT for January is the sum of the sales figures from February of the previous year up to and including January of the current year. It can be calculated quickly from the previous month's MAT as follows: e.g. for June, take the May figure of 6,890, add the sales for the month of June (760) and deduct the sales for June of the previous year (800).

(c) Each December MAT is the same as the cumulative figure for that month.

(d) The cumulative line starts afresh at the beginning of each year.

(e) Although the chart is shown as complete, it would in practice be kept up-to-date each month as the figures become available.

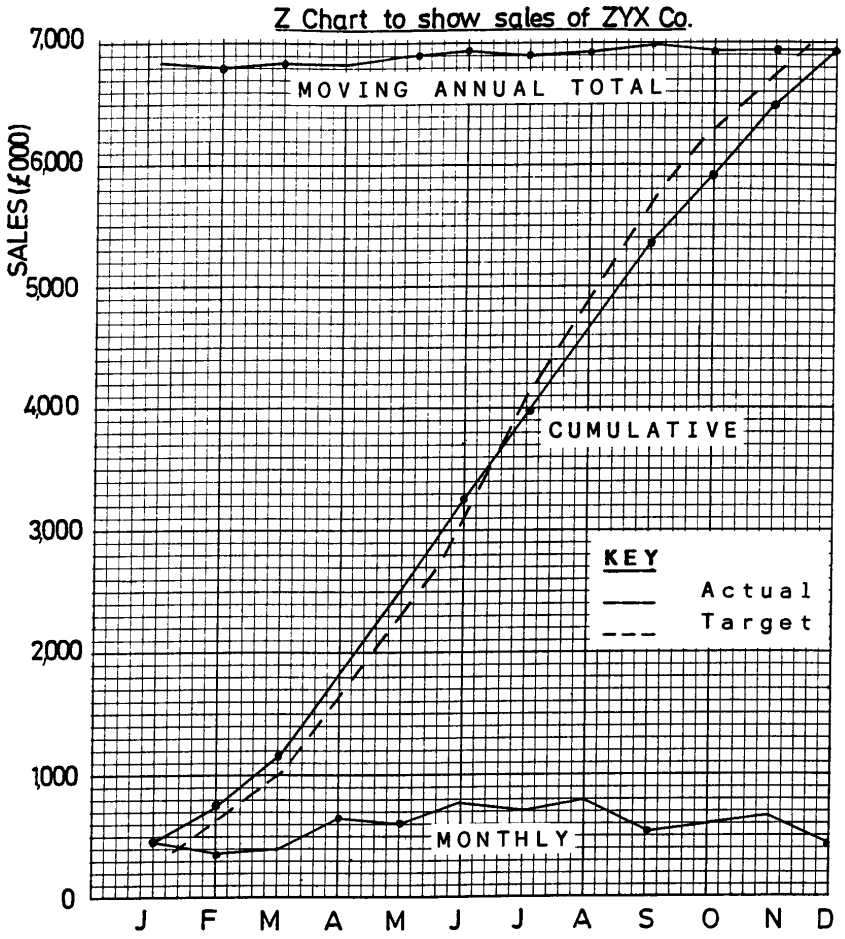


Figure 4.7

(f) The value of the MAT is that it shows at a glance how the current month compares with the same month of last year. If the MAT line slopes up, as it does from August to September, then it shows that this September is better than the previous September. Look at our example: the MAT from August to September goes up from 6,960 to 6,990, showing that this September sales are 30 units higher than last September, which you can check by reference to the "Sales" column. Also, the MAT line acts as a trend line in a time series and gives the general trend for the series.

In the Z chart in Figure 4.7 only one of the scale of "Sales" was used for all the lines. It often happens, however, that the variations in the monthly figures are very small and they do not show up very clearly. In this case a separate scale is used for the monthly figures.

G. LORENZ CURVE

Purpose

One of the problems which frequently confronts the statistician working in economics or industry is that of **concentration**. Suppose that, in a business employing 100 people, the total weekly wages bill is £10,000 and every one of the workers gets £100; there is then an **equal distribution** of wages and there is **no concentration**. In another business employing 100 people and having a total weekly wages bill of £10,000, there are 12 highly skilled experts getting £320 each and 88 unskilled workers

getting £70 each. The wages are not now equally distributed and there is some **concentration** of wages in the hands of the skilled experts. These experts number 12 out of 100 people (i.e. they constitute 12% of the labour force); their share of the total wages bill is 12 x £320 (i.e. £3,840) out of £10,000, which is 38.4%. We can therefore say that 38.4% of the firm's wages is concentrated in the hands of only 12% of its employees.

In the example just discussed there were only two groups, the skilled and the unskilled. In a more realistic case, however, there would be a larger number of groups of people with different wages. For example:

Wages Group (£)	Number of People	Total Wages (£)
0 – 80	205	10,250
80 – 120	200	22,000
120 – 160	35	4,900
160 – 200	30	5,700
200 – 240	20	4,400
240 – 280	10	2,500
	500	49,750

Obviously when we have such a set of figures, the best way to present them is to graph them, which we have done in Figure 4.8. Such a graph is called a **Lorenz Curve**.

Stages in Construction of a Lorenz Curve

(a) Draw up a table giving:

(i) The cumulative frequency

(ii) The percentage cumulative frequency

(iii) The cumulative wages total

(iv) The percentage cumulative wages total

Wages group *(£)*	**Number of people** *(frequency)*	**Cumulative frequency**	**% Cumulative frequency**	**Total wages (£)**	**Cumulative wages total (£)**	**% Cumulative wages total**
0 – 80	205	205	41	10,250	10,250	21
80 – 120	200	405	81	22,000	32,250	65
120 – 160	35	440	88	4,900	37,150	75
160 – 200	30	470	94	5,700	42,850	86
200 – 240	20	490	98	4,400	47,250	95
240 – 280	10	500	100	2,500	49,750	100
	500			49,750		

(b) On graph paper draw scales of 0-100% on both the horizontal and vertical axes. The scales should be the same length on both axes.

(c) Plot the cumulative percentage frequency against the cumulative percentage wages total and join up the points with a smooth curve. Remember that 0% of the employees earn 0% of the total wages, so that the curve will always go through the origin.

(d) Draw in the 45° diagonal. Note that if the wages had been equally distributed, i.e. 50% of the people had earned 50% of the total wages etc., the Lorenz curve would have been this diagonal line.

The graph is shown in Figure 4.8.

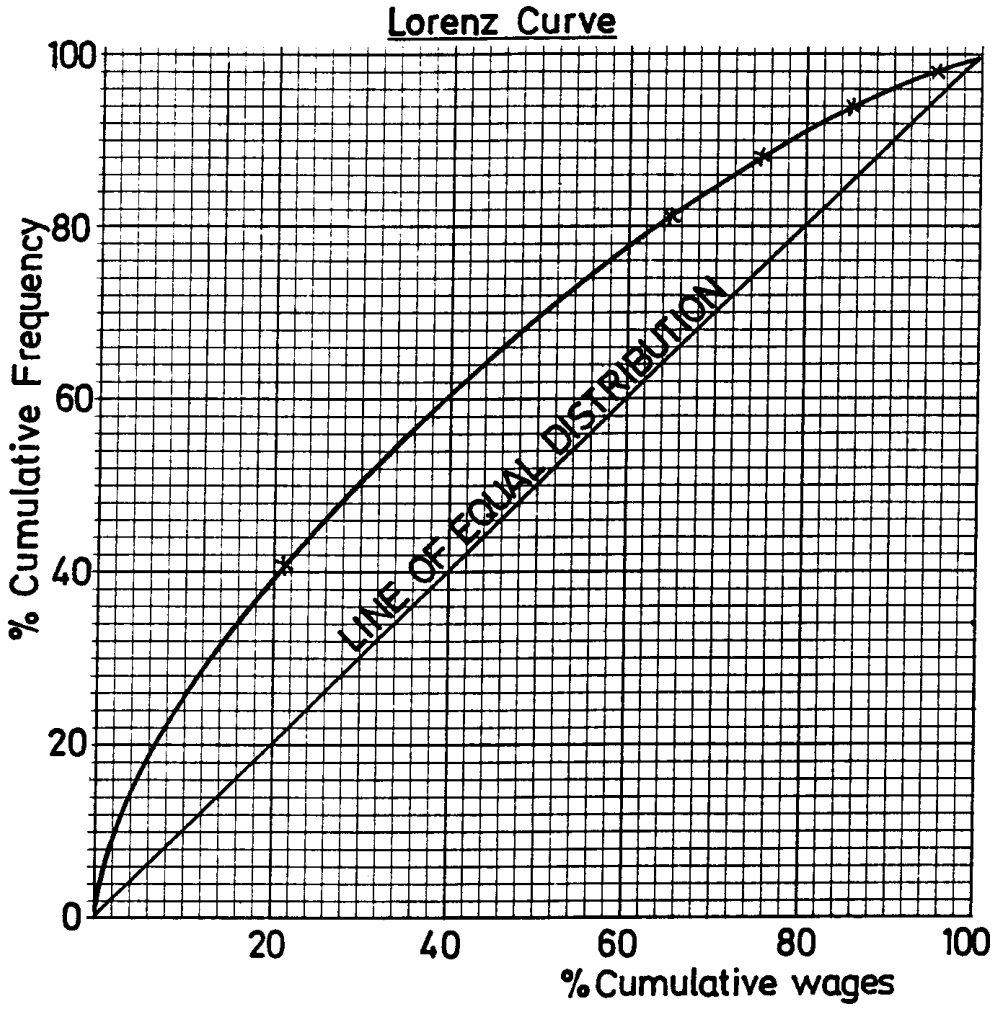


Figure 4.8

H. RATIO SCALES (SEMI-LOG GRAPHS)

Purpose

Look at Figure 4.9 and study the following notes:

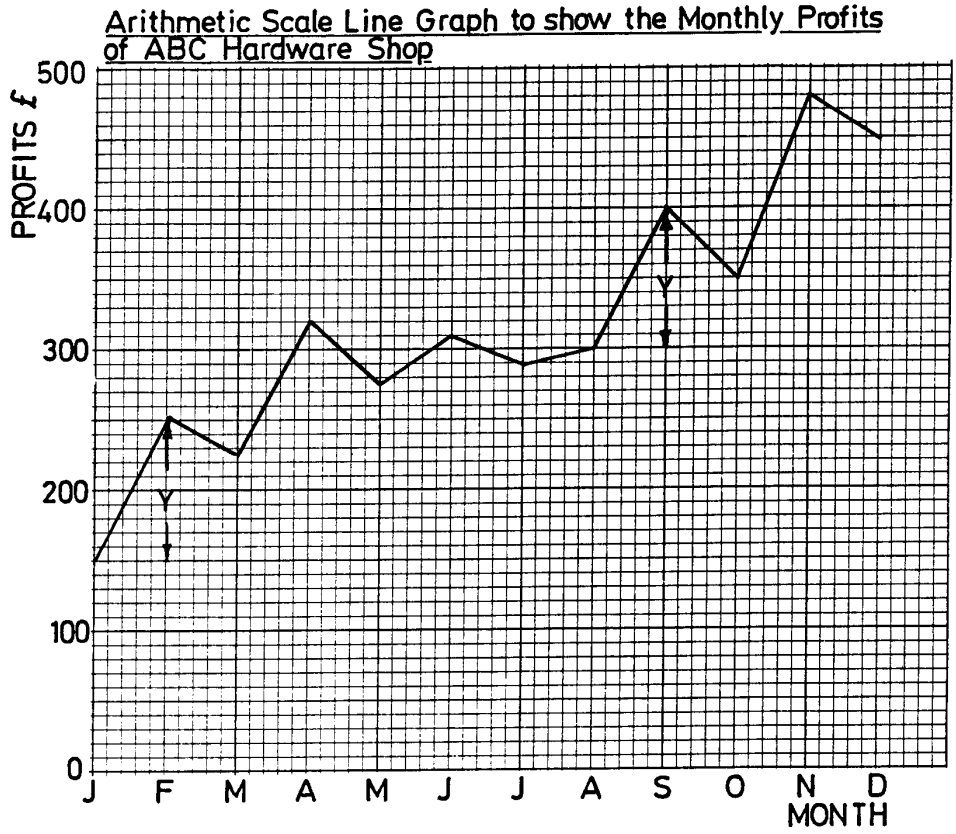


Figure 4.9

(a) On the profits scale, a given distance represents the same change in profits at all parts of the scale. For example, the profits went up by £100 from January to February (from £150 to £250); they also went up by £100 (from £300 to £400) from August to September.

(b) On the graph, the **vertical** distance between January and February is the same as the vertical distance between August and September (shown as Y). In each case Y represents £100.

(c) The angle of slope of the graph from January to February is the same as that from August to September. This indicates that the rate of change is the same in the two cases – £100 per month.

(d) Although the **actual** rate of change is the same in the two cases (£100 per month), the **relative** rate of change is different. £100 is **66⅔**% of January sales (£150) but £100 is only **33⅓**% of August sales (£300) so the relative rate of change from January to February is **66⅔**%, but from August to September is **33⅓**%.

Very often we are more interested in relative changes than in actual changes. It would be convenient, therefore, if we could draw a graph in such a way that equal **percentage** changes looked the same no matter what the actual values were. This can be done by means of a **ratio scale**: examine

Figure 4.10 and you will notice that the vertical scale, instead of going up in steps of equal amounts, goes up by steps of equal multiples. A ratio scale is a logarithmic scale rather than the more usual natural or arithmetic scale used in Figure 4.9.

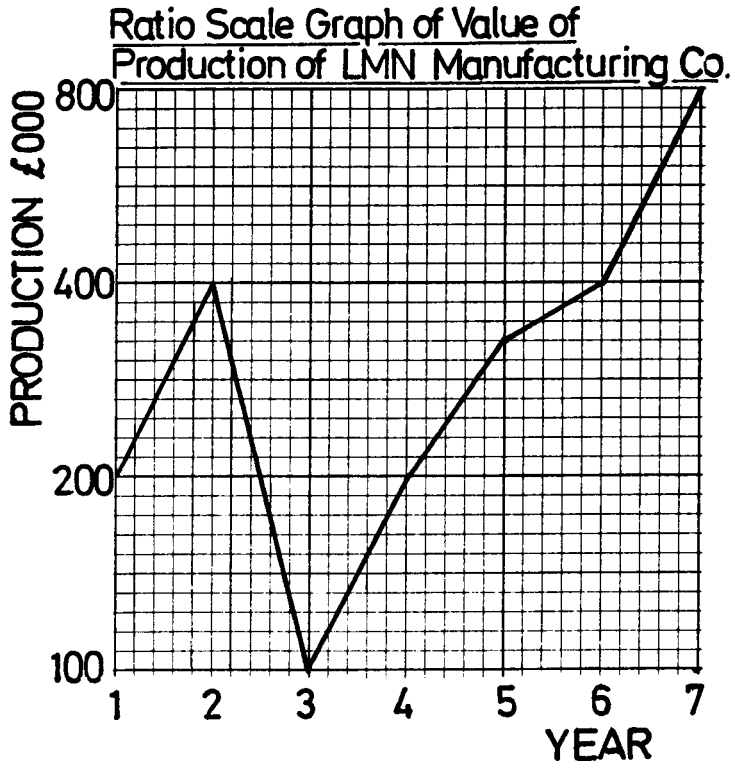


Figure 4.10

On a graph of this kind, equal distances on the vertical scale indicate, not equal amounts, but equal **multiples** or **ratios**. Consequently, changes of equal amounts may look different, but changes of equal proportions (or percentages) look the same. For illustration, three changes of equal percentage but different amounts are shown in the graph.

From Year	To Year	Change in Production	
		Amount *£000*	Percent
1	2	200	100
3	4	100	100
6	7	400	100

Method of Drawing

You will probably be wondering how we draw these ratio scales; it is easy enough if we are dealing in round hundreds as in the above example, but how do we deal with the entire scale of numbers? There are two answers to this question:

(a) It is possible to buy specially printed graph paper on which the scales are ratio scales instead of ordinary scales. It is usually called semi-logarithmic graph paper.

(b) Use ordinary graph paper, but use the **logarithms** of the numbers on the vertical scale instead of the numbers themselves.

As an example of this latter procedure, we will draw the time series of profits (used above) on ratio scales. First, let's compile a table showing the data. The graph is then drawn as in Figure 4.11, with the logarithms of profits on the vertical scale.

Logarithms of Monthly Profits

Month	**Profits** *(£)*	**Log** *(profits)*
J	150	2.1761
F	250	2.3979
M	225	2.3522
A	320	2.5051
M	275	2.4393
J	310	2.4914
J	290	2.4624
A	300	2.4771
S	400	2.6021
O	350	2.5441
N	480	2.6812
D	450	2.6532

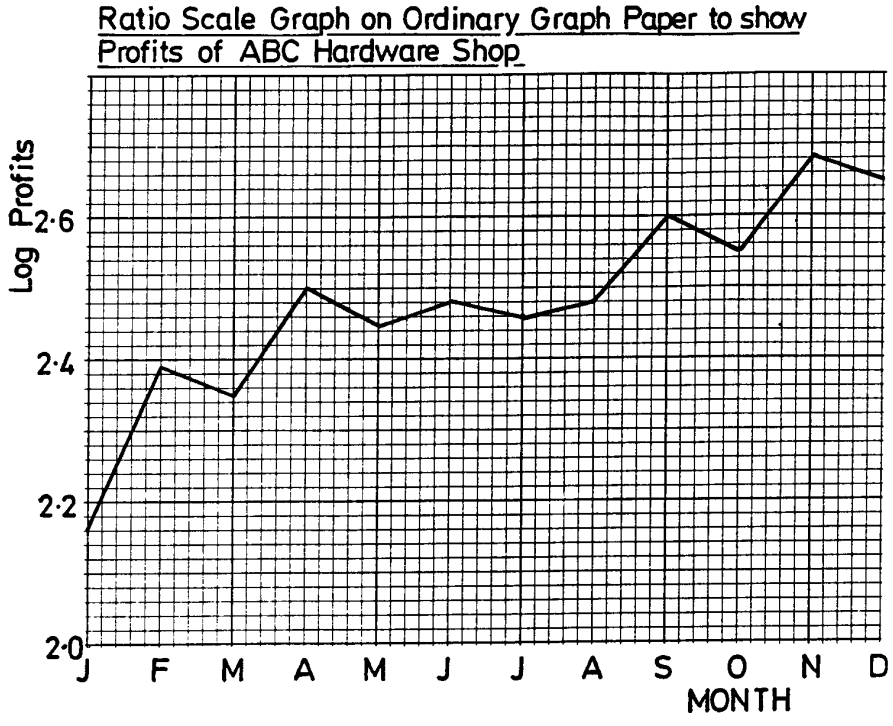


Figure 4.11

Figure 4.11 shows how a ratio scale graph can be drawn on ordinary graph paper using logarithms. If you use semi-logarithmic graph paper, you can plot the values direct without looking up the logarithms.

Whenever you want to see how actual values are changing, use ordinary scales; whenever you want to see percentage changes, use ratio scales.

There is no zero base line on the ratio scale graph because the log of zero is minus infinity, which is impossible to show. Similarly, negative values cannot be plotted. The horizontal axis is scaled in ordinary measure.

The most important feature of a ratio curve is not its position on the graph paper but the degree of slope of the curve. Two graphs with the **same slope** show the **same percentage rate of change**.

You can see that another benefit of a ratio scale is that you can cover a wide range of numbers easily on one graph. It is also straightforward to plot two time series of completely different types and units on the same graph and using the same scale. However, it would be inappropriate to use a ratio scale graph for analysing an aggregate into its constituents. A band chart on arithmetic scale paper is more suitable for this.

Study Unit 5

Measures of Location

Contents *Page*

A. INTRODUCTION

In a previous study unit we constructed frequency distributions from sets of raw data. At the time we noticed four features of these frequency distributions:

- All the available data is used in their construction.
- The result of the tabulation is to rearrange individual measurements according to their size instead of the order in which they were collected.
- The resulting tables can be illustrated graphically in several different ways.
- The type of variable which is measured affects the method of construction.

The two types of variables which you have to learn to recognise are:

(a) **Continuous variables**: these may take all values in a given range. In theory all measurements are continuous variables and in general discussion they are treated as such, e.g. time, length, weight.

(b) **Discrete variables**: may take only specified values, e.g. the number of children in families, shoe sizes, clothes sizes.

In this study unit we will discuss calculated functions of a set of data known as **measures of location** or **measures of central tendency**. These functions describe a set of data by giving the position of its "centre".

We shall be using the **sigma notation** throughout this section. $\sum$ is a Greek letter pronounced "sigma" and is used to denote the **summation** of a number of terms.

Thus, if we wish to add together $X_1 + X_2 + X_3 + X_4$, we could say that we wish to add together all the X_i's for i = 1 to 4 inclusive. This is written as:

$$\sum_{i=1}^{4} X_i$$

Similarly, $X_3 + X_4 + X_5 + X_6 + X_7$ would be written as $\sum_{i=3}^{7} X_i$

The two important rules you need to remember are:

(a) $$\sum_{i=1}^{N} aX = a \sum_{i=1}^{N} X_i$$

(b) $$\sum_{i=1}^{N} \frac{X_i}{n} = \frac{1}{n} \sum_{i=1}^{N} X_i$$

B. USE OF MEASURES OF LOCATION

The main measures of location are:

- Mean
- Median
- Mode

Let's first consider, in general terms, why we need these measures:

(a) Descriptive Use

The main purpose of a statistical analysis is to review unwieldy sets of data so that they may be understood and used in planning economic and business policies. A measure of location describes one feature of a set of data by a single number.

You have to discriminate between the various "centres" as each has its advantages and disadvantages. You must inspect any set of data carefully and choose the "centre" which is best for the problem you have to solve.

(b) Comparison of Distributions

Suppose you wish to compare the distribution of the weights of men and women in a given population. The data has been summarised to give two frequency distributions, and these distributions give the frequency curves shown in Figure 5.1.

The two curves overlap, showing, as you would expect, that some of the women are heavier than some of the men but, in general, the women are lighter than the men. The curves are the same shape and are symmetrical, so by symmetry, without any calculations, you can read off a value of x from each distribution which you could call the "centre". Since every other visible feature of the curves is the same, these two values of x describe their difference.

Weight Distribution of Men and Women

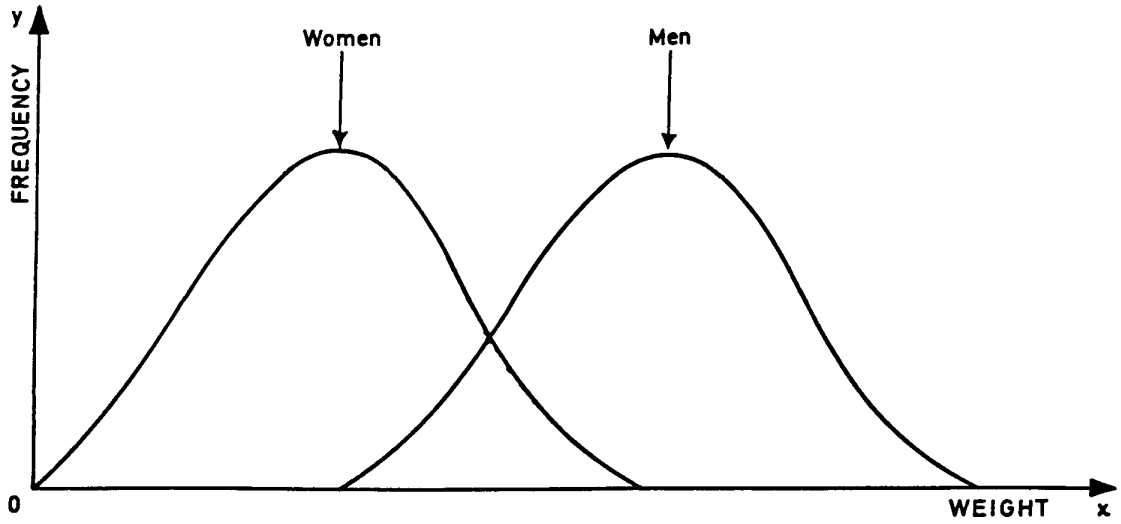


Figure 5.1

C. MEANS

Arithmetic Mean

The arithmetic mean of a set of observations is the **total sum of the observations divided by the number of observations**. This is the most commonly used measure of location and it is often referred to as "the mean".

Example 1

Find the mean monthly rainfall in Town A from the monthly observations given in Table 5.1:

Table 5.1: Monthly Rainfall in Town A

Month	**Rainfall** *(inches)*
Jan	5.4
Feb	6.8
Mar	7.2
Apr	6.5
May	5.2
June	4.2
July	2.1
Aug	2.8
Sept	3.9
Oct	4.5
Nov	4.8
Dec	5.3

We will use the sigma notation to work out this problem.

Remember that:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_n$$

Let x_I = rainfall in inches for the ith month

$\bar{x}$ = mean monthly rainfall

This symbol $\bar{x}$ (read as "x-bar") is commonly used to denote the mean of a set of observations.

Remembering that n = 12:

$$\text{Total rainfall} = \sum_{i=1}^{12} x_i$$

$$\bar{x} = \left(\sum_{i=1}^{12} x_i\right) \div 12 = \frac{58.7}{12} = 4.89 \text{ in}$$

From this example we can deduce the general formula for finding the mean of a set of n observations:

$$\bar{x} = \sum x \Big/ n = \frac{\sum x}{n} = \frac{1}{n}\sum x$$

You will notice that the formula is given in slightly different forms. They are the same, although the third one is most frequently used.

In a simple problem like this with only one set of data, the subscript l and the limits of summation may be omitted.

Example 2

Table 5.2 is the frequency distribution of the number of days on which 100 employees of a firm were late for work in a given month. Using this data, find the mean number of days on which an employee is late in a month.

Table 5.2: Number of Days Employees are Late in a Month

Number of Days Late *(x)*	**Number of Employees** *(f)*	**Number of Days** *(fx)*
1	32	32
2	25	50
3	18	54
4	14	56
5	11	55
Total	100	247

This is a little more difficult than Example 1. Begin by looking closely at the data; it is given in the form of a simple frequency distribution and the variable, x, is discrete and exact. It is almost certain that, in an exam question, you would only be given the first two columns without the symbols x and f. You should construct the full table and include it in your answer; then define the symbols you are using.

Let x_i = the ith value that x can take

f_i = the corresponding ith frequency

Then:

$$\text{Total number of days} = \sum_{i=1}^{5} f_i x_i = 247$$

$$\text{Total number of employees} = \sum_{i=1}^{5} f_i = 100$$

and $$\bar{x} = \sum_{i=1}^{5} f_i x_i \Big/ \sum_{i=1}^{5} f_i = \frac{247}{100} = 2.47$$

To deduce the general formula for a problem of this type, we let n be the number of values that the variable may take. Then:

$$\bar{\mathbf{x}} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{f_i x_i}}{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{f_i}} = \frac{\sum \mathbf{fx}}{\sum \mathbf{f}}$$

(The subscript i may be dropped if there is no possibility of confusion.)

Example 3

You are given a frequency distribution of heights of employees. As in Example 2, the data is grouped into a number of classes, but the variable, x, is a measurement, so it is continuous and it is quite likely that all the 80 values are different. The frequency distribution only tells us the class boundaries, e.g. we know that 7 observations lie between 160 and 165 cm but we do not know any of their values. To calculate the mean we must have a single value of x which is typical of all the values in each class. We choose the **midpoint** of each class as this typical value and assume that each observation in the class is equal to the midpoint value.

Then the value of $\bar{x}$ is calculated in exactly the same way as in Example 2. As long as we realise that $\bar{x}$ is an appropriate value only, this result is accurate enough for statistical analysis.

The first step in the solution of this problem is to construct a table as follows:

Table 5.3: Heights of Employees in cm

Class Boundaries *(cm)*	**Class Midpoints** *(x)*	**Frequency** *(f)*	*(fx)*
160 – under 165	162.5	7	1,137.5
165 – under 170	167.5	11	1,842.5
170 – under 175	172.5	17	2,932.5
175 – under 180	177.5	20	3,550.0
180 – under 185	182.5	16	2,920.0
185 – under 190	187.5	9	1,687.5
		80	14,070.0

Let x_i = midpoint of the ith class

f_i = frequency of the ith class

n = 6 classes

Then $\bar{x} = \sum_{i=1}^{6} f_i x_i \Big/ \sum_{i=1}^{6} f_i$

$$= \frac{14{,}070.0}{80} = 175.875$$

The general formula for calculating the mean of a frequency distribution of a continuous variable is then:

$$\bar{x} = \frac{\sum fx}{\sum f}$$

This formula looks exactly the same as the previous one. It is most important that you recognise the difference between them. It lies in the definition of x_i as the midpoint of the ith class instead of the ith value of the variable.

Note that although the above formula is commonly used in association with continuous variables, it is possible for a set of discrete data to have so many values of x that several of these values are grouped together, e.g. if there are 50 values they may be divided into 10 classes with 5 values in each. Then x_i is defined as the middle value in each class and the formula gives the approximate value of $\bar{x}$.

If you are given a frequency distribution with classes of different widths, make sure that you use the correct class midpoints, e.g. if the first classes above had been combined to give boundaries of 160-under 170, the class midpoint would have been 165.

Using an Assumed Mean

These three examples cover all the types of problems on arithmetic means that you will have to solve. You will see that if x, f and n are large, the amount of arithmetic required can be extensive. Even though you may use a pocket calculator, it is easy to make errors. The best way of avoiding errors is to use small numbers. We can use a short-cut method which reduces the amount of arithmetic and also saves time because you have less writing to do.

We will illustrate this method by a simple example and then apply it to Example 3 above so that you can see the reduction in size of numbers.

Table 5.4 shows the number of shares owned by seven individuals. First examine column 2 of the table and try to make a reasonable guess at the value of the mean number of shares owned; 110 appears to be a suitable value. This is called x_0, the **assumed or working mean**.

Table 5.4: Number of Shares Owned

Person	**Number of Shares** *(x)*	**$(x - x_0)$** *where $x_0 = 110$*
A	75	−35
B	125	+15
C	130	+20
D	80	−30
E	120	+10
F	50	−60
G	120	+10
Total	700	−125 +55 = −70

Assumed mean $x_0 = 110$

The third column is the difference between the actual number of shares owned by each person and the assumed mean. You can see that if your guess is close to the true mean the total of this column will be small.

The formula to use is $\bar{x} = x_0 + \dfrac{\sum(x - x_0)}{n}$

In this example, n = 7, x_0= 110 and $\sum(x - x_0) = -70$. So our guess was too large.

Therefore:

$$\bar{x} = 110 - \frac{70}{7} = 110 - 10 = 100$$

Now check this result by adding column 2, giving $\bar{x} = \frac{700}{7} = 100$

If you had chosen $x_0 = 100$ then $(x - x_0)$ and the positive and negative values in column 3 would cancel each other out. Try out several other values of x_0 to satisfy yourself that $\bar{x}$ always works out to the same value. You will see that the nearer your guess is to the true value, the smaller are the numbers in column 3.

Now we will use Example 3 to show how the method simplifies the arithmetic when it is applied to a grouped frequency distribution:

Table 5.5: Heights of Employees in cm

Heights: Class Boundaries *(cm)*	**Midpoints** *x*	**Frequency** *f*	**$x-x_0$** *where $x_0 = 177.5$*	**$f(x-x_0)$**
160 – under 165	162.5	7	-15	–105
165 – under 170	167.5	11	-10	–110
170 – under 175	172.5	17	-5	–85
175 – under 180	177.5	20	0	0
180 – under 185	182.5	16	5	80
185 – under 190	187.5	9	10	90
Totals		80		–130

The first three columns of Table 5.5 are the same as the first three of Table 5.3. To choose the best value of x_0 you should use the following rules:

(a) Let $x_0 = x_i$, then one of the values in columns 4 and 5 will be zero.

(b) Let i = the number of the class with the **largest frequency**, because it is reasonable to expect that the mean will lie in or near the class which contains the greatest number of observations.

(You may find that the arithmetic is easier if you break rule (b) when the largest frequency is at the top or bottom of the table.)

So, applying the rules to Table 5.5, i = 4 giving $x_0 = 177.5$

Now fill in the values in columns 4 and 5, e.g. for the first row, $x - x_0 = -15$ and $f = 7$ so the total differences in the first class are $f(x - x_0) = 7(-15) = -105$.

The formula to use is $\bar{x} = x_0 + \frac{\sum f(x - x_0)}{\sum f}$

$$\bar{x} = 177.5 - \frac{130}{80}$$

$$= 177.5 - 1.625 = 175.875$$

This is the same value as before.

For practice choose other values of i and satisfy yourself that the value of $\bar{x}$ is always the same but that the arithmetic is not as easy as when i = 4.

In an examination question you may be asked to compare the different measures of location, so we will look at the advantages and disadvantages of the arithmetic mean.

Advantages and Disadvantages of Arithmetic Mean

(a) Advantages

(i) It is easy to calculate as the only information you need is the sum of all the observations and the number of observations.

(ii) It is a well known statistic and it is easily manipulated to calculate other useful statistical measures.

(iii) It uses the values of all the observations.

(b) Disadvantages

(i) A few extreme values can cause distortion which makes it unrepresentative of the data set.

(ii) When the data is discrete it may produce a value which appears to be unrealistic, e.g. in Example 2, the mean number of days on which an employee is late is 2.47.

(iii) It cannot be read from a graph.

Weighted Mean

A firm owns six factories at which the basic weekly wages are given in column 2 of Table 5.6. Find the mean basic wage earned by employees of the firm.

Table 5.6: Basic Weekly Wage at Factories

Factory	Basic Weekly Wage £(x)	Number of Employees (w)	(wx)
A	85	50	4,250
B	105	80	8,400
C	64	40	2,560
D	72	35	2,520
E	96	90	8,640
F	112	75	8,400
Total	534	370	34,770

If you have no further information than column 2 then:

$$\bar{x} = \frac{1}{n}\sum x = \frac{£534}{6} = £89$$

But suppose you also know the number of employees at each factory (column 3), then:

$$\bar{x} = \frac{\text{Total wage bill}}{\text{Number of employees}} = \frac{\sum wx}{\sum w} = \frac{£34{,}770}{370} = £93.97$$

This second result, which takes account of the number of employees, is a much more realistic measure of location for the distribution of the basic wage than the straight mean we found first. The second result is called the **weighted mean** of the basic wage, where the weights are the numbers of employees at each factory.

In Example 2 earlier, the mean calculated could be called the weighted mean of the number of days, where the weights are the frequencies.

The advantage of using a weighted mean is that the weights can be chosen to vary the effect of some values of the variable on the measure of location.

Geometric Mean

The geometric mean is seldom used. It is applied to data for which the ratio of any two consecutive numbers is constant. It is evaluated by taking the nth root of the product of all n observations, i.e.:

$$\mathbf{G = \sqrt[n]{x_1\, x_2\, x_3 \ldots x_n}}$$

Example

The table shows the ratios of increase in pay received by Southampton University and College faculty members for each of several academic years:

Year	Ratio
1	1.061
2	1.085
3	1.105
4	1.052

Geometric mean $= \sqrt[4]{1.061 \times 1.085 \times 1.105 \times 1.052} = 1.075$

Harmonic Mean

Another measure of central tendency which is only occasionally used is the harmonic mean. It is most frequently employed for averaging speeds where the **distances** for each section of the journey are equal.

If the speeds are X_i then:

$$\textbf{Harmonic Mean} = \frac{N}{\sum_{i=1}^{N} \frac{1}{X_i}}$$

Example

An aeroplane travels a distance of 900 miles. If it covers the first third and the last third of the trip at a speed of 250 mph and the middle third at a speed of 300 mph, find the average speed.

$$\text{Average speed} = \frac{3}{\frac{1}{250} + \frac{1}{250} + \frac{1}{300}} = \frac{3}{0.01133} = 264.7$$

Note: If the average speed is to be calculated where the **times** for each section of the journey are the same, the appropriate average is the arithmetic mean.

The geometric mean and harmonic mean are of little practical interest. Both use all the observations, but they are awkward to calculate and cannot be illustrated graphically.

D. MEDIAN

Definition

If a set of n observations is arranged in order of size then, if n is **odd**, the median is the **value of the middle observation**; if n is **even**, the median is the **value of the arithmetic mean of the two middle observations.**

Note that the same value is obtained whether the set is arranged in ascending or descending order of size, though the ascending order is most commonly used. This arrangement in order of size is often called **ranking**.

The rules for calculating the median are:

(a) If n is odd and M is the value of the median then:

M = the value of the $\frac{n+1}{2}$th observation

(b) If n is even, the middle observations are the $\frac{n}{2}$th and the $\left[\frac{n}{2}+1\right]$th observations and then:

M = the value of the mean of these two observations

To show how the median is calculated for data presented in different forms, we will use the data from the three examples used before.

Calculation of the Median

Example 1

When n is relatively small and all the individual observations are listed, begin by ranking the observations.

Arrange the monthly rainfall observations given in Table 5.1 in ascending order of size:

2.1, 2.8, 3.9, 4.2, 4.5, 4.8, 5.2, 5.3, 5.4, 6.5, 6.8, 7.2

n =12 (i.e. even), so $\frac{n}{2}$th observation is the 6th and $\left[\frac{n}{2}+1\right]$th observation is the 7th.

Therefore:

M = mean of 6th and 7th observations:

$$M = \frac{(4.8 + 5.2)}{2} = 5.0$$

Example 2

The data is given in the form of a simple frequency table so the values of the variable have already been arranged in ascending order of size; the smallest value x_1 occurs f_1 times and the general value x_i occurs f_i times.

Table 5.7: Number of Days Employees are Late in a Month

Number of Days Late *(x)*	**Number of Employees** *(f)*	**Cumulative Frequency**
1	32	32
2	25	57
3	18	75
4	14	89
5	11	100

The first two columns of Table 5.7 repeat the simple frequency distribution of Table 5.2. The third column is the cumulative frequency of the distribution. n = 100 so the median value is the mean of the 50th and 51st observations. From the table you can see at once that these values are both 2, so the median value is 2.

Example 3

The data is given in the form of a grouped frequency distribution. Since we do not know the actual values of any of the observations, we cannot order them. We can only say that all the observations in the first class are smaller than all the observations in the second class and so on. In order to calculate the median value we assume that the f_i observations in the ith class are equally spaced across the class. This means that the median value found is an approximation only.

Table 5.8: Height of Employees in cm

Heights: Class Boundaries ***(cm)***	**Frequency**	**Cumulative Frequency**	**Percentage Cumulative Frequency**
160 – under 165	7	7	8.75
165 – under 170	11	18	22.50
170 – under 175	17	35	43.75
175 – under 180	20	55	68.75
180 – under 185	16	71	88.75
185 – under 190	9	80	100.00
	80		

Since n = 80, the median value of this distribution is the mean of the 40th and 41st observation values, and you can see that both these observations lie in the 4th class. The value of the lower class boundary is 175 cm, the first observation in this class is the 36th, and the width of the class is 5 cm. So the 40th and 41st observations are the 5th and 6th in the class, which contains 20 observations.

Therefore, assuming that the observations are evenly spaced in the class:

$$\text{Value of 40th observation} = \left[175 + \frac{5}{20} \times 5\right] \text{cm}$$

$$\text{Value of 41st observation} = \left[175 + \frac{6}{20} \times 5\right] \text{cm}$$

$$\text{Median} = \frac{1}{2}(176.25 + 176.5)\ \text{cm} = 176.375\ \text{cm}$$

The values in the fourth column of Table 5.8 are the percentage cumulative frequencies and are found by expressing each of the cumulative frequencies as a percentage of the total frequencies (80).

Note that if the total frequency is 100, the cumulative and percentage cumulative frequencies are the same. Percentage cumulative frequencies are useful for comparing different sets of data and for calculating quantiles (see later).

The median can be found graphically by plotting the **ogive**. If the y axis is the percentage cumulative frequency axis, then the median can be read very easily by finding the value of x corresponding to the 50% value on the y axis, as shown in Figure 5.2. This gives M = 176 cm, which is very close to the calculated approximate value.

Advantages and Disadvantages of the Median

(a) Advantages

(i) Its value is not distorted by extreme values, open-ended classes or classes of irregular width.

(ii) All the observations are used to order the data even though only the middle one or two observations are used in the calculation.

(iii) It can be illustrated graphically in a very simple way.

(b) Disadvantages

(i) In a grouped frequency distribution the value of the median within the median class can only be an estimate, whether it is calculated or read from a graph.

(ii) Although the median is easy to calculate it is difficult to manipulate arithmetically. It is of little use in calculating other statistical measures.

Ogive of Heights

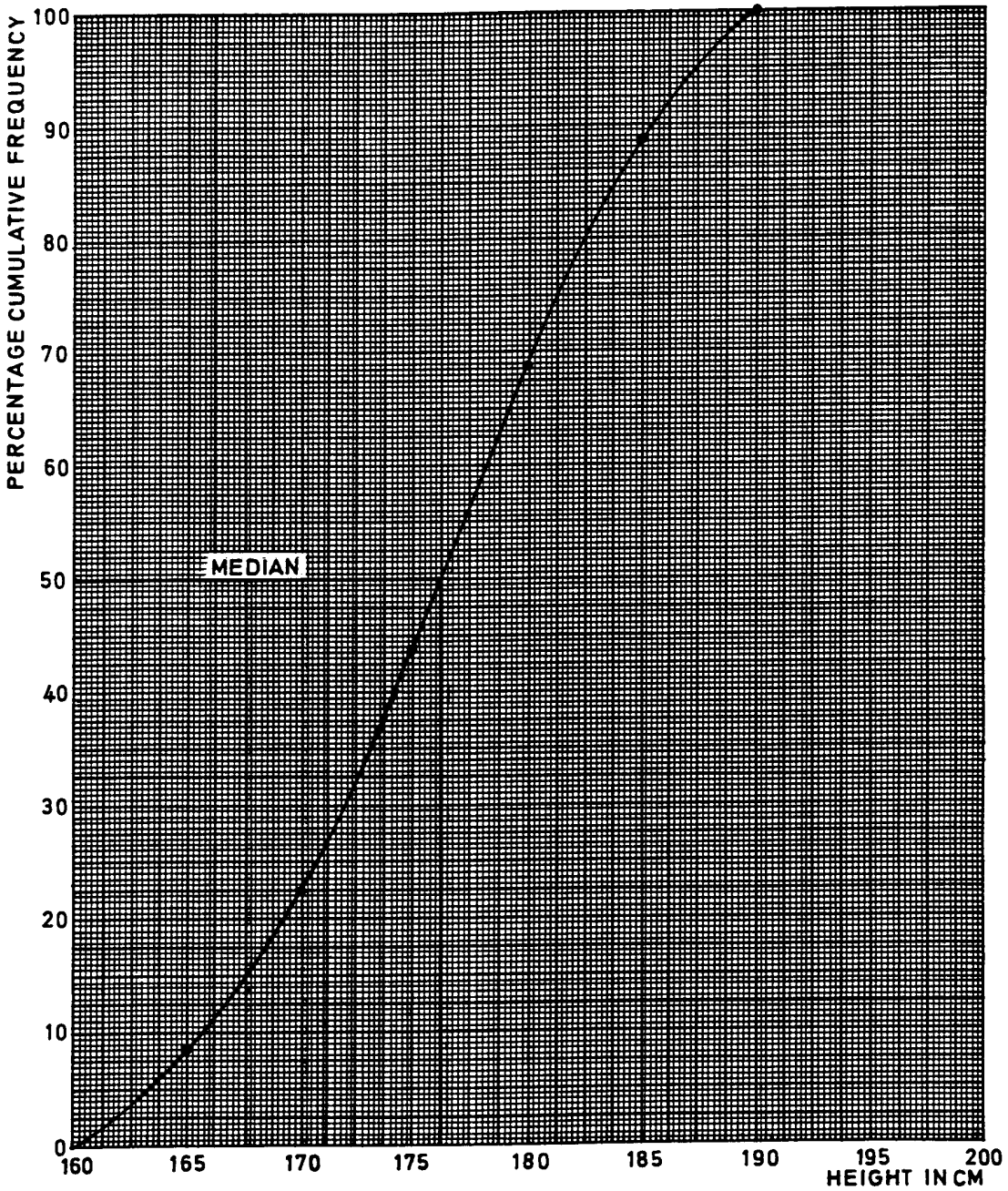


Figure 5.2

E. QUANTILES

Definitions

If a set of data is arranged in ascending order of size, quantiles are the values of the observations which divide the number of observations into a given number of **equal parts**.

They cannot really be called measures of central tendency, but they are measures of location in that they give the position of specified observations on the x axis.

The most commonly used quantiles are:

(a) Quartiles

These are denoted by the symbols Q_1, Q_2 and Q_3 and they divide the observations into four equal parts:

Q_1 has 25% below it and 75% above it.

Q_2 has 50% below it and 50% above, i.e. it is the **median** and is more usually denoted by M.

Q_3 has 75% below it and 25% above.

(b) Deciles

These values divide the observations into 10 equal parts and are denoted by D_1, D_2,...D_9, e.g. D_1 has 10% below it and 90% above, D_2 has 20% below it and 80% above, etc.

(c) Percentiles

These values divide the observations into 100 equal parts and are denoted by P_1, P_2, P_3, ...P_{99}, e.g. P_1 has 1% below it and 99% above.

Note that D_5 and P_{50} are both equal to the median (M).

Calculation of Quantiles

Example

Table 5.9 shows the grouped distribution of the overdraft sizes of 400 bank customers. Find the quartiles, the 4th decile and the 95th percentile of this distribution.

Table 5.9: Size of Overdraft of Bank Customers

Size (£)	Number of Customers	Cumulative Frequency	Percentage Cumulative Frequency
less than 100	82	82	20.5
100 but less than 200	122	204	51.0
200 but less than 300	86	290	72.5
300 but less than 400	54	344	86.0
400 but less than 500	40	384	96.0
500 but less than 600	16	400	100.0
	400		

The values of these quantiles may be approximated by reading from the ogive as shown in Figure 5.3.

The arithmetic calculations are:

Q_1 = average of the 100th and 101st customer

$$= £\left\{\left[100 + \frac{18}{122} \times 100\right] + \left[100 + \frac{19}{122} \times 100\right]\right\} \div 2$$

$$= £115.16$$

Q_2 = average of the 200th and 201st customer

$$= £\left\{\left[100 + \frac{118}{122} \times 100\right] + \left[100 + \frac{119}{122} \times 100\right]\right\} \div 2$$

$$= £197.13$$

Q_3 = average of the 300th and 301st customer

$$= £\left\{\left[300 + \frac{10}{54} \times 100\right] + \left[300 + \frac{11}{54} \times 100\right]\right\} \div 2$$

$$= £319.44$$

D_4 = average of the 160th and 161st customer

$$= £\left\{\left[100 + \frac{78}{122} \times 100\right] + \left[100 + \frac{79}{122} \times 100\right]\right\} \div 2$$

$$= £164.34$$

P_{95} = average of the 380th and 381st customer

$$= £\left\{\left[400 + \frac{36}{40} \times 100\right] + \left[400 + \frac{37}{40} \times 100\right]\right\} \div 2$$

$$= £491.21$$

Ogive of Size of Overdrafts

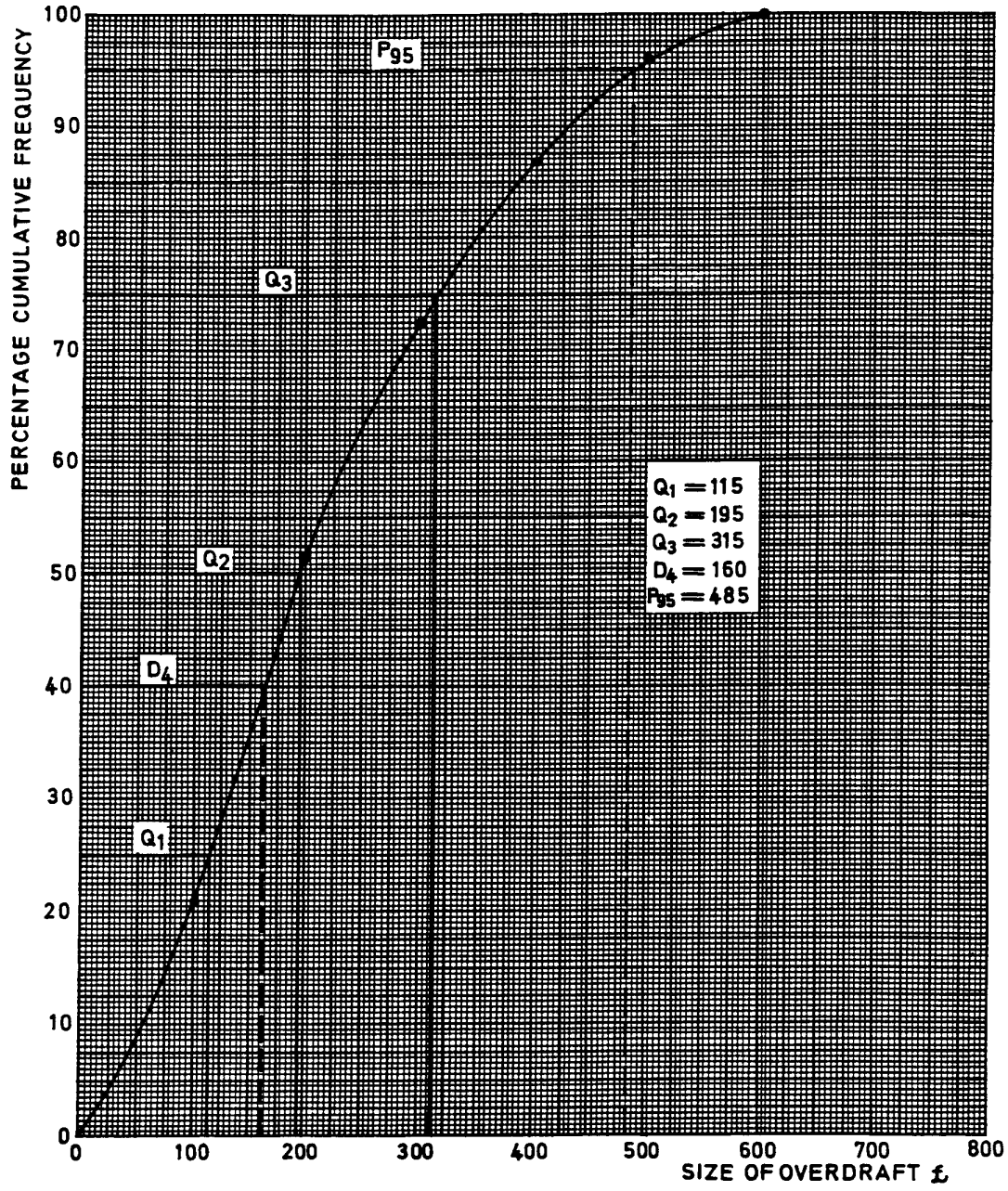


Figure 5.3

You can see from Figure 5.3 that the two methods give approximately the same results.

F. MODE

Definition

If the variable is **discrete**, the mode is that **value of the variable which occurs most frequently**. This value can be found by ordering the observations or inspecting the simple frequency distribution or its histogram.

If the variable is **continuous**, the mode is located in the **class interval with the largest frequency**, and its value must be estimated.

As it is possible for several values of the variable or several class intervals to have the same frequency, a set of data may have several modes.

- A set of observations with one mode is called **unimodal**.
- A set of observations with two modes is called **bimodal**.
- A set of observations with more than two modes is called **multimodal**.

Calculation of Mode

(a) Discrete Variable

Example 1

The following is an ordered list of the number of complaints received by a telephone supervisor per day over a period of a fortnight:

3, 4, 4, 5, 5, 6, 6, 6, 6, 7, 8, 9, 10, 12

The value which occurs most frequently is 6, therefore:

Mode = 6

Suppose one 6 is replaced by a 5, then 5 and 6 both occur three times and the data is bimodal, with modal values 5 and 6.

Example 2

For the dimple frequency distribution shown in table 5.2, the number of days late with the greatest frequency is 1. Therefore:

Mode = 1

(b) Continuous Variable

Example

Find the modal value of the height of employees from the data shown in Table 5.3.

The largest frequency is 20, in the fourth class, so this is the modal class, and the value of the mode lies between 175 and 180 cm.

There are various methods of estimating the modal value (including a graphical one). A satisfactory result is obtained easily by using the following formula:

$$\textbf{Mode} = \mathbf{L} + \frac{\mathbf{f_l}}{\mathbf{f_\ell} + \mathbf{f_u}} \times \mathbf{c}$$

where: L = lower boundary of the modal class

c = width of class interval

f_ℓ = frequency in class interval below modal class

f_u = frequency in class interval above modal class

So in this example

$$\text{Mode} = \left[175 + \frac{17}{17 + 16} \times 5\right] \text{cm}$$

$$= 177.6 \text{ cm to 1 dp}$$

Note: The formula assumes that all the class intervals are the same width. If they are **not**, you must adjust the frequencies to allow for this before you use the formula. If you cannot remember how to do this, look back at the section on frequency distributions and histograms in the previous study unit.

Advantages and Disadvantages of the Mode

(a) Advantages

(i) It is not distorted by extreme values of the observations.

(ii) It is easy to calculate.

(b) Disadvantages

(i) It cannot be used to calculate any further statistic.

(ii) It may have more than one value (although this feature helps to show the shape of the distribution).

G. CHOICE OF MEASURE

We have now discussed all the common measures of location. If you have a choice (particularly in an examination question), which measure should you choose? Before you decide, inspect the data carefully and consider the problem you have to solve.

- The arithmetic mean, although it involves the most arithmetic, is the most obvious choice. It is easy to understand, it uses all the data and it is needed in the calculation of a number of other statistics.
- However, the arithmetic mean is not perfect, if it is distorted by extreme values or uneven class intervals, the median or the mode may be better. If a large number of observations are concentrated at one end of the distribution, the median is likely to be the best measure. If there are concentrations of observations at intervals over the distribution, it may be multimodal and so the mode will give a more realistic picture.
- The mode will be a more useful measure, for example, for a manufacturer of dresses, than the mean, i.e. he would rather know that the greatest demand is for size 14 than that the mean size is 15.12.
- You should **not** choose the geometric or harmonic means as they are difficult to calculate and of little practical interest.
- You use the weighted mean whenever you calculate the arithmetic mean from a frequency distribution.
- Apart from the median, you will use other quantiles only when the question specifically asks for them.

If you are asked to comment on a result, you can often give a better answer if you compare several measures. Remember that all these measures of location replace the details of the raw data by one

number, which summarises the details to tell you **one** characteristic of the distribution. To get a good summary of the data you will need other statistics as well.

Study Unit 6

Measures of Dispersion

Contents ***Page***

A. INTRODUCTION

In the previous study unit we defined ways of describing distributions by calculating measures of location. However, one value of a measure of location can represent several sets of data; the distributions shown in Figure 6.1 all have the same "centre" but their shapes are quite different. So we need another number to show this difference, and this number is called a measure of dispersion. This describes the way in which the observations are **spread** about the "centre", i.e. it is a measure of the variability of the data. A summary of any set of data is not complete unless **both** a measure of location and a measure of dispersion are given.

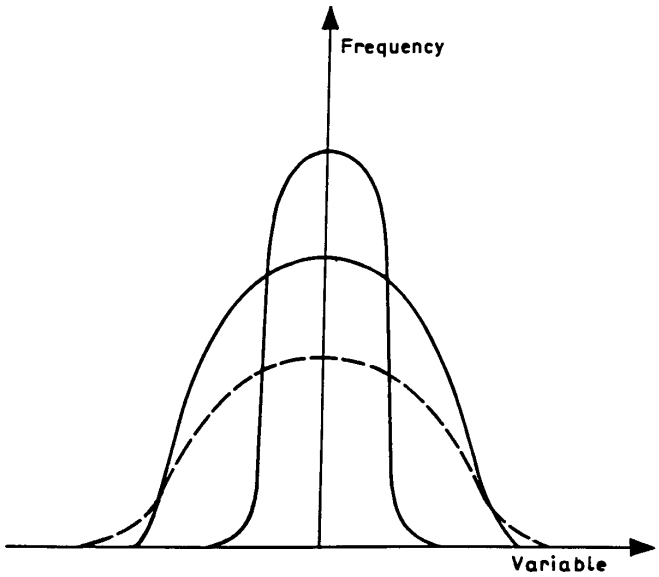


Figure 6.1

If the observations in a set of data are very variable, some of them will be a long way from the "centre", the curve will be wide and flat, and the value of the measure of dispersion will be large. If the observations are not very variable, they will all be close to the "centre", the curve will be narrow and tall, and the measure of dispersion will be small.

You can see from Figure 6.1 that, in these distributions, the part of the curve to the right of the centre is a mirror image of the part to the left. Distributions which have this characteristic are said to be **symmetrical**. You can see without any calculation that if a distribution is symmetrical, the value of its mean, median and mode will be equal. If a distribution does not have this characteristic, it is said to be **skewed** or **asymmetrical**; and the mean, median and mode will all have different values.

There are several different measures of dispersion. The most important of these (which we will describe in this study unit) are:

- Range
- Quartile deviation
- Mean deviation
- Standard deviation and variance

Two further measures, the coefficient of variation and skewness, will also be discussed.

B. RANGE

Definition and Calculation

The range of a distribution is the difference between the values of the largest and the smallest observations in the set of data.

Example

Look back to the data given in Table 5.1 in the last study unit.

The largest monthly rainfall for Town A is 7.2 inches, in March, and the smallest is 2.1 inches, in July. Therefore:

Range = (7.2 – 2.1) inches = 5.1 inches

Table 6.1 gives the monthly rainfall in another town:

Table 6.1: Monthly Rainfall in Town B

Month	Rainfall in Inches
Jan	6.2
Feb	6.6
Mar	10.6
Apr	5.1
May	4.3
June	2.2
July	0.4
Aug	2.8
Sept	3.4
Oct	4.8
Nov	5.4
Dec	6.9

The largest monthly rainfall for Town B is 10.6 inches, in March, and the smallest is 0.4 inches, in July. Therefore:

Range = (10.6 – 0.4) inches = 10.2. inches

i.e. the range for Town B is double that for Town A.

However, if you calculate the mean rainfall for Town B, you find that

$$\bar{x} = \sum x \big/ 12 = 58.7 \big/ 12 = 4.89 \text{ inches},$$

which is exactly the same as for Town A.

If only the means are known, you would say that the rainfall distributions for these two towns are identical, but when the range is known you can see that they are different.

- If the data is given in the form of a **simple frequency** distribution, the range is the difference between the **largest** and **smallest possible values of the variable**.

- If the data is given in the form of a grouped frequency distribution, the range is the difference between the **highest upper class boundary** and the **lowest lower class boundary**.

Advantages and Disadvantages

(a) Advantages

(i) It is easy to understand.

(ii) It is simple to calculate.

(iii) It is a good measure for comparison as it spans the whole distribution.

(b) Disadvantages

(i) It uses only two of the observations and so can be distorted by extreme values.

(ii) It does not indicate any concentrations of the observations.

(iii) It cannot be used in calculating other functions of the observations.

C. QUARTILE DEVIATION

Definition and Calculation

The quartile deviation is half the difference between the third quartile and the first quartile, and for this reason it is often called the **semi-interquartile range** (SIQR).

$$\textbf{Quartile deviation} = \tfrac{1}{2}(Q_3 - Q_1)$$

To calculate the quartile deviation, you must first order the set of observations, so this measure of dispersion is related to the median.

Example 1

Find the quartile deviation of the monthly rainfall for the two towns whose rainfall is given in Tables 5.1 and 6.1. The ordered amounts of rainfall are:

Town A: 2.1, 2.8, 3.9, 4.2, 4.5, 4.8, 5.2, 5.3, 5.4, 6.5, 6.8, 7.2

Town B: 0.4, 2.2, 2.8, 3.4, 4.3, 4.8, 5.1, 5.4, 6.2, 6.6, 6.9, 10.6

Since n (=12) is divisible by 4, none of the quartiles is equal to an observation. Q_1 is halfway between the 3rd and 4th observations, Q_2 is halfway between the 6th and 7th observations and Q_3 is halfway between the 9th and 10th observations.

For Town A:

$Q_1 = \tfrac{1}{2}(3.9 + 4.2)$ $\quad Q_3 = \tfrac{1}{2}(5.4 + 6.5)$

$= 4.05$ $\quad = 5.95$

Quartile deviation $= \tfrac{1}{2}(5.95 - 4.05)$ inches

$= 0.95$ inches

For Town B:

$Q_1 = \tfrac{1}{2}(2.8 + 3.4)$ $\quad Q_3 = \tfrac{1}{2}(6.2 + 6.6)$

$= 3.1 \qquad = 6.4$

Quartile deviation $= \frac{1}{2}(6.4 - 3.1)$ inches

$= 1.65$ inches

It is interesting to compare the medians.

We have already calculated that for Town A, $M = Q_2 = 5.0$ inches. For Town B:

$M = Q_2 = \frac{1}{2}(4.8 + 5.1) = 4.95$ inches.

These two values are very close but not identical, while the quartile deviation of Town B is still much larger, but not quite double, that of Town A.

Example 2

Find the quartile deviation of the number of days late, from the simple frequency distribution given in Table 5.7 in the previous study unit.

Here n = 100, so Q_1 is halfway between the 25th and 26th observations; both of these are 1 so $Q_1 = 1$.

Q_3 is halfway between the 75th and 76th observations; the 75th observation is 3 and the 76th is 4 so $Q_3 = 3.5$.

Therefore:

Quartile deviation $= \frac{1}{2}(3.5 - 1) = 1.25$ days.

Example 3

Find the quartile deviation for the size of overdrafts using the quantiles calculated in the previous study unit from Table 5.9.

$Q_1 = £115.16 \qquad Q_3 = £319.44$

Quartile deviation $= £\frac{1}{2}(319.44 - 115.16) = £102.14$

Note that this method must be used because you are asked to **calculate** the value. Otherwise the graphical method for finding Q_1 and Q_3 would be acceptable, though not quite as accurate.

Advantages and Disadvantages

(a) Advantages

(i) The calculations are simple and quite quick to do.

(ii) It covers the central 50% of the observations and so is not distorted by extreme values.

(iii) It can be illustrated graphically.

(b) Disadvantages

(i) The lower and upper 25% of the observations are not used in the calculation so it may not be representative of all the data.

(ii) Although it is related to the median, there is no direct arithmetic connection between the two.

(iii) It cannot be used to calculate any other functions of the data.

D. MEAN DEVIATION

Definition and Calculation

The mean deviation (MD) is the arithmetic mean of the absolute differences between the observations and their arithmetic mean. Written in symbols, this is:

$$MD = \frac{\sum |x_i - \bar{x}|}{n}$$

The term "absolute difference" means that the difference is always taken as positive.

The symbol $|x_i - \bar{x}|$ is read as "the absolute difference between x_i and $\bar{x}$" or "mod $(x_i - \bar{x})$".

Example 1

Find the mean deviation of the monthly rainfall for the Towns A and B using the data from section C Example 1. We know that $\bar{x} = 4.89$, correct to 2 dp, for both distributions.

Table 6.2 shows the calculations for this problem. When you are asked to work out a mean deviation, columns 3 and 5 in the table you construct would be headed "$|x_i - \bar{x}|$" and all the differences would be positive. In this example these differences have been given their correct signs so that you can see why the absolute values must be used. If they are not, the sum of the two columns 3 and 5 would be zero and so would give a zero measure of dispersion.

Table 6.2: Monthly Rainfall in Towns A and B

Month	**Rainfall Town A** *(inches)*	**Deviations** $(x_i - \bar{x})$	**Rainfall Town B** *(inches)*	**Deviations** $(x_i - \bar{x})$
Jan	5.4	+0.51	6.2	+1.31
Feb	6.8	+1.91	6.6	+1.71
Mar	7.2	+2.31	10.6	+5.71
Apr	6.5	+1.61	5.1	+0.21
May	5.2	+0.31	4.3	–0.59
June	4.2	–0.69	2.2	–2.69
July	2.1	–2.79	0.4	–4.49
Aug	2.8	–2.09	2.8	–2.09
Sep	3.9	–0.99	3.4	–1.49
Oct	4.5	–0.39	4.8	–0.09
Nov	4.8	–0.09	5.4	+0.51
Dec	5.3	+0.41	6.9	+2.01
		+7.06		+11.46
		–7.04		–11.44

(Theoretically, the minus and plus totals of the deviations should be equal in size, but as $\bar{x}$ is rounded off to 2 dp there is a small rounding error.)

For Town A:

$$MD = \frac{\sum |x_i - \bar{x}|}{n}$$

$$= \frac{14.1}{12} \text{ (since total absolute deviation is (7.06 + 7.04))}$$

$$= 1.18 \text{ in}$$

For Town B:

$$MD = \frac{\sum |x_i - \bar{x}|}{n}$$

$$= \frac{22.9}{12} \quad \left[\text{i.e. } \frac{11.46 + 11.44}{12}\right]$$

$$= 1.91 \text{ in}$$

The formula for calculating the mean deviation for a grouped frequency distribution is:

$$\mathbf{MD = \frac{\sum |f(x - \bar{x})|}{f}}$$

Example 2

Table 6.3 shows the calculations for finding the mean deviation of the heights of employees using the grouped frequency distribution given in the last study unit.

Table 6.3: Heights of Employees

Height *(cm)*	**Midpoint** *(x)*	**Frequency** *(f)*	$\lvert x - \bar{x} \rvert$	$f\lvert x - \bar{x} \rvert$
160 – under 165	162.5	7	13.375	93.625
165 – under 170	167.5	11	8.375	92.125
170 – under 175	172.5	17	3.375	57.375
175 – under 180	177.5	20	1.625	32.500
180 – under 185	182.5	16	6.625	106.000
185 – under 190	187.5	9	11.625	104.625
Total		80		486.250

$$MD = \frac{\sum |f(x - \bar{x})|}{\sum f} \text{ given } \bar{x} = 175.875 \text{ and } \sum f = 80$$

$$= \frac{486.25}{80} = 6.08 \text{ cm}$$

Advantages and Disadvantages

(a) Advantages

(i) It uses all the observations in the set of data.

(ii) It is directly related to the mean, i.e. a measure of location.

(iii) It is useful for simple comparison between sets of data.

(b) Disadvantages

(i) Absolute values are difficult to manipulate arithmetically.

(ii) It is not used in any further statistical analysis.

E. STANDARD DEVIATION AND VARIANCE

These two measures of dispersion can be discussed in the same section because the standard deviation is the positive square root of the variance. So even if you are asked to find the standard deviation of a set of data, you will have to find the variance first.

The variance is of great theoretical importance in advanced statistical work but all you need to know about it is its definition, how to calculate it, and its relationship to the standard deviation.

Since the variance is a function of the squares of the observations, the unit in which it is measured is the square of the unit in which the observations are measured. So its square root, i.e. the standard deviation, is measured in the **same** unit as the observations.

Definition and Calculation

We will define the standard deviation first as you are more likely to be asked for this in your examination than the variance.

The standard deviation is the **positive square root of the mean of the squares of the differences between all the observations and their mean.** You will find this definition quite easy to understand when you see it written in symbols.

Let σ (the small Greek letter "sigma") be the standard deviation. You will sometimes find "s" or "sd" used.

The formula used to calculate the standard deviation is:

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

Formula (a)

where: x_i = value of the observation

$\bar{x}$ = mean of the observations

n = number of observations

You will remember that some of the differences $(x_i - \bar{x})$ will be positive and some will be negative, so that:

$$\sum |x_i - \bar{x}| = 0$$

In section D on the mean deviation we dealt with this problem by using the absolute value of $(x_i - \bar{x})$. Here we deal with it by adding the squares of $(x_i - \bar{x})$ and using the **positive** square root of this sum.

As: Standard deviation = $\sqrt{\text{Variance}}$

σ^2 = Variance

Example 1

Using the data of Table 6.4 find the standard deviation of the monthly rainfall of Town A.

Table 6.4: Monthly Rainfall for Table A

Month	Rainfall in Inches	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
Jan	5.4	+0.51	0.2601
Feb	6.8	+1.91	3.6481
Mar	7.2	+2.31	5.3361
Apr	6.5	+1.61	2.5921
May	5.2	+0.31	0.0961
June	4.2	–0.69	0.4761
July	2.1	–2.79	7.7841
Aug	2.8	–2.09	4.3681
Sep	3.9	–0.99	0.9801
Oct	4.5	–0.39	0.1521
Nov	4.8	–0.09	0.0081
Dec	5.3	+0.41	0.1681
			25.8692

Table 6.4 shows the calculation of $\sum (x_i - \bar{x})^2$, using x = 4.89, which we have already calculated and so can assume to be known, and n = 12.

Then, substituting in the formula (a) gives:

$$\sigma = \sqrt{\frac{25.8692}{12}} = \sqrt{2.16} = 1.47 \text{ in to 2dp}$$

Variance = σ^2 = $(1.47)^2$ = 2.16 in

By expanding the expression $\sum (x_i - \bar{x})^2$ we can rewrite the formula used so far (formula (a)) in two alternative and sometimes more useful forms.

Remembering that $\bar{x} = \frac{\sum x_i}{n}$, the alternative forms are:

$$\boxed{\sigma = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2}}$$ ***Formula (b)***

$$\sigma = \sqrt{\frac{\sum x^2}{n} - \left[\frac{\sum x}{n}\right]^2}$$ ***Formula (c)***

The choice of the formula to use depends on the information that you already have and the information that you are asked to give. In any calculation you should keep the arithmetic as simple as possible and the rounding errors as small as possible.

- If you already know $\bar{x}$ and it is a small integer, there is no reason why you should not use formula (a).
- If you already know $\bar{x}$ but it is not an integer, as in Example 1, then formula (b) is the best to use.
- If you do not know $\bar{x}$ then you should use formula (c), particularly if you are not asked to calculate $\bar{x}$.

When you are given the data in the form of a simple or grouped frequency distribution then the formula for calculating the standard deviation is:

$$\sigma = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i}}$$ ***Formula (d)***

$$\sigma = \sqrt{\frac{1}{n}\left[\sum f_i x_i^2 - \frac{\left(\sum f_i x_i\right)^2}{n}\right]}$$ ***Formula (e)***

where: $n = \sum f_i$

Formula (d), like formula (a), is derived directly from the definition of σ. Formula (e) is obtained by using the sigma notation as in formula (c) and this is the one to use in calculations.

Example 2

Using the data of Table 5.3, find the standard deviation of the heights of employees.

The first four columns of table 6.5 are exactly the same as those in Table 5.3. Column five is added so that we have all the sums required in formula (e). So from the table:

$$n = \sum f = 80; \quad \sum fx = 14{,}070.0; \quad \sum fx^2 = 2{,}478{,}750$$

Table 6.5: Heights of Employees in cm

Height *(cm)*	**Midpoint** *(x)*	**Frequency** *(f)*	**fx**	**fx^2**
160 – under 165	162.5	7	1,137.5	184,843.75
165 – under 170	167.5	11	1,842.5	308,618.75
170 – under 175	172.5	17	2,932.5	505,856.25
175 – under 180	177.5	20	3,550.0	630,125.00
180 – under 185	182.5	16	2,920.0	532,900.00
185 – under 190	187.5	9	1,687.5	316,406.25
Totals		80	14,070.0	2,478,750.00

Then: $\sigma = \sqrt{\frac{1}{80}\left[2{,}478{,}750 - \frac{(14{,}070)^2}{80}\right]}$

$$\sigma = \sqrt{\frac{1}{80}(2{,}478{,}750 - 2{,}474{,}561)}$$

$$\sigma = \sqrt{\frac{4189}{80}} = \sqrt{52.36} = 7.24$$

Short-cut Method of Calculation

As you can see, the arithmetic involved in calculating a standard deviation is extensive even with a small number of classes and a pocket calculator. So we use a development of the short-cut method explained in the previous study unit.

The short-cut method is easy to use and to remember if you break it down into a number of simple steps, as follows:

(a) Choose for the assumed or working mean x_0, the value of the midpoint of the class with the largest frequency.

(b) Work out the differences $(x_i - x_0)$.

(If you only need to find the mean, carry on exactly as in Study Unit 5, but if you need the variance and standard deviation carry on with the next steps instead.)

(c) Divide each of the differences by c, the width of the class interval, and call the result d, i.e.

$$d = \frac{x_i - x_0}{c}.$$

(d) Multiply each value of d by the frequency of the class it represents, add all these products and divide the result by $\sum f$. You now have the quantity:

$$\frac{\sum fd}{f} = \bar{d}$$

(e) Square each value of d, multiply it by the corresponding frequency, add all these products and divide the result by $\sum f$. You now have the quantity:

$$\frac{\sum fd^2}{\sum f}$$

(f) Work out $\frac{\sum fd^2}{\sum f} - \left[\frac{\sum fd}{\sum f}\right]^2$. This is the variance of d.

(g) Work out $\sqrt{\frac{\sum fd^2}{\sum f} - \left[\frac{\sum fd}{\sum f}\right]^2}$. This is the standard deviation of d.

(h) Find the variance of x by multiplying the variance of d by c^2:
i.e. if σ^2 is the variance of x, $\sigma^2 = c^2x$ (variance of d).

(j) Find the standard deviation of x by multiplying the standard deviation of d by c:
i.e. $\sigma^2 = cx$ (standard deviation of d).

Example 1

Now apply these steps to the data on employees' heights, giving Table 6.6:

Table 6.6: Heights of Employees

Midpoints x_i	**Frequency** f_i	$\mathbf{x_i - x_0}$	$\mathbf{d = \frac{x_i - x_0}{c}}$	**fd**	$\mathbf{d^2}$	$\mathbf{fd^2}$
162.5	7	–15	–3	–21	9	63
167.5	11	–10	–2	–22	4	44
172.5	17	–5	–1	–17	1	17
177.5	20	0	0	0	0	0
182.5	16	5	1	16	1	16
187.5	9	10	2	18	4	36
Total	80			–26		176

From the above, $x_0 = 177.5$, c = 5. Therefore:

$$\sum fd^2 = 176 \text{ and } d = \frac{x_i - 177.5}{5}$$

Substitute these values in the formula.

For the variance: $\sigma^2 = c^2\left[\frac{\sum fd^2}{\sum f} - \left[\frac{\sum fd}{f}\right]^2\right]$

$$= 25\left[\frac{176}{80} - \left[\frac{-26}{80}\right]^2\right]$$

$$= 25(2.2 - 0.106) = 52.35$$

For the standard deviation: $\sigma = \sqrt{52.35} = 7.24$

Example 2

Find the variance of the wages for the company wage distribution in table 6.7:

Table 6.7: Wages of Employees

Weekly Wage *(£)*	**Frequency** *f*	**Midpoint** *x*	$\mathbf{x_i - x_0}$	$\frac{\mathbf{x_i - x_0}}{\mathbf{c}} = \mathbf{d}$	**fd**	$\mathbf{d^2}$	$\mathbf{fd^2}$
under 20	20	10	−40	−2	−40	4	80
20 but under 40	45	30	−20	−1	−45	1	45
40 but under 60	65	50	0	0	0	0	0
60 but under 80	35	70	20	1	35	1	35
80 but under 100	25	90	40	2	50	4	100
100 and over	10	110	60	3	30	9	90
Total	200				+30		350

The last class of this distribution is open-ended so we have to make an assumption about its width in order to find its midpoint. The most convenient assumption is to make it the same width as the other classes, giving a midpoint of 110.

$$x_0 = 50,\ c = 20,\ d = \frac{x_i - 50}{20},\ \sum f = 200,\ \sum fd = 30,\ \sum fd^2 = 350$$

Substituting in the formula for the variance gives:

$$\sigma^2 = c^2\left[\frac{\sum fd^2}{\sum f} - \left[\frac{\sum fd}{\sum f}\right]^2\right]$$

$$= 20^2\left[\frac{350}{200} - \left(\frac{30}{200}\right)^2\right]$$

$$= 400(1.75 - 0.0225)$$

$$= 400 \times 1.7275$$

$$= 691$$

An Important Use of the Standard Deviation

The standard deviation is one of the measures used to describe the variability of a distribution. It has an additional use which makes it more important than the other measures of dispersion. It is used as a unit to measure the **distance between any two observations**. For example, in the distribution of employees' heights, the distance in the distribution between a height of 160 cm and a height of 189 cm is 29 cm. But as the standard deviation is 7.24 cm, we can also say that the distance between the two observations is (29 ÷7.24), or just over 4 standard deviations.

Advantages and Disadvantages of the Standard Deviation

(a) Advantages

(i) It uses all the observations.

(ii) It is closely related to the most commonly used measure of location, i.e. the mean.

(iii) It is easy to manipulate arithmetically.

(b) Disadvantages

(i) It is rather complicated to define and calculate.

(ii) Its value can be distorted by extreme values.

F. COEFFICIENT OF VARIATION

The standard deviation is an **absolute** measure of dispersion and is expressed in the units in which the observations are measured. The coefficient of variation is a **relative** measure of dispersion, i.e. it is independent of the units in which the standard deviation is expressed.

The coefficient of variation is calculated by expressing the standard deviation as a percentage of the mean:

$$\text{Coefficient of variation (CV)} = \frac{\sigma}{\bar{x}} \times 100\%$$

By using the coefficient of variation you can compare dispersions of various distributions, such as heights measured in cm with a distribution of weights measured in kg.

Example

Compare the dispersion of the monthly rainfall for Town A with the dispersion of employees' heights.

Town A: $\sigma = 1.47$ inches $\bar{x} = 4.89$ inches

$$\textit{therefore: } \text{CV} = \frac{1.47}{4.89} \times 100\% = 30.06\%$$

Employees' heights: $\sigma = 7.24$ cm $\bar{x} = 175.875$ cm

$$\textit{therefore: } \text{CV} = \frac{7.24}{175.875} \times 100\% = 4.12\%$$

This shows that the rainfall distribution is much more variable than the employees' heights distribution.

G. SKEWNESS

When the items in a distribution are dispersed equally on each side of the mean, we say that the distribution is **symmetrical**. Figure 6.2 shows two symmetrical distributions.

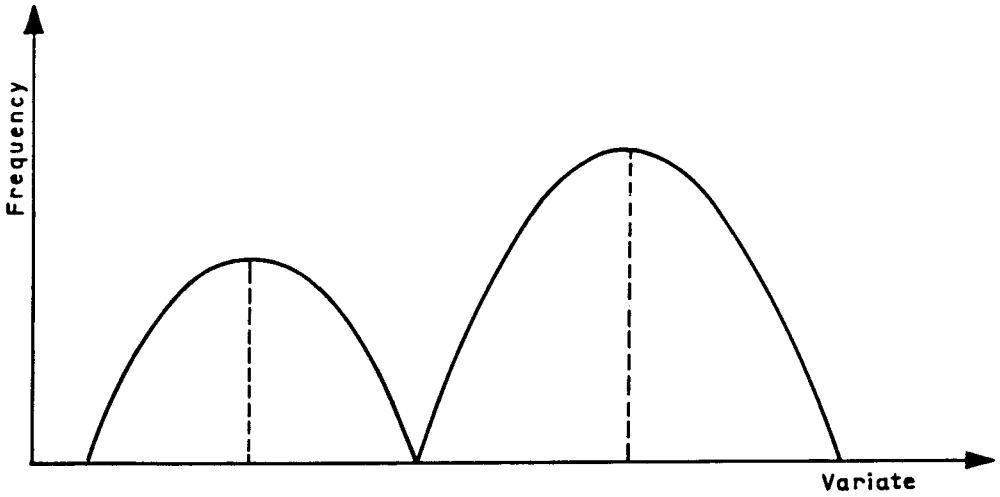


Figure 6.2

When the items are not symmetrically dispersed on each side of the mean, we say that the distribution is **skew** or asymmetric. Two skew distributions are shown in Figure 6.3. A distribution which has a tail drawn out to the right, as in Figure 6.3 (a), is said to be **positively skew**, while one with a tail to the left, like (b), is **negatively skew**.

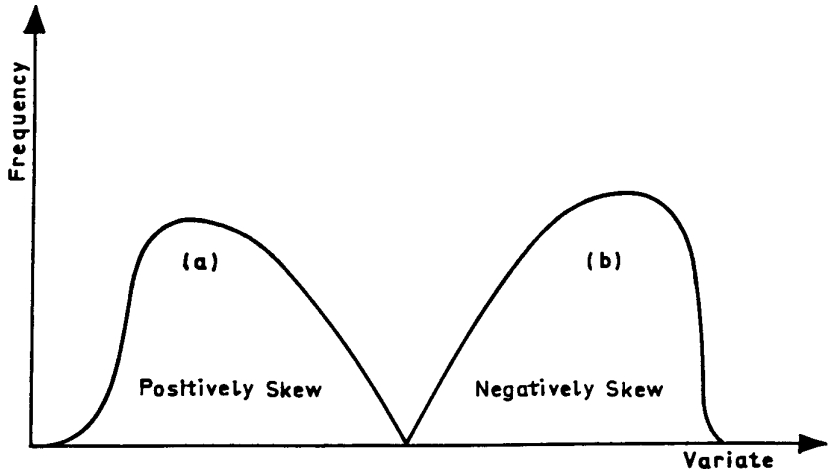


Figure 6.3

Two distributions may have the same mean and the same standard deviation but they may be differently skewed. This will be obvious if you look at one of the skew distributions and then look at the **same one** through from the other side of the paper! What, then, does skewness tell us? It tells us that we are to expect a few unusually high values in a positively skew distribution or a few unusually low values in a negatively skew distribution.

If a distribution is symmetrical, the mean, mode and the median all occur at the same point, i.e. right in the middle. But in a skew distribution the mean and the median lie somewhere along the side with the "tail", although the mode is still at the point where the curve is highest. The more skew the distribution, the greater the distance from the mode to the mean and the median, but these two are always in the same order; working outwards from the mode, the median comes first and then the mean, as in Figure 6.4:

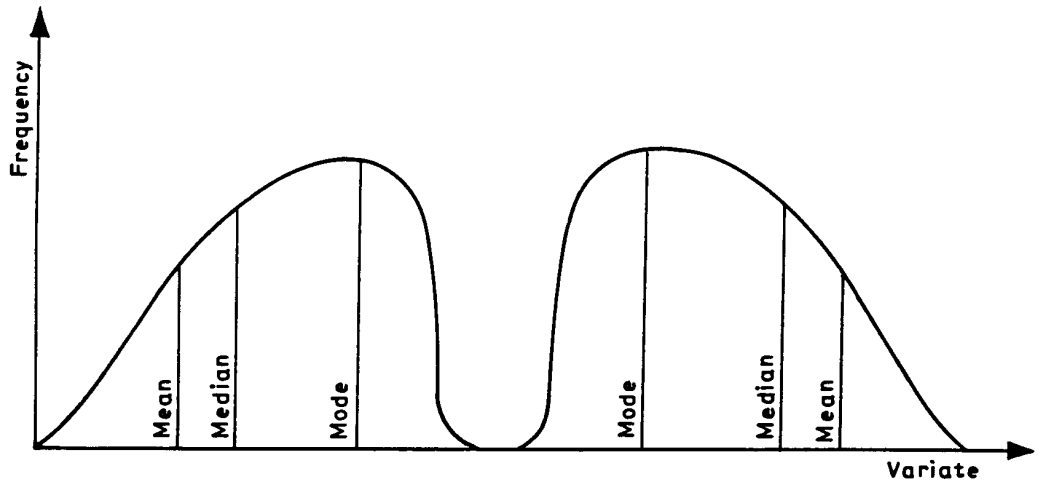


Figure 6.4

For most distributions, except for those with very long tails, the following relationship holds approximately:

Mean – Mode = 3 (Mean – Median)

The more skew the distribution, the more spread out are these three measures of location, and so we can use the amount of this spread to measure the amount of skewness. The most usual way of doing this is to calculate:

$$\text{Pearson's First Coefficient of Skewness} = \frac{\text{Mean - Mode}}{\text{Standard Deviation}}$$

However, the mode is not always easy to find and so we use the equivalent formula:

$$\text{Pearson's Second Coefficient of Skewness} = \frac{3(\text{Mean - Median})}{\text{Standard Deviation}}$$

(See the next study unit for more details of Pearson.)

You are expected to use one of these formulae when an examiner asks for the skewness (or coefficient of skewness) of a distribution. When you do the calculation, remember to get the correct sign (+ or –) when subtracting the mode or median from the mean and then you will get negative answers for negatively skew distributions, and positive answers for positively skew distributions. The value of the coefficient of skewness is between –3 and +3, although values below –1 and above +1 are rare and indicate very skew distributions.

Example of variates with positive skew distributions include size of incomes of a large group of workers, size of households, length of service in an organisation, and age of a workforce. Negative skew distributions occur less frequently. One such example is the age at death for the adult population of the UK.

H. SUMMARY

The definitions and formulae introduced in this study unit are very important in statistical analysis and interpretation and you are likely to get questions on them in your examination. You should learn all the formal definitions and the formulae thoroughly. Make sure you know when each of the formulae should be used, and that you can distinguish between those formulae which are only a statement of the definition in symbols and those which are used to calculate the measures.

Remember that even if you are allowed to use a pocket calculator, you must show all the steps in a calculation.

Examination questions will often ask you to comment on the results you have obtained. You will be able to make sensible comments if you have studied the use plus the advantages and disadvantages of each of the measures.

Study Unit 7

Correlation

Contents *Page*

A. INTRODUCTION

When studying frequency distributions, we were always handling only **one variable**, e.g. height or weight. Having learned how to solve problems involving only one variable, we should now discover how to solve problems involving **two variables** at the same time.

If we are comparing the weekly takings of two or more firms, we are dealing with only one variable, that of takings; if we are comparing the weekly profits of two or more firms, we are dealing with only one variable, that of profits. But if we are trying to assess, for one firm (or a group of firms), whether there is any relationship between takings and profits, then we are dealing with two variables, i.e. takings and profits.

B. SCATTER DIAGRAMS

Examples of Correlation

Suppose we have measured the height and weight of 6 men. The results might be as follows:

Man	**Height** *(in)*	**Weight** *(lb)*
A	66	150
B	72	159
C	65	138
D	69	145
E	64	128
F	70	165

A **scatter diagram** or scattergram is the name given to the method of representing these figures graphically. On the diagram, the horizontal scale represents one of the variables (let's say height) while the other (vertical) scale represents the other variable (weight). Each **pair** of measurements is represented by one point on the diagram, as shown in Figure 7.1:

Scattergram of Men's Heights and Weights

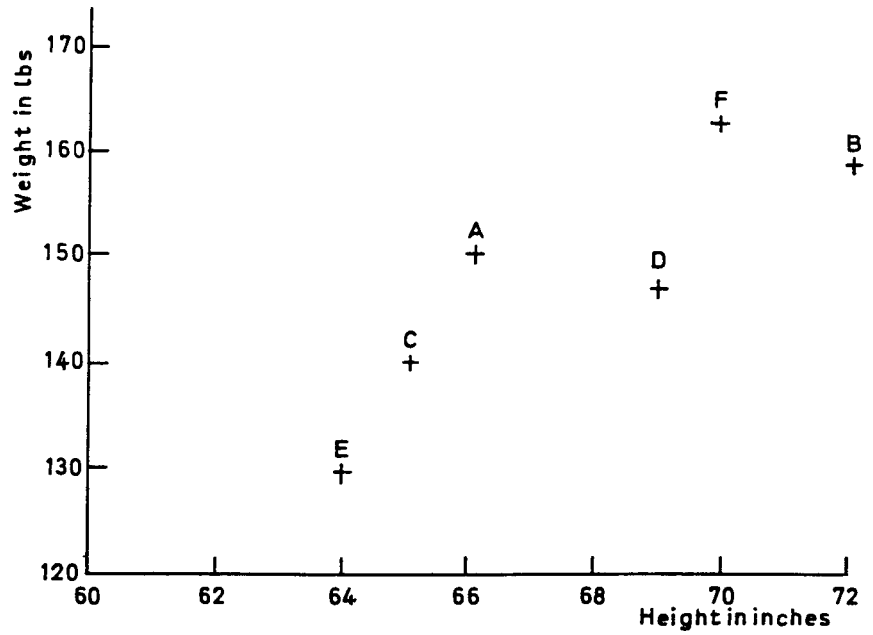


Figure 7.1

Make sure that you understand how to plot the points on a scatter diagram, noting especially that:

- Each point represents a **pair** of corresponding values.
- The two scales relate to the two variables under discussion.

The term scatter diagram or scattergram comes from the scattered appearance of the points on the chart.

Examining the scatter diagram of heights and weights, you can see that it shows up the fact that, by and large, tall men are heavier than short men. This shows that some relationship exists between men's heights and weights. We express this in statistical terms by saying that the two variables, height and weight are **correlated**. Figure 7.2 shows another example of a pair of correlated variables (each point represents one production batch):

Cost of Production Compared with Impurity Contents

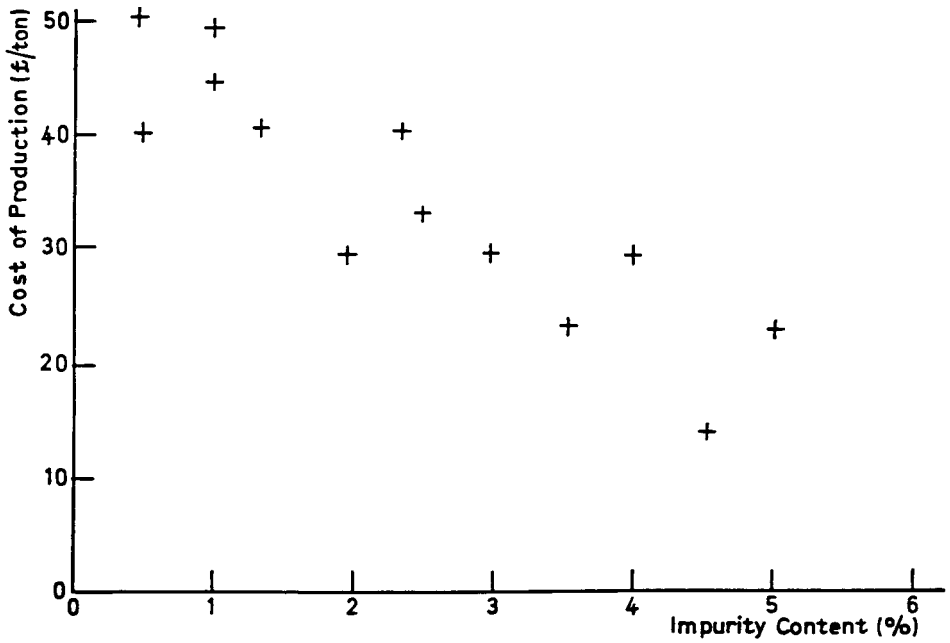


Figure 7.2

Here you see that, in general, it costs more to produce material with a low impurity content than it does to produce material with a high impurity content. However, you should note that correlation does not necessarily mean an **exact** relationship, for we know that, while tall men are usually heavy, there are exceptions, and it is most unlikely that several men of the same height will have exactly the same weight!

Degrees of Correlation

In order to generalise our discussion, and to avoid having to refer to particular examples such as height and weight or impurity and cost, we will refer to our two variables as x and y. On scatter diagrams, the horizontal scale is always the x scale and the vertical scale is always the y scale. There are three degrees of correlation which may be observed on a scatter diagram. The two variables may be:

(a) Perfectly Correlated

When the points on the diagram all lie exactly on a straight line (Figure 7.3):

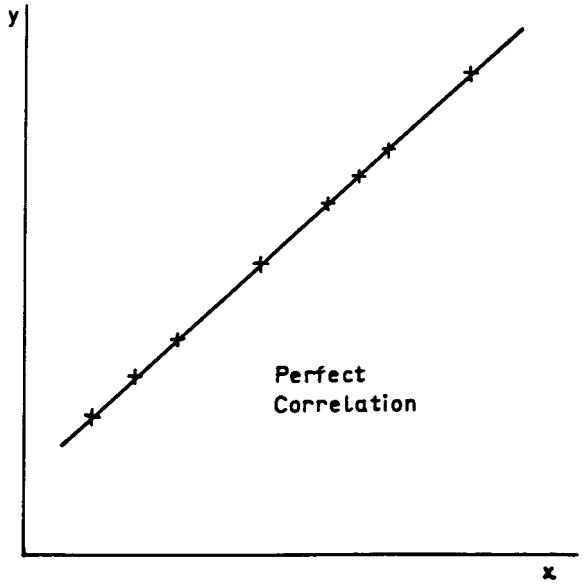


Figure 7.3

(b) Uncorrelated

When the points on the diagram appear to be randomly scattered about, with no suggestion of any relationship (Figure 7.4):

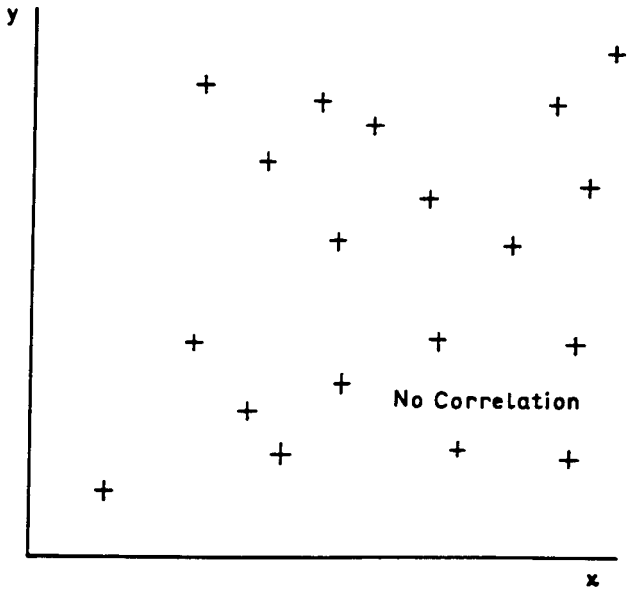


Figure 7.4

(c) Partly Correlated

When the points lie scattered in such a way that, although they do not lie exactly on a straight line, they do display a general tendency to be clustered around such a line (Figure 7.5):

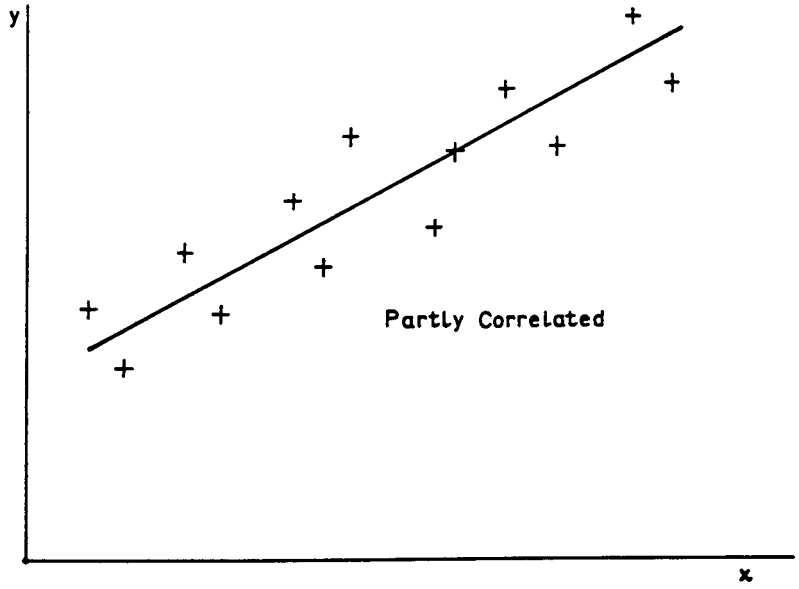


Figure 7.5

Different Types of Correlation

There is a further distinction between correlations of the height/weight type and those of the impurity/cost type. In the first case, high values of the x variable are associated with high values of the y variable, while low values of x are associated with low values of y. On the scatter diagram (Figure 7.6 (a)), the points have the appearance of clustering about a line which slopes **up to the right**. Such correlation is called **positive** or **direct** correlation.

In the other case (like the impurity/cost relationship) high values of the x variable are associated with low values of the y variable and vice versa; on the scatter diagram (Figure 7.6 (b)) the approximate line slopes **down to the right**. This correlation is said to be **negative** or **inverse**.

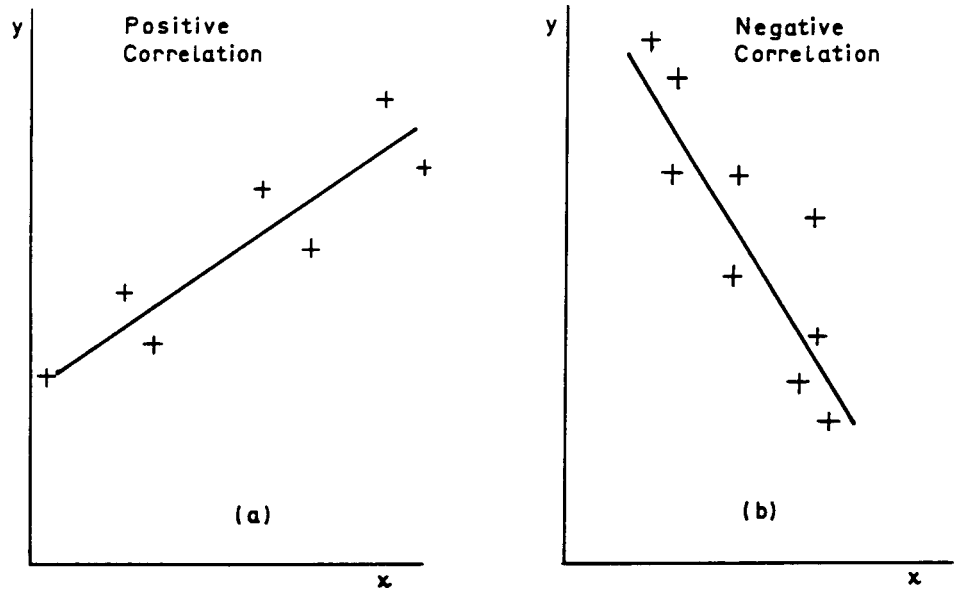


Figure 7.6

(a) Linear Correlation

The correlation is said to be linear when the relationship between the two variables is linear. In other words all the points can be represented by straight lines. For example, the correlation between car ownership and family income may be linear as car ownership is related in a linear fashion to family income.

(b) Non-linear Correlation

Non-linear correlation is outside the scope of this course but it is possible that you could be required to define it in an examination question. It occurs when the relationship between the two variables is non-linear. An example is the correlation between the yield of a crop, like carrots, and rainfall. As rainfall increases so does the yield of the crop of carrots, but if rainfall is too large the crop will rot and yield will fall. Therefore, the relationship between carrot production and rainfall is non-linear.

C. THE CORRELATION COEFFICIENT

General

If the points on a scatter diagram all lie very close to a straight line, then the correlation between the two variables is stronger than it is if the points lie fairly widely scattered away from the line.

To measure the strength, or intensity, of the correlation in a particular case, we calculate a **linear correlation coefficient**, which we indicate by the small letter r. In textbooks and examination papers you will sometimes find this referred to as Pearson's Product Moment Coefficient of Linear Correlation, after the English statistician who invented it. It is also known as the product-moment correlation coefficient.

For an illustration of the method used to calculate the correlation coefficient, suppose we are given the following pairs of values of x and y:

x	10	14	7	12	5	6
y	5	3	5	2	7	8

We shall plot these on a scatter diagram so that we can make some qualitative assessment of the type of correlation present (Figure 7.7). We see from the scatter diagram that some negative correlation appears to be present:

Scatter Diagram

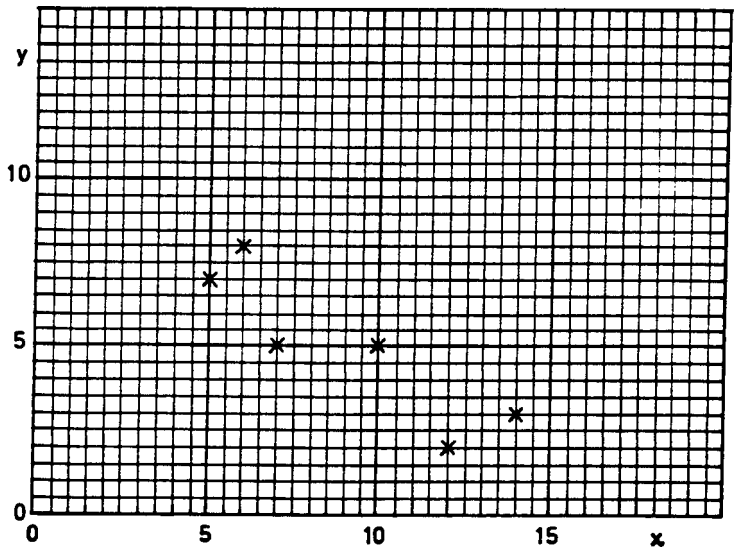


Figure 7.7

Formula

The formula for Pearson's product-moment correlation coefficient (the proof is beyond the scope of this course) is:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}}$$

where: n is the number of pairs of readings.

It is a good idea to set out the calculation in tabular form:

x	y	x^2	y^2	xy
10	5	100	25	50
14	3	196	9	42
7	5	49	25	35
12	2	144	4	24
5	7	25	49	35
6	8	36	64	48
$\sum x = 54$	$\sum y = 30$	$\sum x^2 = 550$	$\sum y^2 = 176$	$\sum xy = 234$

n = 6

Therefore: $r = \dfrac{6 \times 234 - 54 \times 30}{\sqrt{\left(6 \times 550 - 54^2\right)\left(6 \times 176 - 30^2\right)}}$

$$= \frac{1{,}404 - 1{,}620}{\sqrt{(3{,}300 - 2{,}916)(1{,}056 - 900)}}$$

$$= \frac{-216}{\sqrt{384 \times 156}} = \frac{-216}{\sqrt{59{,}904}} = \frac{-216}{244.75} = -0.88 \text{ to 2 dec. places}$$

This result (r = –0.88) shows that x and y are negatively correlated.

Characteristics of a Correlation Coefficient

We know what the + and – signs of the correlation coefficient tell us: that the relationship is positive (increase of x goes with increase of y) or negative (increase of x goes with decrease of y). But what does the actual numerical value mean? Note the following points (the proofs are again beyond the scope of this course):

(a) The correlation coefficient is always between –1 and +1 inclusive. If you get a numerical value bigger than 1, then you've made a mistake!

(b) A correlation coefficient of –1.0 occurs when there is **perfect negative correlation**, i.e. all the points lie **exactly** on a straight line sloping down from left to right.

(c) A correlation of 0 occurs when there is **no correlation**.

(d) A correlation of +1.0 occurs when there is **perfect positive correlation**, i.e. all the points lie **exactly** on a straight line sloping upwards from left to right.

(e) A correlation of between 0 and + 1.0. indicates that the variables are **partly correlated**. This means that there is a relationship between the variables but that the results have also been affected by other factors.

In our example (r = –0.88), we see that the two variables are quite strongly negatively correlated. If the values of r had been, say, –0.224, we should have said that the variables were only slightly negatively correlated. For the time being, this kind of interpretation is all that you need consider.

Significance of the Correlation Coefficient

Correlation analysis has been applied to data from many business fields and has often proved to be extremely useful. For example, it has helped to locate the rich oil fields in the North Sea and also helps the stockbroker to select the best shares in which to put his clients' money.

Like many other areas of statistical analysis, correlation analysis is usually applied to sample data. Thus the coefficient, like other statistics derived from samples, must be examined to see how far they can be used to make generalised statements about the population from which the samples were drawn. **Significance tests** for the correlation coefficient are possible to make, but they are beyond the scope of this course, although you should be aware that they exist.

We must be wary of accepting a high correlation coefficient without studying what it means. Just because the correlation coefficient says there is some form of association, we should not accept it without some other supporting evidence. We must also be wary of drawing conclusions from data that does not contain many pairs of observations. Since the sample size is used to calculate the coefficient, it will influence the result and, whilst there are no hard and fast rules to apply, it may well be that a correlation of 0.8 from 30 pairs of observations is a more reliable statistic than 0.9 from 6 pairs.

Another useful statistic is r^2 (r squared); this is called the **coefficient of discrimination** and may be regarded as the percentage of the variable in y directly attributable to the variation in x. Therefore, if you have a correlation coefficient of 0.8, you can say that approximately 64 per cent (0.8^2) of the variation in y is explained by variations in x. This figure is known as the **explained variation** whilst the balance of 36% is termed the **unexplained variation**. Unless this unexplained variation is small there may be other causes than the variable x which explain the variation in y, e.g. y may be influenced by other variables or the relationship may be non-linear.

In conclusion, then, the coefficient of linear correlation tells you only part of the nature of the relationship between the variables; it shows that such a relationship exists. You have to interpret the coefficient and use it to deduce the form and find the significance of the association between the variables x and y.

Note on the Computation of r

Often the values of x and y are quite large and the arithmetic involved in calculating r becomes tedious. To simplify the arithmetic and hence reduce the likelihood of numerical slips, it is worth noting the following points:

(a) We can take any constant amount off every value of x

(b) We can take any constant amount off every value of y

(c) We can divide or multiply every value of x by a constant amount

(d) We can divide or multiply every value of y by a constant amount

all without altering the value of r. This also means that the value of r is independent of the units in which x and y are measured.

Let's consider the above example as an illustration. We shall take 5 off all the x values and 2 off all the y values to demonstrate that the value of r is unaffected. We call the new x and y values, x' (x-dash) and y' respectively:

x	y	x′	y′	$(x')^2$	$(y')^2$	x′y′
10	5	5	3	25	9	15
14	3	9	1	81	1	9
7	5	2	3	4	9	6
12	2	7	0	49	0	0
5	7	0	5	0	25	0
6	8	1	6	1	36	6
Totals		24	18	160	80	36

n = 6

Therefore: $$r = \frac{n\sum x'y' - \sum x' \sum y'}{\sqrt{\left[n\sum (x')^2 - \left(\sum x'\right)^2\right]\left[n\sum (y')^2 - \left(\sum y'\right)^2\right]}}$$

$$= \frac{6 \times 36 - 24 \times 18}{\sqrt{\left(6 \times 160 - 24^2\right)\left(6 \times 80 - 18^2\right)}}$$

$$= \frac{216 - 432}{\sqrt{(960 - 576)(480 - 324)}}$$

$$= \frac{-216}{\sqrt{384 \times 156}} = -0.88 \text{ to 2 dec. places}$$

Thus the result is identical and the numbers involved in the calculation are smaller, taken overall.

D. RANK CORRELATION

General

Sometimes, instead of having actual measurements, we only have a record of the **order** in which items are placed. Examples of such a situation are:

(a) We may arrange a group of people in order of their heights, without actually measuring them. We could call the tallest No. 1, the next tallest No. 2, and so on.

(b) The results of an examination may show only the order of passing, without the actual marks; the highest-marked candidate being No. 1, the next highest being No. 2, and so on.

Data which is thus arranged in order of merit or magnitude is said to be **ranked**.

Relationship between Ranked Variates

Consider, as an example, the case of eight students who have take the same two examinations, one in Mathematics and one in French. We have not been told the actual marks obtained in the examination, but we have been given the relative position (i.e. the **rank**) of each student in each subject:

Student	Relative Position	
	French	*Mathematics*
A	8	6
B	5	5
C	3	4
D	6	7
E	7	8
F	2	1
G	1	3
H	4	2

We see from this table of ranks that student F was top in Mathematics but only second in French. Student G was top of the class in French, student E was bottom of the class (rank 8) in Mathematics, and so on.

A question which naturally arises is, "Is there any relationship between the students' performances in the two subjects?" This question can be put into statistical terms by asking: "Is there any correlation between the students' ranks in Mathematics and their ranks in French?" The answer to the question will fall into one of the following three categories:

(a) **No correlation**: no connection between performance in the Mathematics examination and performance in the French examination.

(b) **Positive correlation**: students who do well in one of the subjects will, generally speaking, do well in the other.

(c) **Negative correlation**: students who do well in one of the subjects will, generally speaking, do poorly in the other.

We will start our analysis by drawing the scatter diagram as in Figure 7.8. It does not matter which subject we call x and which y.

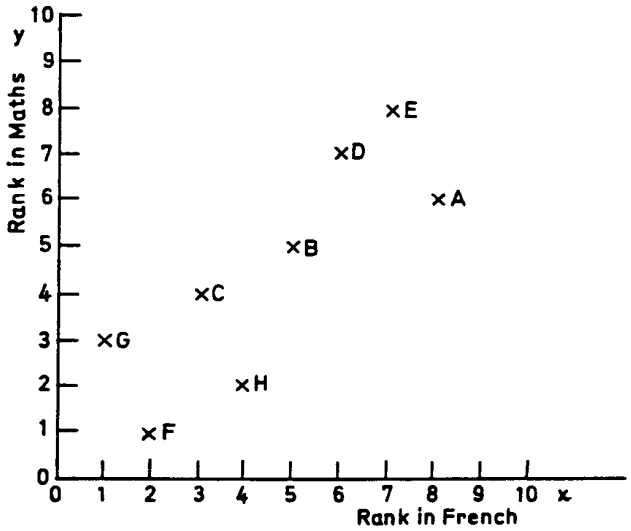


Figure 7.8: Scatter Diagram of Students' Results

The general impression given by the scatter diagram is that there is positive correlation. To find out how strong this correlation is, we calculate the correlation coefficient:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

n = 8

Student	Rank in French *(x)*	Rank in Maths *(y)*	x^2	y^2	xy
A	8	6	64	36	48
B	5	5	25	25	25
C	3	4	9	16	12
D	6	7	36	49	42
E	7	8	49	64	56
F	2	1	4	1	2
G	1	3	1	9	3
H	4	2	16	4	8
Total	36	36	204	204	196

$$r = \frac{8 \times 196 - (36)^2}{\sqrt{\left[8 \times 204 - (36)^2\right]\left[8 \times 204 - (36)^2\right]}} = \frac{1{,}568 - 1{,}296}{1{,}632 - 1{,}296}$$

$$= \frac{272}{336} = 0.81$$

Ranked Correlation Coefficients

With ranked variates, there are simpler methods of calculating a correlation coefficient.

(a) Spearman's Rank Correlation Coefficient

This is usually denoted by the letter r_s. Its formula is:

$$r_s = 1 - \frac{6\sum d^2}{n^3 - n} \text{ i.e. } r_s = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$

In some books you may find R or the Greek letter ρ (pronounced "roe") used instead of r_s but you will recognise Spearman's coefficient by its formula.

In this formula, d is the difference between the two ranks for any one item, and n is the number of items involved. In the above example, n = 8. You can follow the calculation of r_s in the following table:

Student	Rank in:		d	d^2
	Maths	*French*		
A	6	8	−2	4
B	5	5	0	0
C	4	3	1	1
D	7	6	1	1
E	8	7	1	1
F	1	2	−1	1
G	3	1	2	4
H	2	4	−2	4
Total		*(Check)*	0	16

$$r_s = 1 - \frac{6 \times 16}{8^3 - 8} = 1 - \frac{96}{512 - 8} = 1 - \frac{96}{504} = 1 - \frac{12}{63}$$

$$= 1 - 0.19 = +0.81$$

When there is perfect agreement between the ranks of the two variates, then all the values of d will be 0 and so the rank correlation coefficient will be +1.0. When there is complete disagreement between the ranks, the values of d will be at their maximum and the rank correlation coefficient is −1.0.

(b) Kendall's Rank Correlation Coefficient

This is usually denoted by the Greek letter τ (pronounced "taw"). It does not give exactly the same answer as Spearman's method. Its formula is:

$$\tau = \frac{S}{\frac{1}{2}n(n-1)}$$

where, as before, n is the number of pairs of observations. S is referred to as the score of the ranks.

To work out the score, we first arrange the students in order of their French ranks. We then consider for each student in turn whether the differences in French rankings between him and students lower down the list have the same signs as the differences in their Mathematics rankings. If the signs are the same, a pair of students is said to be **concordant**. If the signs are different, the pair is **discordant**. The score, S, is $(n_c - n_d)$ where n_c is the total number of concordant pairs and n_d is the total number of discordant pairs. It is easiest to set out the calculation in a table:

Student	Rank in:		n_c	n_d	$n_c - n_d$
	French	*Mathematics*			
G	1	3	5	2	3
F	2	1	6	0	6
C	3	4	4	1	3
H	4	2	4	0	4
B	5	5	3	0	3
D	6	7	1	1	0
E	7	8	0	1	–1
A	8	6	0	0	0
Total					18

Compared with Student G, whose French rank is 1, all other French ranks have a higher numerical value. Students G's Maths rank is 3, however, so there are 5 Maths ranks with a higher numerical value and 2 with a lower numerical value. Thus $n_c = 5$ and $n_d = 2$. Similarly, for Student F, all French ranks below him in the table have higher numerical values and so do all the Maths ranks so $n_c = 6$ and $n_d = 0$. n_c and n_d are found similarly for the other students. Each student should be compared only with those **lower down** the table, so that each pair of French and Maths rankings is considered once only.

$$\tau = \frac{18}{\frac{1}{2} \times 8 \times 7} = \frac{36}{56} = 0.64 \text{ to 2 dp}$$

This value, being relatively large and positive, again shows a tendency for a high mark in French to be associated with a high mark in Maths, although the agreement is not perfect.

Tied Ranks

Sometimes it is not possible to distinguish between the ranks of two or more items. For example, two students may get the same mark in an examination and so they have the same rank. Or, two or more people in a group may be the same height. In such a case, we give all the equal ones an average rank and then carry on **as if we had given them different ranks.**

You will see what this means by studying the following examples:

(a) First two equal out of eight:

1½ **1½** 3 4 5 6 7 8

Average of 1 & 2

(b) Three equal out of nine, but not at the ends of the list:

1 2 3 **5** **5** **5** 7 8 9

Average of 4, 5 & 6

(c) Last two equal out of eight:

1	2	3	4	5	6	**7½**	**7½**
						Average of 7 & 8	

(d) Last four equal out of eleven:

1	2	3	4	5	6	7	**9½**	**9½**	**9½**	**9½**
							Average of 8, 9, 10 & 11			

Strictly speaking, a rank correlation coefficient should not be used in these cases without making some adjustment for tied ranks. But the formula for the adjustments are a little complex and are outside the scope of this course. The best way for you to deal with tied ranks in practice is to calculate the ordinary (Pearson's) correlation coefficient. If, in an examination, you are specifically asked to calculate a rank correlation coefficient when there are tied ranks, then of course you must do so; but you might reasonably add a note to your answer to say that, because of the existence of tied ranks, the calculated coefficient is only an approximation, although probably a good one.

Final note: Rank correlation coefficients may be used when the actual observations (and not just their rankings) **are** available. We first work out the rankings for each set of data and then calculate Spearman's or Kendall's coefficient as above. This procedure is appropriate when we require an approximate value for the correlation coefficient. Pearson's method using the **actual** observations is to be preferred in this case, however, so calculate a rank correlation coefficient only if an examination question specifically instructs you to do so.

Study Unit 8

Linear Regression

Contents *Page*

A. INTRODUCTION

We've seen how the correlation coefficient measures the degree of relationship between two variates. With perfect correlation (r = +1.0 or r = –1.0), the points of the scatter diagram all lie exactly on a straight line. It is sometimes the case that two variates are perfectly related in some way such that the points would lie exactly on a line, but not a **straight** line. In such a case r would not be 1.0. This is a most important point to bear in mind when you have calculated a correlation coefficient; the value may be small, but the reason may be that the correlation exists in some form other than a straight line.

The correlation coefficient tells us the extent to which the two variates are linearly related, but it does not tell us how to find the particular straight line which represents the relationship. The problem of determining which straight line best fits the points of a particular scatter diagram comes under the heading of **linear regression** analysis.

Remember that a straight-line graph can always be used to represent an equation of the form $y = mx + c$. In such an equation, y and x are the variables while m and c are the constants. Figure 8.1 shows a few examples of straight-line graphs for different values of m and c. Note the following important features of these linear graphs:

- The value of c is always the value of y corresponding to x = 0.
- The value of m represents the **gradient** or **slope** of the line. It tells us the number of units change in y per unit change in x. Larger values of m mean steeper slopes.

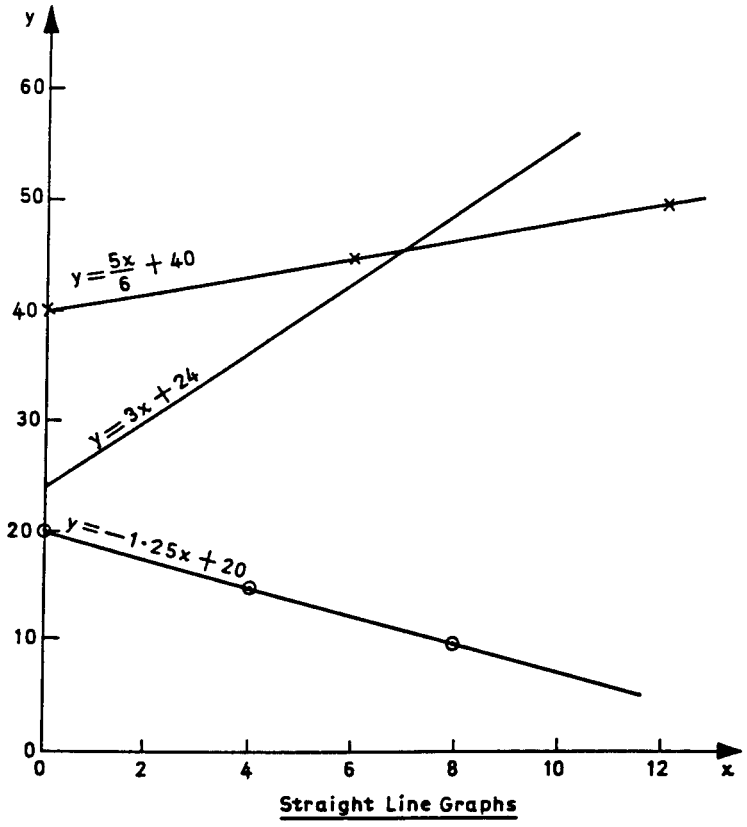


Figure 8.1

- Negative values of the gradient, m, mean that the line slopes **downwards** to the right; positive values of the gradient, m, mean that the line slopes **upwards** to the right.

So long as the equation linking the variables y and x is of the form $y = mx + c$, it is always possible to represent it graphically by a straight line. Likewise, if the graph of the relationship between y and x is a straight line, then it is always possible to express that relationship as an equation of the form $y = mx + c$.

Often in regression work the letters a and b are used instead of c and m, i.e. the regression line is written as $y = a + bx$. You should be prepared to meet both forms.

If the graph relating y and x is **not** a straight line, then a more complicated equation would be needed. Conversely, if the equation is **not** of the form $y = mx + c$ (if, for example, it contains terms like x^2 or log x) then its graph would be a curve, not a straight line.

B. REGRESSION LINES

Nature of Regression Lines

When we have a scatter diagram whose points suggest a straight-line relationship (though not an exact one), and a correlation coefficient which supports the suggestion (say, r equal to more than about 0.4 or 0.5), we interpret this by saying that there is a linear relationship between the two variables but there are other factors (including errors of measurement and observation) which operate to give us a scatter of points around the line instead of exactly on it.

In order to determine the relationship between y and x, we need to know what **straight line** to draw through the collection of points on the scatter diagram. It will not go through all the points, but will lie somewhere in the midst of the collection of points and it will slope in the direction suggested by the points. Such a line is called a **regression line**.

In Figure 8.2 x is the monthly output of a factory and y is the total monthly costs of the factory; the scatter diagram is based on last year's records. The line which we draw through the points is obviously the one which we think best fits the situation, and statisticians often refer to regression lines as **lines of best fit**. Our problem is how to draw the best line.

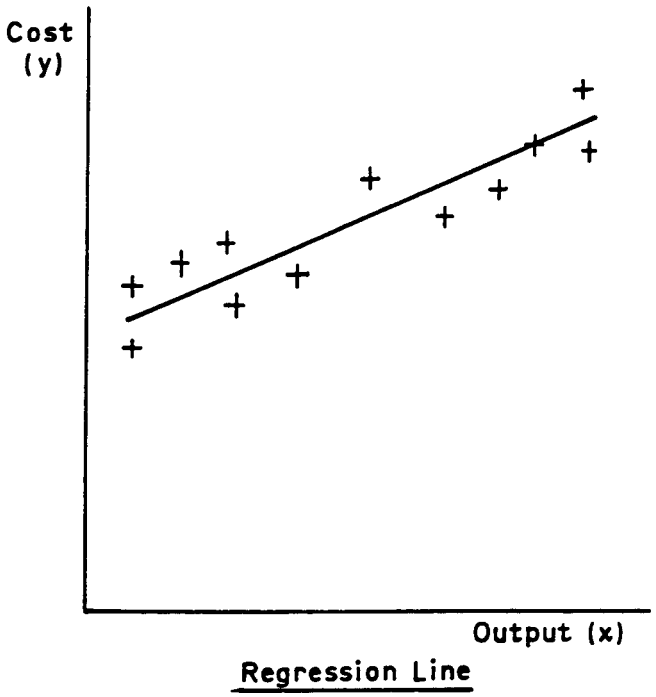


Figure 8.2

There are two methods available – a graphical method and a mathematical method.

Graphical Method

It can be proved mathematically (but you don't need to know how!) that the regression line **must** pass through the point representing the arithmetic means of the two variables. The graphical method makes use of this fact, and the procedure is as follows:

(a) Calculate the means $\bar{x}$ and $\bar{y}$ of the two variables.

(b) Plot the point corresponding to this pair of values on the scatter diagram.

(c) Using a ruler, draw a straight line through the point you have just plotted and lying, as evenly as you can judge, among the other points on the diagram.

In Figure 8.3 the above procedure was followed using the data from the section on the correlation coefficient in the previous study unit. If someone else (you, for example) were to do it, you might well get a line of a slightly different slope, but it would still go through the point of the means (marked +).

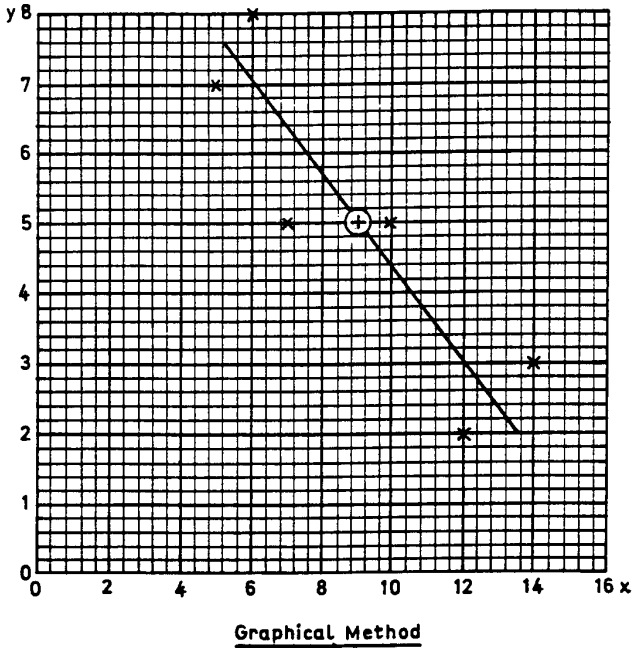


Figure 8.3

Quite obviously, this method is not exact (no graphical methods are) but it is often sufficient for practical purposes. The stronger the correlation, the more reliable this method is, and with perfect correlation there will be little or no error involved.

Mathematical Method

A more exact method of determining the regression line is to find mathematically the values of the constants m and c in the question y = mx + c, and this can be done very easily. This method is called the **least squares** method, as the line we obtain is that which **minimises the sum of the squares of the vertical deviations of the points from the line.** The equation of the least squares line is:

$y = mx + c$, although this is sometimes written as $y = a + bx$

where: $m = \dfrac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$

$c = \bar{y} - m\bar{x}$ or $\dfrac{\sum y - m\sum x}{n}$

n = number of pairs of readings.

We will now apply these formulae to the example we used when talking about the correlation coefficient. If you look back at the last study unit you will see that we had the following figures:

$\sum x = 54; \quad \sum y = 30; \quad \sum x^2 = 550 \quad \sum xy = 234; \quad n = 6$

Therefore: $\bar{x} = 9$, and

$\bar{y} = 5$

Applying the formulae, we get:

$$m = \frac{6 \times 234 - 54 \times 30}{6 \times 550 - (54)^2} = \frac{-216}{384} = -0.5625$$

$$c = 5 - (-0.5625)9 = 5 + 5.0625 = 10.0625$$

m and c are termed the **regression coefficients** (and m also represents the gradient, as previously stated).

The equation for the regression line in this case is therefore:

$$y = 10.0625 - 0.5625x$$

To draw this line on the scatter diagram, choose two values of x, one towards the left of the diagram and one towards the right. Calculate y for each of these values of x, plot the two points and join them up with a straight line. If you have done the calculations correctly, the line will pass through the $(\bar{x}, \bar{y})$ point.

For drawing the regression line, we will choose values of x which are convenient, e.g. x = 0 and x = 16. The corresponding values of y are:

For x = 0, y = 10.0625 – 0 = 10.0625

For x = 16, y = 10.0625 – 16(0.5625) = 10.0625 – 9.0 = 1.0625

The two points marked . are shown in the scatter diagram in Figure 8.4, together with the individual points $(\times)$, the regression line (drawn as an unbroken line) and the mean point (+).

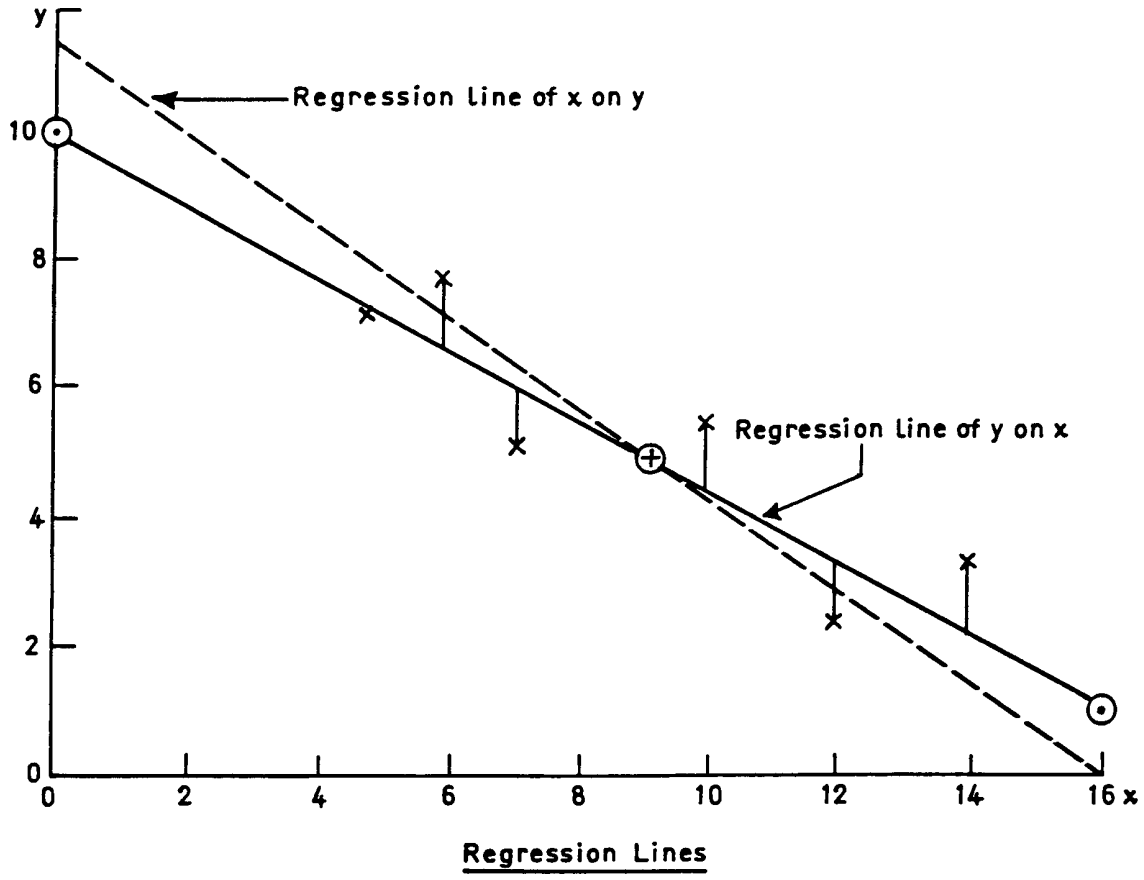


Figure 8.4

The regression line which we have drawn, and the equation which we have determined, represent the **regression of y upon x**. We could, by interchanging x and y, have obtained the regression of x on y. This would produce a different line and a different equation. This latter line is shown in Figure 8.4 by a broken line. The question naturally arises, "Which regression line should be used?". The statistician arrives at the answer by some fairly complicated reasoning but, for our purposes, the answer may be summed up as follows:

(a) Always use the regression of y on x. That is, use the method described in detail above, putting y on the **vertical** axis and x on the **horizontal** axis.

(b) If you intend to use the regression line to predict one thing from another, then the thing you want to predict is treated as y; the other thing is x. For example, if you wish to use the regression line (or its equation) to predict costs from specified outputs, then the outputs will be the x and the costs will be the y.

(c) If the regression is not to be used for prediction, then the x should be the variate whose value is known more reliably.

C. USE OF REGRESSION

The main use of a regression line is to calculate values of the dependent variable not observed in the data set. Take as our example that of employees' heights with a regression equation of:

$$y = 2.87(x) - 345.33$$

where x is height.

Of the 12 people measured and weighed there was nobody of height 181 cm; therefore, if we wanted to know the weight of somebody of this height, it would be impossible to read it from the data available. However, by assuming that a linear relationship exists between weight and height it is possible, by using the regression equation, to calculate an estimate of the weight:

$$x = 181$$

$$y = 2.87(181) - 345.33 = 174.14 \text{ lb}$$

Therefore, the estimated weight of somebody of height 181 cm is 174.14 lb.

Since the value of x (181 cm) lies **within the observed range** of x from the 12 people, we say that we have estimated the value of y by **interpolation**.

However, if we wish to use a regression equation to forecast a result from values which are **outside the range of observations** from which the line is calculated, we have to consider carefully the validity of the estimate obtained. This use of the regression line is called **extrapolation** and we have to assume that the same linear relationship will exist for observations beyond those from which it has been formulated. For example, say we want to estimate the weight of somebody whose height is 194 cm, this value is outside the range of the 12 people measured but y can still be calculated as:

$$x = 194$$

$$y = 2.87(194) - 345.33 = 211.45 \text{ lb}$$

This result seems reasonable, but common sense suggests that values of x much smaller than 160 cm or much larger than 186 cm would be rather improbable.

Sometimes this assumption of the same linear relationship is incorrect, as the factors that influenced the two variables may not remain constant outside the range from which the regression equation is formed, or some extra factor may be introduced.

Consider the relationship between time and the average working wage; if a regression line calculated from data that is collected during years where inflation is very low is used to estimate the wage for years of high inflation, the predicted figure will be much lower than the actual figure, i.e. the change in inflation will change the relationship between the variables. This emphasises that extrapolation gives reliable results only for values **close to the ends of the observed range**.

D. CONNECTION BETWEEN CORRELATION AND REGRESSION

The degree of correlation between two variables is a good guide to the likely accuracy of the estimates made from the regression equation. If the correlation is high then the estimates are likely to be reasonably accurate, and if the correlation is low then the estimates will be poor as the unexplained variation is then high.

You must remember that both the regression equations and the correlation coefficient are calculated from the same data, so both of them must be used with caution when estimates are predicted for values outside the range of the observations, i.e. when values are predicted by extrapolation or the correlation coefficient is assumed to remain constant under these conditions. Also remember that the values calculated for both correlation and regression are influenced by the number of pairs of observations used. So results obtained from a large sample are more reliable than those from a small sample.

Questions on correlation and regression are frequently set in examinations and they are also in practical use in many business areas. Therefore a thorough knowledge of both topics is important.

Study Unit 9

Time Series Analysis

Contents *Page*

A. INTRODUCTION

Businesses and governments use statistical analysis of information collected at regular intervals over extensive periods of time to plan future policies. For example, sales values or unemployment levels recorded at yearly, quarterly or monthly intervals are examined in an attempt to predict their future behaviour. Such sets of values observed at regular intervals over a period of time are called **time series**.

The analysis of this data is a complex problem as many variable factors may influence the changes. The first step is to plot the observations on a scattergram, which differs from those we have considered previously, as the points are evenly spaced on the time axis in the order in which they are observed, and the time variable is always the independent variable. This scattergram gives us a good visual guide to the actual changes but is very little help in showing the component factors causing these changes or in predicting future movements of the dependent variable.

Statisticians have constructed a number of mathematical models to describe the behaviour of time series, and several of these will be discussed in this study unit.

B. STRUCTURE OF A TIME SERIES

These models assume that the changes are caused by the variation of four main factors; they differ in the relationship between these factors. It will be easier to understand the theory in detail if we relate it to a simple time series so that we can see the calculations necessary at each stage.

Consider a factory employing a number of people in producing a particular commodity, say thermometers. Naturally, at such a factory during the course of a year some employees will be absent for various reasons. The following table shows the number of days lost through sickness over a five year period. Each year has been broken down into four quarters of three months. We have assumed that the number of employees at the factory remained constant over the five years.

Days Lost Through Sickness at a Thermometer Factory

Year	Quarter	Days Lost	Year	Quarter	Days Lost
19x0	1	30	19x1	1	40
	2	20		2	25
	3	15		3	18
	4	35		4	45
19x2	1	45	19x3	1	50
	2	30		2	32
	3	22		3	28
	4	55		4	60
19x4	1	60			
	2	35			
	3	30			
	4	70			

We will begin by plotting the scattergram for the data, as shown in Figure 9.1.

The scattergram of a time series is often called a **historigram**. (Do not confuse this with a histogram, which is a type of bar chart.) Note the following characteristics of a historigram:

(a) It is usual to join the points by **straight** lines. The only function of these lines is to help your eyes to see the pattern formed by the points.

(b) Intermediate values of the variables **cannot** be read from the historigram.

(c) A historigram is simpler than other scattergrams since no time value can have more than one corresponding value of the dependent variable.

(d) Every historigram will look similar to this, but a careful study of the change of pattern over time will suggest which model should be used for analysis.

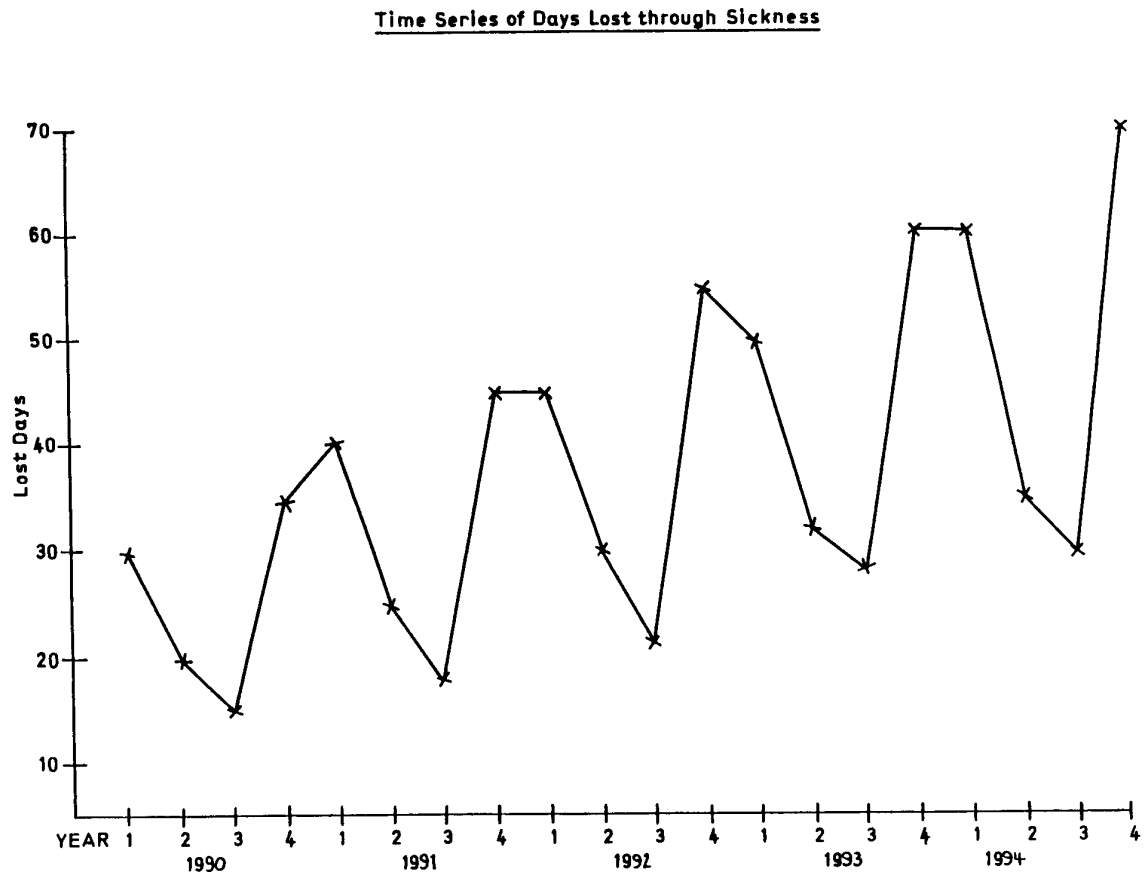


Figure 9.1

There are four factors that influence the changes in a time series – trend, seasonal variations, cyclical fluctuations, irregular or random fluctuations. Now we will consider each in turn.

Trend

This is the change in general level over the whole time period and is often referred to as the **secular trend**. You can see in Figure 9.1 that the trend is definitely upwards, in spite of the obvious fluctuations from one quarter to the next.

A trend can thus be defined as a **clear tendency for the time series data to travel in a particular direction** in spite of other large and small fluctuations. An example of a linear trend is shown in

Figure 9.2. There are numerous instances of a trend, for example the amount of money collected from UK taxpayers is always increasing; therefore any time series describing income from tax would show an upward trend.

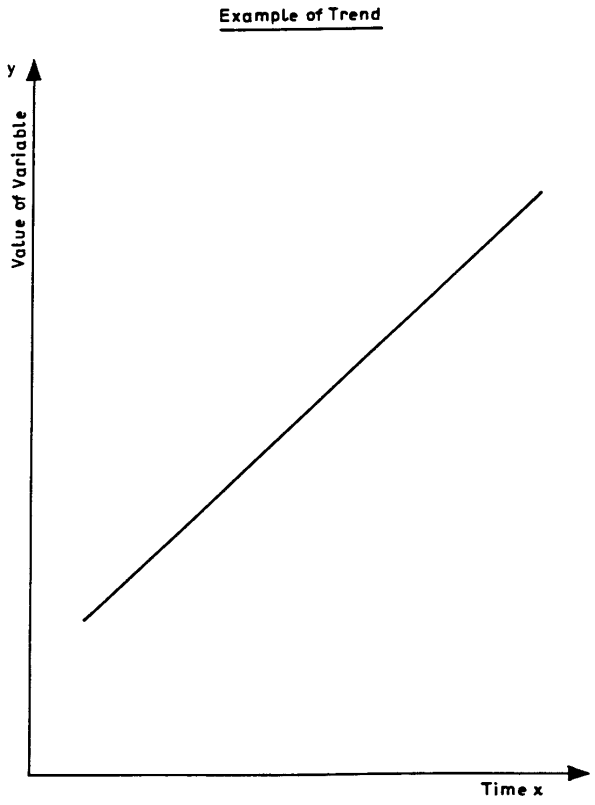


Figure 9.2

Seasonal Variations

These are variations which are repeated over relatively short periods of time. Those most frequently observed are associated with the seasons of the year, e.g. ice-cream sales tend to rise during the summer months and fall during the winter months. You can see in our example of employees' sickness that more people are sick during the winter than in the summer.

If you can establish the variation throughout the year then this **seasonal variation** is likely to be similar from one year to the next, so that it would be possible to allow for it when estimating values of the variable in other parts of the time series. The usefulness of being able to calculate seasonal variation is obvious as, for example, it allows ice-cream manufacturers to alter their production schedules to meet these seasonal changes. Figure 9.3 shows a typical seasonal variation that could apply to the examples above.

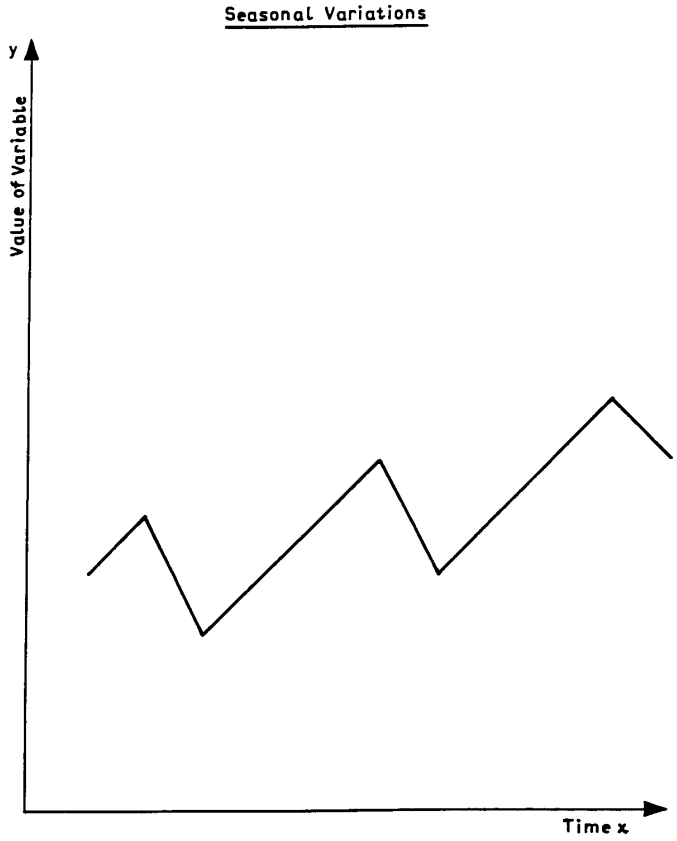


Figure 9.3

Cyclical Fluctuations

These are long-term but fairly regular variations. They are difficult to observe unless you have access to data over an extensive period of time during which external conditions have remained relatively constant. For example, it is well known in the textile trade that there is a cycle of about three years, during which time demand varies from high to low. This is similar to the phenomena known as the **trade cycle** which many economists say exists in the trading pattern of most countries but for which there is no generally accepted explanation.

Figure 9.4 shows how such a cyclical fluctuation would relate to an upward trend. In our example on sickness, a cyclical fluctuation could be caused by, say, a two-year cycle for people suffering from influenza.

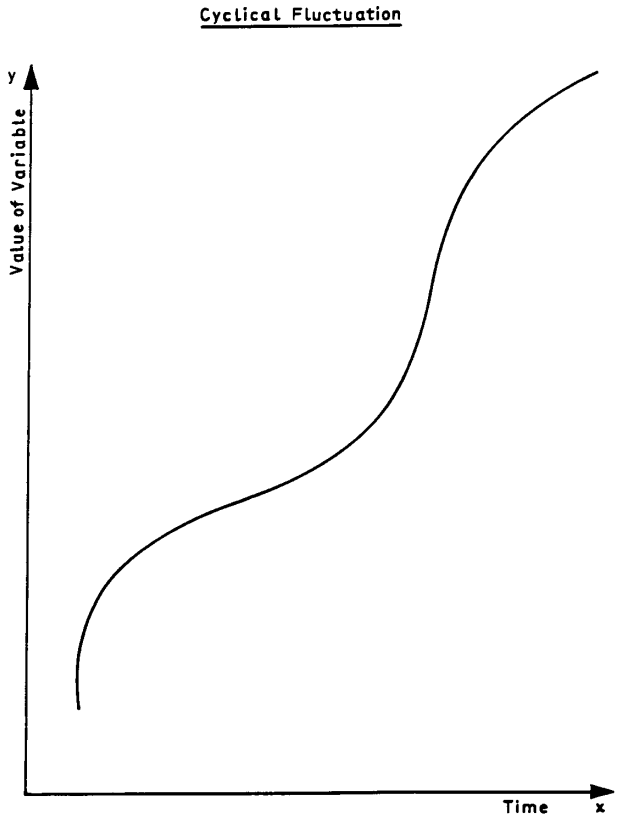


Figure 9.4

As this type of fluctuation is difficult to determine, it is often considered with the final (fourth) element, and the two together are called the **residual variation**.

Irregular or Random Fluctuations

Careful examination of Figure 9.1 shows that there are other relatively small irregularities which we have not accounted for and which do not seem to have any easily seen pattern. We call these irregular or random fluctuations and they may be due to errors of observation or to some one-off external influence which is difficult to isolate or predict. In our example there may have been a measles epidemic in 19x3, but it would be extremely difficult to predict when and if such an epidemic would occur again.

Summary

To sum up, a time series (Y) can be considered as a combination of the following four factors:

- Trend (T)
- Seasonal variation (S)
- Cyclical fluctuation (C)
- Irregular fluctuations (I)

It is possible for the relationship between these factors and the time series to be expressed in a number of ways through the use of different mathematical models. We are now going to look in detail at the **additive model** before looking briefly at the multiplicative and logarithmic models.

C. CALCULATION OF COMPONENT FACTORS FOR THE ADDITIVE MODEL

The additive model can be expressed by the equation:

Time Series = Trend + Seasonal Variation + Cyclical Fluctuations + Random Fluctuations

i.e. $Y = T + S + C + I$

Usually the cyclical and random fluctuations are put together and called the "residual" (R), i.e.:

$$Y = T + S + R$$

Trend

The **most important factor** of a time series is the trend, and before deciding on the method to be used in finding it, we must decide whether the conditions that have influenced the series have remained stable over time. For example, if you have to consider the production of some commodity and want to establish the trend, you should first decide if there has been any significant change in conditions affecting the level of production, such as a sudden and considerable growth in the national economy. If there has, you must consider breaking the time series into sections over which the conditions have remained stable.

Having decided the time period you will analyse, you can use any one of the following methods to find the trend. The basic idea behind most of these methods is to average out the three other factors of variation so that you are left with the long-term trend.

(a) Graphical Method

Once you have plotted the historigram of the time series, it is possible to draw in by eye a line through the points to represent the trend. The result is likely to vary considerably from person to person, unless the plotted points lie very near to a straight line, so it is not a satisfactory method.

(b) Semi-Averages Method

This is a simple method which involves very little arithmetic. The time period is divided into equal parts, and the arithmetic means of the values of the dependent variable in each half are calculated. These means are then plotted at the quarter and three-quarters position of the time series. The line adjoining these two points represents the trend of the series. Note that this line will pass through the overall mean of the values of the dependent variable.

In our example which consists of five years of data, the midpoint of the whole series is mid-way between quarter 2 and quarter 3 of 19x2.

For the mean of the first half:

Year	Quarter	Days Lost
19x0	1	30
	2	20
	3	15
	4	35
19x1	1	40
	2	25
	3	18
	4	45
19x2	1	45
	2	30
	Total	303

Mean = 30.3

For the mean of the second half:

Year	Quarter	Days Lost
19x2	3	22
	4	55
19x3	1	50
	2	32
	3	28
	4	60
19x4	1	60
	2	35
	3	30
	4	70
	Total	442

Mean = 44.2

These values are plotted on the historigram in Figure 9.5. You will notice that 30.3 days, as it is the mean for the first half, is plotted halfway between quarters 1 and 2 of 19x1, and likewise 44.2 days is plotted halfway between quarters 3 and 4 of 19x3. The trend line is then drawn between these two points and it can be extrapolated beyond these points as shown by the dotted line.

If there is an odd number of observations in the time series, the middle observation is ignored and the means of the observations on each side of it are calculated.

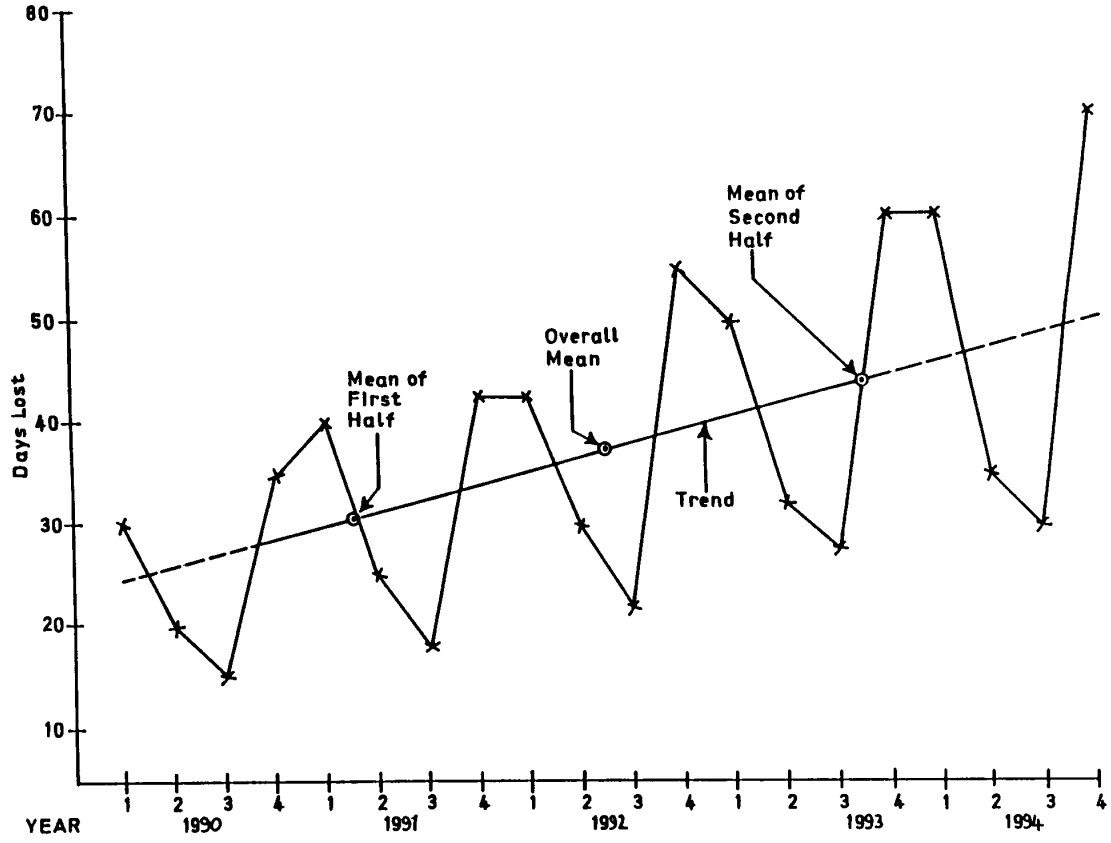


Figure 9.5

(c) Least Squares Method

The trend line is calculated using the formula in Study Unit 8 Section B (the mathematical method of calculating regression lines). In fact the trend line is the regression line of y on x where y is the dependent variable and x is the time variable. Since in a time series the observations are always recorded at equally-spaced time intervals, we can represent x by the first n positive integers, where n is the number of observations. We never calculate the other regression line in time series analysis as it has no significance. Thus the equation of the trend is:

$$y = a + bx \qquad (1)$$

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \qquad (2)$$

$$a = \frac{\sum y - b\sum x}{n} \qquad (3)$$

Using the data given in our earlier example, we set up a table of calculations as follows:

Year	Quarter	x	Days Lost y	x^2	xy
19x0	1	1	30	1	30
	2	2	20	4	40
	3	3	15	9	45
	4	4	35	16	140
19x1	1	5	40	25	200
	2	6	25	36	150
	3	7	18	49	126
	4	8	45	64	360
19x2	1	9	45	81	405
	2	10	30	100	300
	3	11	22	121	242
	4	12	55	144	660
19x3	1	13	50	169	650
	2	14	32	196	448
	3	15	28	225	420
	4	16	60	256	960
19x4	1	17	60	289	1,020
	2	18	35	324	630
	3	19	30	361	570
	4	20	70	400	1,400
Total		210	745	2,870	8,796

n = 20

Therefore: $b = \frac{20(8,796) - 210(745)}{20(2,870) - (210)^2} = \frac{175,920 - 156,450}{57,400 - 44,100}$

$= \frac{19,470}{13,300} = 1.46$

and $a = \frac{745 - 1.46(210)}{20} = \frac{438.4}{20} = 21.92$

So the equation of the trend line is:

$y = 21.92 + 1.46x$

where y is the number of days lost owing to sickness and x is the number given to the quarter required.

We can now draw the line represented by this equation on the time series historigram as shown in Figure 9.6. This method uses all the available information, but it suffers from the same limitations as other regression lines if it is used for prediction by extrapolation.

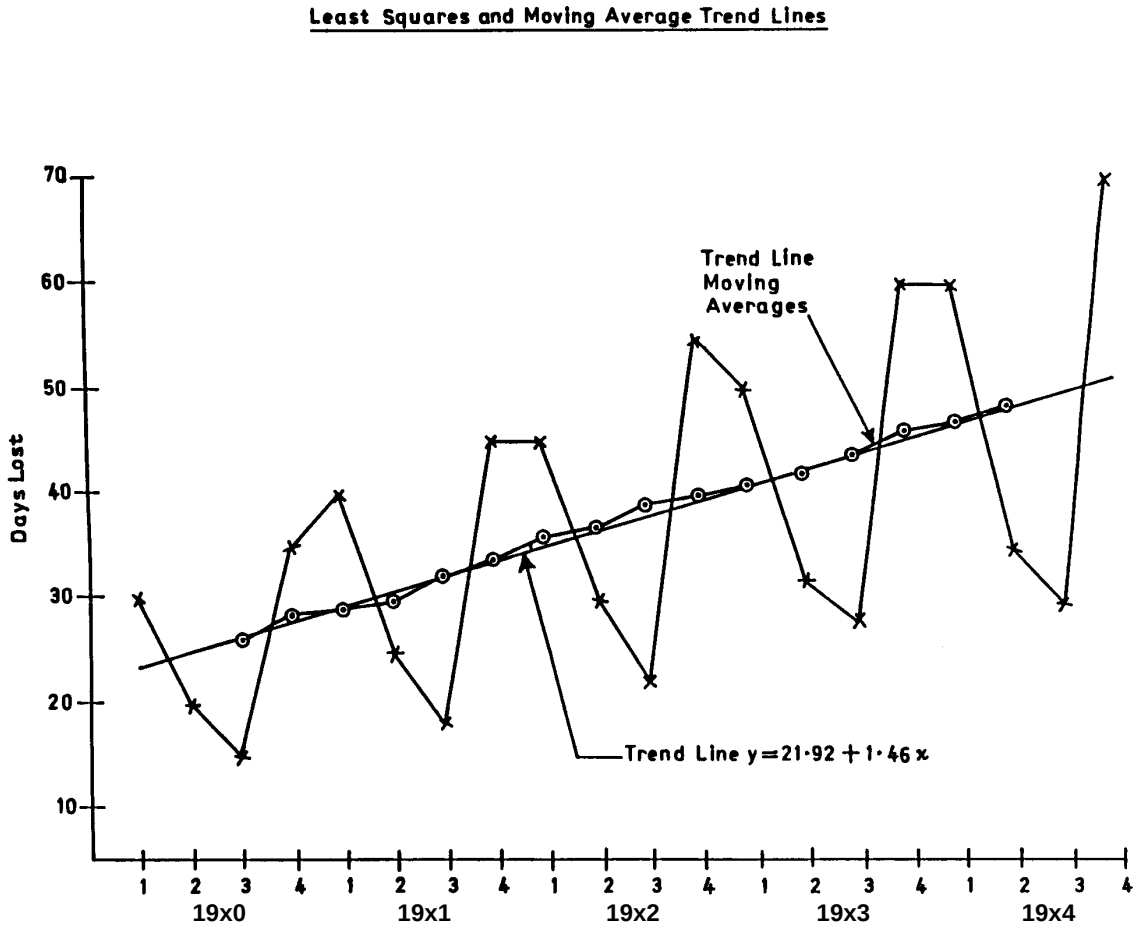


Figure 9.6

(d) Moving Averages Method

So far, the methods we have discussed for finding trends have resulted in a straight line, but the actual trend may be a curve or a series of straight segments. The method of moving averages gives a way of calculating and plotting on the historigram a trend point corresponding to each observed point. These points are calculated by averaging a number of consecutive values of the dependent variable so that variations in individual observations are reduced. The number of consecutive values selected will depend on the length of the short-term or seasonal variation shown on the historigram.

The method of calculating a set of moving averages is illustrated by the following simple example. Consider the seven numbers 6, 4, 5, 1, 9, 5, 6 and take the number of time periods covered by the fluctuations to be four as in quarterly figures, then a moving average of order four is needed.

Step 1: Find the average of the first to fourth numbers.

$$\text{Average} = \frac{6+4+5+1}{4} = 4$$

Step 2: Find the average of the second to fifth numbers.

$$\text{Average} = \frac{4+5+1+9}{4} = 4.75$$

Step 3: Find the average of the third to sixth numbers.

$$\text{Average} = \frac{5+1+9+5}{4} = 5$$

Step 4: Find the average of the fourth to seventh numbers.

$$\text{Average} = \frac{1+9+5+6}{4} = 5.25$$

Hence the moving averages of order 4 are 4, 4.75, 5, 5.25. For monthly data a moving average of order 12 would be needed; for daily data the order would be 7, and so on.

Using the data of the earlier example, we calculate the trend values and plot them on Figure 9.6 so that we can compare the two trend lines. The table of calculations follows:

Year	Quarter	Days Lost	4-Quarter Total	Moving Average	Trend
(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
19x0	1	30			
	2	20			
			100	25	
	3	15			26.3
			110	27.5	
	4	35			28.1
			115	28.75	
19x1	1	40			29.1
			118	29.5	
	2	25			30.8
			128	32.0	
	3	18			32.6
			133	33.25	
	4	45			33.9
			138	34.5	
19x2	1	45			35.0
			142	35.5	
	2	30			36.8
			152	38.0	
	3	22			38.6
			157	39.25	
	4	55			39.5
			159	39.75	
19x3	1	50			40.5
			165	41.25	
	2	32			41.9
			170	42.5	
	3	28			43.8
			180	45.0	
	4	60			45.4
			183	45.75	
19x4	1	60			46.0
			185	46.25	
	2	35			47.5
			195	48.75	
	3	30			
	4	70			

The trend is given correct to one decimal place as this is the greatest accuracy justified by the accuracy of the data. Notice how the table of calculations is set out, with the numbers in columns (4) and (5) placed **midway between** two quarterly readings. This is because we were

averaging over an even number of values, so the moving average would have to be plotted in this position on the historigram and would not correspond to any particular quarter. Thus it is necessary to add column (6) which gives the mean of successive pairs of moving averages and these numbers are the trend values plotted. (The values in column (6) are often called the **centred moving averages.**)

If we were calculating a moving average with an odd number of values it would not be necessary to carry out this final stage as the moving averages would be centred on an actual observation and so would be the trend values, e.g. daily observation over a number of weeks or data with a short-term cycle of an odd number of years.

The main advantage of this method is that the trend values take into account the **immediate** changes in external factors which the trend lines, using the previous two methods, are unable to do. However, this method has three disadvantages:

(i) The trend line cannot be found for the whole of the time series. As you can see from our example, there are no trend values for quarters at the beginning and end of the series.

(ii) Problems can be encountered in deciding the order number, i.e. the period of fluctuation. Unless the seasonal or cyclical movement is definite and clear cut, the moving method of deriving the trend may yield a rather unsatisfactory line.

(iii) Since the trend is calculated as a simple arithmetic mean it can be unduly influenced by a few extreme values.

Seasonal Variation

As we are assuming at present that the additive model is satisfactory, once we have found the trend by one of the methods described in the previous section we can find the value of the remaining factors for each value of the dependent variable from the equation for the additive model by subtraction: i.e.

$$Y = T + S + C + I$$

and so $Y - T = S + C + I = S + R$

($C + I = R$ since we cannot usually separate C and I)

Column (5) of the following table shows the value of this difference for all the quarters from 19x0 quarter 3 to 19x4 quarter 2.

Year	Quarter	Days Lost Y	Trend T	Y – T
(1)	(2)	(3)	(4)	(5)
19x0	3	15	26.3	–11.3
	4	35	28.1	6.9
19x1	1	40	29.1	10.9
	2	25	30.8	–5.8
	3	18	32.6	–14.6
	4	45	33.9	11.1
19x2	1	45	35.0	10.0
	2	30	36.8	–6.8
	3	22	38.6	–16.6
	4	55	39.5	15.5
19x3	1	50	40.5	9.5
	2	32	41.9	–9.9
	3	28	43.8	–15.8
	4	60	45.4	14.6
19x4	1	60	46.0	14.0
	2	35	47.5	–12.5

One of the assumptions we make for the additive model is that the seasonal variations are the same for corresponding quarters in each year. You can see that this is not the case in column (5) except that for each year the first and fourth quarters give a positive result and the second and third a negative one. The variation must be caused by the residual (R), and this factor can be eliminated by calculating the adjusted average for each quarter as shown in the next table:

Year	1st Qtr	2nd Qtr	3rd Qtr	4th Qtr	
19x0			–11.3	6.9	
19x1	10.9	–5.8	–14.6	11.1	
19x2	10.0	–6.8	–16.6	15.5	
19x3	9.5	–9.9	–15.8	14.6	
19x4	14.0	–12.5			
Total	44.4	–35.0	–58.3	48.1	
Average	11.1	–8.8	–14.6	12.0	(–0.3)
Adjusted Average	11.175	–8.725	–14.525	12.075	

The average fluctuations should add up to zero, but as you can see in the example above, because of rounding errors they do not; therefore a minor adjustment is carried out in the last row. This is done by subtracting a quarter of the total outstanding from each average (in this case 0.25 of – 0.3 = – 0.075).

Therefore the values 11.2, –8.7, –14.5 and 12.1 (all correct to 1 dp) are the seasonal fluctuations of the four quarters for the time series of days lost through sickness at a factory.

Deseasonalised Data and Residual

The remaining results that are needed for this analysis are the deseasonalised values (Y – S) and the residuals (Y – S – T). These are shown in columns (4) and (6) of the following table:

Year and Qtr	Days Lost	Seasonal Adjustment	Deseasonalised Data	Trend	Residual
	Y	**S**	**Y – S**	**T**	**R = Y – S – T**
(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
19x0 3	15	–14.5	29.5	26.3	3.2
4	35	12.1	22.9	28.1	–5.2
19x1 1	40	11.2	28.8	29.1	–0.3
2	25	–8.7	33.7	30.7	3.0
3	18	–14.5	32.5	32.6	–0.1
4	45	12.1	32.9	33.9	–1.0
19x2 1	45	11.2	33.8	35.0	–1.2
2	30	–8.7	38.7	36.7	2.0
3	22	–14.5	36.5	38.6	–2.1
4	55	12.1	42.9	39.5	3.4
19x3 1	50	11.2	38.8	40.5	–1.7
2	32	–8.7	40.7	41.9	–1.2
3	28	–14.5	42.5	43.7	–1.2
4	60	12.1	47.9	45.4	2.5
19x4 1	60	11.2	48.8	46.0	2.8
2	35	–8.7	43.7	47.5	–3.8

As you can see, there is no pattern to the residuals but they are fairly small, i.e. they can be considered as random errors of observation and rounding, though they may contain a systematic cyclic element.

D. OTHER MODELS

Multiplicative Model

We have seen that the **additive model** is defined as:

$$Y = T + S + C + I$$

By contrast, the **multiplicative model** is defined as:

$$Y = T \times S \times C \times I$$

The difference between these models rests in the **assumptions** made for each. In the case of the additive model it is assumed that the magnitude of all the factors other than the trend is not affected by the trend, but with the multiplicative model it is assumed that their magnitude is directly proportional to the trend. This means that the effect of a given factor is to change the overall total of the time series by a constant multiple as opposed to a fixed amount. For example, the multiplicative model would assume that sales of umbrellas are 30% above the yearly average in winter, whereas the additive model would assume that sales were 2,000 above the yearly average in winter.

The assumptions made for the **additive model** will be satisfactory as long as the **trend is linear** or alters only slightly during the period of analysis. Under other conditions the multiplicative model is likely to give more reliable results. For example, a company with a turnover of £200,000 per annum is likely to experience considerably greater seasonal fluctuations than one with a turnover of only £10,000 per annum, so that the multiplicative model would be more applicable.

Logarithmic Model

The other type of model quite often used is called the **logarithmic model** and is defined as:

$$\log Y = \log T + \log S + \log C + \log I$$

Basically this is a special case of the additive model and is used when the trend of the original data does not appear to be linear, but when a historigram of the **logarithms** of the data is plotted it is very similar to that shown in Figure 9.1, so that a **linear trend of the logarithms** can reasonably be assumed.

Example of a Multiplicative Model

We will look at the sales of premium bonds at a large post office over five years. The data and calculations are shown in the following table and plotted on the historigram shown in Figure 9.7. You can see that in this data there is a high growth rate, particularly over the years 19x1 and 19x2, and it is in this type of time series that the multiplicative model is best used.

Sales of Premium Bonds

Year and Quarter		Sales *(£000)*	Moving Total	Moving Average	Trend	Sales/ Trend
19x0	1	320				
	2	200				
			1,070	268		
	3	150			278	0.54
			1,150	288		
	4	400			289	1.38
			1,160	290		
19x1	1	400			294	1.36
			1,190	298		
	2	210			304	0.69
			1,240	310		
	3	180			360	0.50
			1,640	410		
	4	450			434	1.04
			1,830	458		
19x2	1	800			478	1.67
			1,990	498		
	2	400			554	0.72
			2,440	610		
	3	340			629	0.54
			2,590	648		
	4	900			661	1.36
			2,690	673		
19x3	1	950			681	1.40
			2,750	688		
	2	500			698	0.72
			2,830	708		
	3	400				
	4	980				

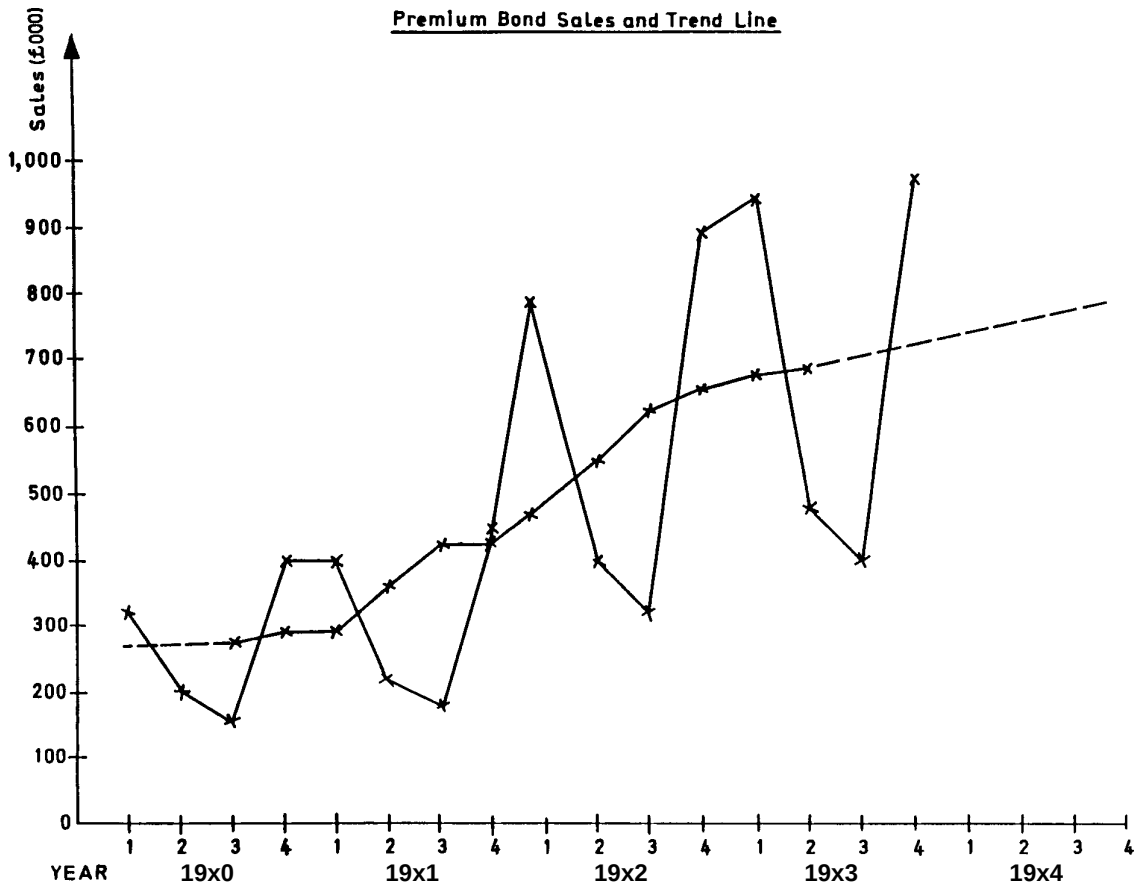


Figure 9.7

The calculation of the trend can be carried out by any of the methods described previously, but the moving average method is usually preferred. The trend is removed from the series by dividing the sales value by the trend, i.e.:

Since $Y = T \times S \times C \times I$

$$\frac{Y}{T} = S \times C \times I$$

Now we need to remove the cyclical and irregular fluctuations from the time series in order to calculate the seasonal fluctuations. This is achieved in a similar way as with the additive model by calculating the average of the Sales/Trend ratios and adjusting where necessary for rounding errors.

Year	Qtr 1	Qtr 2	Qtr 3	Qtr 4	
19x0			0.54	1.38	
19x1	1.36	0.69	0.50	1.04	
19x2	1.67	0.72	0.54	1.36	
19x3	1.40	0.72			
Total	4.43	2.13	1.58	3.78	
Average	1.48	0.71	0.53	1.26	(3.98)
Adjusted Average	1.49	0.71	0.53	1.27	

If this averaging has successfully removed the residual variations, the average ratios should add up to 4.0 units. As they do not, they must be adjusted by multiplying each of them by the ratio:

$$\frac{4.00}{3.98}$$

The analysis is completed by calculating the residual variations which, as before, consist of the cyclical and irregular fluctuations, i.e.:

$$\frac{Y}{T} = S \times R$$

$$\frac{Y}{T \times S} = R$$

The residual variations calculated in this way are shown in the final column of the following table:

Year	Qtr	Sales	Trend	Seasonal	Residual
19x0	3	150	278	0.53	1.02
	4	400	289	1.27	1.09
19x1	1	400	294	1.49	0.91
	2	210	304	0.71	0.97
	3	180	360	0.53	0.94
	4	450	434	1.27	0.82
19x2	1	800	478	1.49	1.12
	2	400	554	0.71	1.02
	3	340	629	0.53	1.02
	4	900	661	1.27	1.07
19x3	1	950	681	1.49	0.94
	2	500	698	0.71	1.01

The final two columns give the proportional changes for seasonal fluctuations and residual variations and they show that seasonal fluctuations can account for up to 50% of the changes in the trend

figures. The residual variations calculated show that the trend and seasonal sales could vary by as much as 18%.

E. FORECASTING

Assumptions

The reason for isolating the trend within a time series is to be able to make a prediction of its future values and thus estimate the movement of the time series. Before looking at the various methods available to carry out this process, we must state two assumptions that must be made when forecasting:

(a) That Conditions Remain Stable

Those conditions and factors which were apparent during the period over which the trend was calculated must be assumed to be unchanged over the period for which the forecast is made. If they do change, then the trend is likely to change with them, thus making any predictions inaccurate, e.g. forecasts of savings trends based on given interest rates will not be correct if there is a sudden change either up or down in these rates.

(b) That Extra Factors Will Not Arise

It is sometimes the case that, when trends are predicted beyond the limits of the data from which they are calculated, extra factors will arise which influence the trend. For example, there is a limit to the number of washing machines that can be sold within a country. This capacity is a factor that must be considered when making projections of the future sales of washing machines. Therefore, in forecasting from a time series it must be assumed that such extra factors will not arise.

These assumptions are similar to those mentioned when we looked at the extrapolation of a regression line.

Methods of Forecasting

There are two main methods of forecasting, although both are primarily concerned with short-term forecasts because the assumptions mentioned previously will break down gradually for periods of longer than about a year.

(a) Moving Averages Method

This method involves extending the moving average trend line drawn on the historigram of the time series. The trend line is extended by assuming that the gradient remains the same as that calculated from the data. The further forward you extend it, the more **unreliable** becomes the forecast.

When you have read the required trend value from the graph, the appropriate seasonal fluctuation is added to this and allowance is made for the residual variation. For example, consider the premium bond sales shown in Figure 9.7. On this figure the moving average trend line stops at the final quarter of 19x4. If this line is extrapolated with the same gradient to the first quarter of 19x5 then:

19x5 1st Qtr: Trend = 750

This is multiplied by the seasonal variation as it is a multiplicative model, i.e.
$750 \times 149 = 1{,}118$, and the residual variation which varied by as much as $\pm$ 18% is added to

this. Therefore the final short-term estimate for the sales of premium bonds for the first quarter of 19x5 is £1,118,000 ± £201,000.

Although fairly easy to calculate, this forecast, like all others, must be treated with caution, because it is based on the value of the trend calculated for the final quarter of 19x4, so if this happens to be an especially high or low value then it would influence the trend, and thus the forecast, considerably.

(b) Least Squares Method

If the line of best fit, $y = a + bx$, is used as the trend line and drawn on a historigram, it can be extended to give an estimate of the trend. Preferably the required value of x can be substituted in the equation to give the trend value. The seasonal fluctuation and residual variations must be added as in (a).

Using the results of the earlier example involving days lost through sickness at a factory, the trend line was:

$$y = 21.92 + 1.46x$$

where x took all the integer values between 1 and 20.

Now suppose we want to estimate the number of days lost in the first quarter of 19x5, i.e. when x = 21. The value of the trend would be:

$$y = 21.92 + 1.46(21)$$

$$= 52.58$$

$$= 53 \text{ days}$$

(This result could also be read from the graph in Figure 9.6.)

To this must be added, as it is an additive model, the seasonal fluctuation for a first quarter, which was about 11 days, making a total of 64 days. The residual variation for this series was a maximum of ± 5 days. Therefore the forecast for days lost through sickness for the first quarter of 19x5 is between 59 and 69 days.

This forecast again is not entirely reliable, as the trend is depicted by one straight line of a fixed gradient. It is a useful method for short-term forecasting, although like the previous method it becomes more **unreliable** the further the forecast is extended into the future.

There are no hard and fast rules to adopt when it comes to choosing a forecast method. Do not think that the more complicated the method the better the forecast. It is often the case that the simpler, more easily understood methods produce better forecasts, especially when you consider the amount of effort expended in making these forecasts. Remember that, whatever the method used for the forecast, it is only an educated guess as to future values.

F. THE Z-CHART

We will conclude this study unit with a short description of a particular type of chart which plots a time series, called a Z-Chart. (We mentioned this also in Study Unit 4.) It is basically a means of showing three sets of data relating to the performance of an organisation over time. The three sets of data are plotted on the same chart and should be kept up-to-date. The graphs are:

- The plot of the current data, be it monthly, quarterly or daily.
- The cumulative plot of the current data.
- The moving total plot of the data.

It is often used to keep senior management informed of business developments. As an example we will plot a Z-Chart for the sales of premium bonds in 19x2 using the data of the table below with the sales broken down into months. The table also shows the cumulative monthly sales and the moving annual totals. Note that the scale used for (a) is shown on the right of the chart and is twice that used for (b) and (c) so that the fluctuations in monthly sales show up more clearly. This is a device often used so that the chart is not too large.

Year	Month	Sales	Cumulative Sales	Moving Annual Total
				1,240
19x2	Jan	150	150	1,290
	Feb	350	500	1,460
	Mar	300	800	1,640
	Apr	100	900	1,670
	May	150	1,050	1,730
	June	150	1,200	1,830
	July	120	1,320	1,890
	Aug	120	1,440	1,940
	Sept	100	1,540	1,990
	Oct	300	1,840	2,140
	Nov	400	2,240	2,340
	Dec	200	2,440	2,440

These totals are presented in Figure 9.8. It is called a Z-Chart because the position of the three graphs on the chart makes it look like the letter Z.

This is a useful chart because management can see at a glance how production is progressing from one month to the next. It is also possible to compare the current year's performance with a set target or with the same periods in previous years.

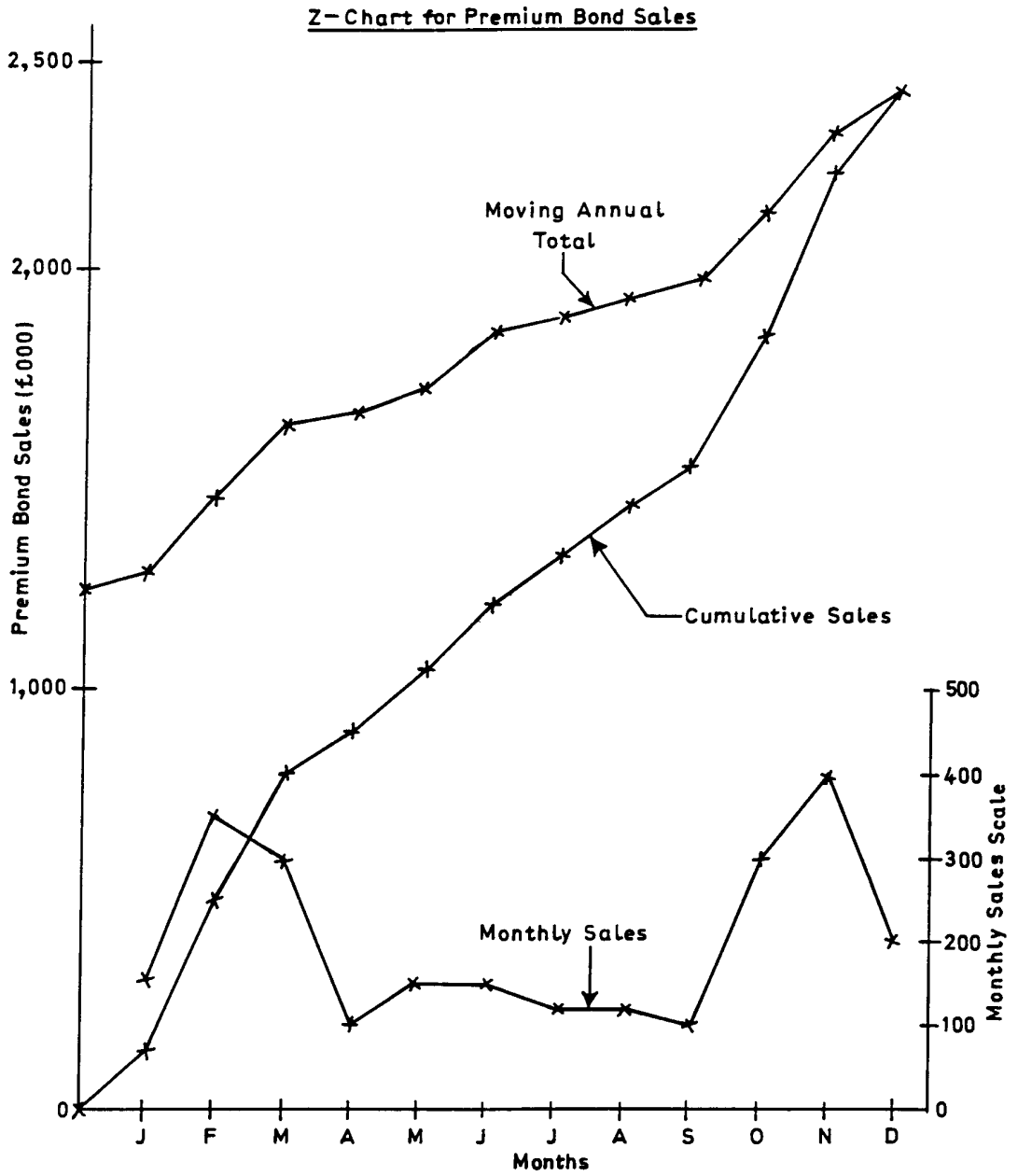


Figure 9.8

G. SUMMARY

In this study unit we discussed the main models used to analyse time series. We began by identifying the various factors into which a time series may be divided in order to use these models, and went on to show how to separate a time series into these **constituent** factors. This is an important subject and you should particularly note the following points:

- Set out all calculations systematically in tables.
- The layout of the table used for calculation of centred moving averages is very important for all models.
- You must learn thoroughly the method of calculating and adjusting seasonal variations for all models.

Study Unit 10

Index Numbers

Contents *Page*

(continued over)

A. THE BASIC IDEA

Table 10.1 shows the monthly profits of Firm X for a period of one year. We could plot profits against time (i.e. each month) and draw a graph. However, if we are interested in **changes** in profits rather than in the actual level of profits, we can use one month's figures, say January, as standard and express all the others as percentages of this standard. Because we are dealing with percentages, we use a standard figure of 100.

In Table 10.1, the right-hand column shows January set to the standard figure of 100 and all the other profit values set to percentages of this standard.

Table 10.1: Monthly Profits of Firm X

Month	Profit	Profit Based on Jan = 100
Jan	512	100
Feb	520	102
Mar	530	104
Apr	531	104
May	546	107
Jun	549	107
Jul	560	109
Aug	565	110
Sep	568	111
Oct	573	112
Nov	584	114
Dec	585	114

The percentage figures in the right-hand column are called **index numbers** of profits and, in this case, January is known as the **base** month against which all others are compared.

The essentials of an index number then are that it illustrates **changes** by expressing the items in a time series as **percentages** of the item at a chosen **base** period.

B. BUILDING UP AN INDEX NUMBER

In commercial and economic affairs there are some very important quantities which are too complex to be measured directly; such things as the "level of industrial production" or the "cost of living". These are real enough things, but they are made up of a very large number of component parts, all affecting the main issue in different ways or to different extents. Index numbers are especially suited to dealing with such matters.

You should note that an index number is sometimes called an **index** (plural: **indices**).

Simple Index

Let us take first of all a much over-simplified family food bill from which we want to show the effects of changes in price over a period of time, i.e. a very elementary cost of living index. Suppose we consider four items of food – milk, butter, tea and potatoes – and we know the average weekly consumption of these items in a typical household for Year 1 and Year 10 and the price of each item, as set out in Table 10.2:

Table 10.2

Item	Year 1		Year 10	
	Average Weekly Consumption	**Price**	**Average Weekly Consumption**	**Price**
Milk	8 pints	5p/pt	5 pints	21p/pt
Butter	1 kg	40p/kg	500g	100p/kg
Tea	250g	60p/kg	125g	200p/kg
Potatoes	8 lbs	2p/lb	6 lbs	7p/lb

We somehow want to combine the price per given unit of each of the items so that we have a single index number comparing prices in Year 10 with those in Year 1.

- Because we are combining the four food items into a single "shopping basket", the index is called an **aggregative index**.
- It makes no allowance for the different quantities of item used – butter as compared to tea, for example – and so is called a **simple index.**
- Finally, because we are comparing prices, it is called a **price index**.

All we do in this extremely simple situation is total the prices/given unit for each year and express that for Year 10 as a percentage of that for Year 1:

Simple aggregative price index (Year 10 compared to Year 1)

$$= \left[\frac{21+100+200+7}{5+40+60+2}\right]\times 100 \;=\; \frac{328}{107}\times 100 \;=\; \mathbf{306.5}$$

This tells us that prices as a whole were more than three times higher in Year 10 than in Year 1, i.e. prices had increased by 206.5% in that period.

Notes

(a) We shall work out all the index numbers in this study unit correct to one decimal place. This is precise enough for most comparisons, and particularly in times of rapid inflation when there are large changes in price indices.

(b) There are no units to an index number as we are expressing one price as a percentage of another price. We must remember to have our prices in the same units, in this case pence, when calculating an aggregative index.

(c) Year 1 is the base year in this example. Instead of having to state this every time in words, it is customary to write: Price index for Year 10 = 306.5 (Year 1 = 100).

You may already have some criticisms to make of this simple approach to constructing an index. Firstly, it depends on the units of the commodities given. Suppose for tea we had said that its price was 30p/500g and 100p/500g instead of 60p/kg and 200p/kg, then:

Simple aggregative price index for Year 10 (Year 1 = 100)

$$= \left[\frac{21+100+100+7}{5+40+30+2}\right]\times 100 = \frac{228}{77}\times 100 = \mathbf{296.1}$$

In other words, we get a completely different value for the index, which is obviously unsatisfactory.

Price Relatives

We can get round this problem by using the **ratio of prices** of a given item rather than the actual prices themselves. Thus, the price of a pint of milk in Year 10 as a percentage of its price in Year 1 is:

$$\frac{21}{5}\times 100 = 420.0.$$

This ratio, 420.0, is called the **price relative** for milk in Year 10 (Year 1 = 100).

Similarly, we can work out price relatives for the other items. (Remember, all we are doing is making the current price into a percentage of the base year price.)

Table 10.3

Commodity	**Price Relatives in Year 10 (Year 1 = 100)**
Milk	$\frac{21}{5}\times 100 = 420.0$
Butter	$\frac{100}{40}\times 100 = 250.0$
Tea	$\frac{200}{60}\times 100 = 333.3$
Potatoes	$\frac{7}{2}\times 100 = 350.0$

From these price relatives we can now construct another index number called the **mean of relatives index**, which is just the arithmetic mean of the price relatives, i.e.:

Mean of relatives index number of Year 10 (Year 1 = 100)

$$= \frac{420.0+250.0+333.3+350.0}{4} = \frac{1{,}353.3}{4} = \mathbf{338.3}$$

In other words, on this basis prices in general appear to have risen 238% over the given period.

Another advantage of this price-relative type of index number is that the prices of all the commodities do not have to be in the same units, although the prices of **each individual item** must be in the same units. This is a useful feature if you are dealing with results from different countries.

C. WEIGHTED INDEX NUMBERS (LASPEYRES AND PAASCHE INDICES)

You may think that the mean of relatives index is still not very satisfactory, in that all items are treated as of equal importance and no account has been taken of the different quantities of the items consumed. For instance, the average family is much more concerned about a 5p increase in the price of a loaf of bread than a 10p increase in the price of a drum of pepper, as far more bread is consumed than pepper.

If you look back at Table 10.2 you will see that we are, in fact, given the average weekly consumption of each item in Year 1 and Year 10. You can see that the consumption pattern, as well as the prices, has changed over the 10-year period. We are interested in calculating an index for **prices**, so we have to be careful not to overemphasise the increase in prices by incorporating the changes in consumption.

Weighted Aggregative Index Numbers

We can adopt either of two approaches:

(a) We can consider the consumption pattern in Year 1 as typical and:

(i) Work out the total expenditure on the four items in Year 1; then,

(ii) Work out what the total expenditure would have been in Year 10 if the family had consumed at Year 1 levels; and finally,

(iii) Express the sum in (ii) as a percentage of (i) to form an index number.

This index is called a **base-weighted aggregative index** and in our example we work as follows:

Year 1 values are (Year 1 consumption × Year 1 prices)

Year 10 values are (Year 1 consumption × Year 10 prices)

In other words, we assume the consumption has not changed, only the prices.

The resulting table of values is:

Table 10.4

Item	Year 1	Year 10
	Expenditure Using Year 1 Consumption	**Expenditure Using Year 1 Consumption**
Milk	40	168
Butter	40	100
Tea	15	50
Potatoes	16	56
Total	111	374

Base-weighted aggregative index of prices in Year 10 (Year 1 = 100)

$$= \frac{374}{111} \times 100 = \mathbf{336.9}$$

This type of index, where the weights are derived from quantities or values consumed in the base period, is known as a **Laspeyres index**, after the 19th-century economist of that name.

The main defect of a Laspeyres index is that the weights become out-of-date as the pattern of demand changes. A Laspeyres index tends to **overstate** the change in prices, as it takes no account of the fall in consumption when prices rise.

(b) The alternative method is to regard Year 10 as typical and to work all the figures as before, except that this time assume Year 10 consumption in Year 1.

This index is called the **current-weighted aggregative index**. For our example we have:

Table 10.5

Item	Year 1	Year 10
	Expenditure Using Year 10 Consumption	**Expenditure Using Year 10 Consumption**
Milk	25	105
Butter	20	50
Tea	7.5	25
Potatoes	12	42
Total	64.5	222

Current-weighted aggregative index of prices in Year 10 (Year 1 = 100)

$$= \frac{222}{64.5} \times 100 = \mathbf{344.2}$$

This type of index, where the weights are derived from quantities or values consumed in the current period, is known as a **Paasche index** after the 19th-century economist of that name.

The main defect of a Paasche index is that new weights have to be ascertained each time the index is calculated, and this involves time-consuming and expensive survey work. A Paasche index tends to **understate** the changes in prices, as most people tend to buy less of those commodities which have gone up in price.

Weighted Price-relative Index Numbers

We can also form base-weighted or current-weighted price-relative index numbers. As before, we work out the price relatives for each commodity and as we now want to take into account the relative importance of each item in the family budget, we use as weight the actual **expenditure** on each item. The expenditure is used rather than the quantities consumed, to avoid exaggeration of variations arising from the change in consumption pattern rather than the change in price.

(a) Base-weighted Price-relative Index Number (Laspeyres)

Table 10.6

Item	**Price Relative**	**Expenditure in Year 1 (Weight)** *(pence)*	**Price Relative × Weight** *(pence)*
Milk	420.0	40	16,800
Butter	250.0	40	10,000
Tea	333.3	15	5,000
Potatoes	350.0	16	5,600
	Total	111	37,400

Base-weighted price-relative index for Year 10 (Year 1 = 100)

$$= \frac{\sum(\text{Price relative} \times \text{Weight})}{\sum \text{Weights}} = \frac{37{,}400}{111} = \mathbf{336.9}$$

(b) Current-weighted Price-relative Index Number (Paasche)

Table 10.7

Item	**Price Relative**	**Expenditure in Year 1 (Weight)** *(pence)*	**Price Relative × Weight** *(pence)*
Milk	420.0	105	44,100
Butter	250.0	50	12,500
Tea	333.3	25	8,333
Potatoes	350.0	42	14,700
	Total	222	79,633

Current-weighted price-relative index for Year 10 (Year 1 = 100)

$$= \frac{\sum(\text{Price relative} \times \text{Weight})}{\sum \text{Weights}} = \frac{79{,}633}{222} = \mathbf{358.7}$$

D. FISHER'S IDEAL INDEX

The American economist Irving Fisher made a major study of index numbers in the early 1920s. He proposed a number of tests, which could be applied to decide if an index number was acceptable or not. One of these was called the factor reversal test. This states that if the prices and quantities used in an index number are exchanged, then the product of the two index numbers should be an index of total expenditure.

In practice, none of the commonly used index numbers can satisfy this test; however, an index number designed by Fisher and called by him, the "ideal index" does meet the test. It also meets another test of a similar nature called the time reversal test. This index is found using the formula:

$$\text{Index number} = \sqrt{\frac{\Sigma(p_1q_0)}{\Sigma(p_0q_0)} \cdot \frac{\Sigma(p_1q_1)}{\Sigma(p_0q_1)}} \times 100$$

It is in fact the geometric mean of the Laspeyres and Paasche index numbers.

This index has not been widely used, not least because it requires much heavier computation than many other index numbers, and cannot be easily "explained" in the way that an aggregative index number can. Although the computations could now be carried out without trouble using a computer, the ideal index has not found any more favour than it did when it was first suggested.

E. FORMULAE

It will be useful at this stage to summarise our results so far by using formulae. We use the normal notation:

p_o = base year price

p_1 = current year price

q_o = base year quantity

q_1 = current year quantity

n = number of commodities considered

We have the following results:

- Price relative for current year $= \dfrac{\text{Current price}}{\text{Base price}} \times 100$

$$= \frac{p_1}{p_0} \times 100 \qquad (1)$$

- Simple aggregative price index $= \dfrac{\Sigma\text{Current price}}{\Sigma\text{Base price}} \times 100$

$$= \frac{\Sigma p_1}{\Sigma p_0} \times 100 \qquad (2)$$

- Simple mean of price relatives $= \dfrac{\Sigma\text{Price relatives}}{n}$ *(3)*

- Base-weighted aggregative price index (Laspeyres)

$$= \frac{\Sigma(\text{Current price} \times \text{Base quantity})}{\Sigma(\text{Base price} \times \text{Base quantity})} \times 100$$

$$= \frac{\Sigma(p_1q_0)}{\Sigma(p_0q_0)} \times 100 \qquad (4)$$

- Current-weighted aggregative price index (Paasche)

$$= \frac{\sum(\text{Current price} \times \text{Current quantity})}{\sum(\text{Base price} \times \text{Current quantity})} \times 100$$

$$= \frac{\sum(p_1 q_1)}{\sum(p_0 q_1)} \times 100 \qquad (5)$$

- Weighted price-relative index $= \dfrac{\sum(\text{Price relatives} \times \text{Weight})}{\sum \text{Weight}}$ *(6)*

- Ideal price index (Fisher's) $= \sqrt{\dfrac{\sum(p_1 q_0)}{\sum(p_0 q_0)} \cdot \dfrac{\sum(p_1 q_1)}{\sum(p_0 q_1)}} \times 100$ *(7)*

In (6), for a base-weighted price-relative index use (Base price × Base quantity) as the weight. And, for a current-weighted price-relative index, use (Current price × Current quantity) as the weight.

In trying to remember these it is probably simplest to memorise the price-relative, Laspeyres and Paasche formulae and to deduce the others from their descriptive names.

F. QUANTITY OR VOLUME INDEX NUMBERS

You must not think that we are always concerned with **price** indices. Often we are interested in **volume** or **quantity** indices as, for instance, in the Index of Industrial Production which seeks to measure the changes in volume of output in a whole range of industries over a period of time. We can calculate such quantity index numbers in exactly the same sort of way as we dealt with the price indices, for example:

- Quantity relative of a commodity in current year relative to base year

$$= \frac{q_1}{q_0} \times 100$$

- Base-weighted aggregative quantity index (Laspeyres)

$$= \frac{\sum(q_1 p_0)}{\sum(q_0 p_0)} \times 100$$

- Base-weighted quantity-relative index (Paasche)

$$= \frac{\sum\left(\frac{q_1}{q_0} \times 100\right)(p_0 q_0)}{\sum(p_0 q_0)}$$

NB There is no need to memorise these as they are really the same formulae with quantity substituted for price.

Notes

(a) The **price** of a commodity is now used as the weight for an aggregative quantity index and the **expenditure** on that commodity is used as the weight for a quantity-relative index.

(b) It is usual, if we are considering the situation from a producer's point of view rather than the consumer's, to call the index numbers **volume** indices and $\sum(p_0 q_0)$, for example, will be the total **value** of production in the base year.

(c) Remember that for any commodity at any one time:

Value = Price × Volume (producer's view)

Expenditure = Price × Quantity (consumer's view)

Worked Example

Table 10.8 shows UK imports of board from Finland. Calculate a base-weighted **price** Laspeyres index for all types of board for Year 3 (Year 1 = 100).

Table 10.8

Type	Year 1		Year 3	
	Quantity (q_o) *(000 tonnes)*	**Value (p_oq_o)** *(£m)*	**Quantity (q_1)** *(000 tonnes)*	**Value (p_1q_1)** *(£m)*
Machine glazed	90	300	180	650
Folding box board	70	250	10	30
Kraft board	180	550	240	650
Woodpulp board	90	250	80	230
Other board	40	100	100	250

As we are asked for a price index, we must first calculate the price per tonne for each type of board using:

$$\text{Value} = \text{Price} \times \text{Quantity} \quad \text{i.e.} \quad \text{Price} = \frac{\text{Value}}{\text{Quantity}}$$

Table 10.9

Year 1	Year 3
Price (p_o) *(£000/tonne)*	**Price (p_1)** *(£000/tonne)*
3.33	3.61
3.57	3.00
3.06	2.71
2.78	2.88
2.50	2.50

We have now to decide whether to use an aggregative index or a price-relative index. We are asked to find a base-weighted index. Interestingly, we should obtain the same answer whichever method we choose. However, there is less calculation involved in this particular example if we choose an aggregative type, so this is the one we shall work first; we will try the other later.

Base-weighted aggregative price index for Year 3 (Year 1 = 100)

$$= \frac{\text{Total value at Year 3 prices and Year 1 quantities}}{\text{Total value at Year 1 prices and Year 1 quantities}} \times 100$$

We have the Year 1 values in column two of Table 10.8 so we need only sum that column to get the denominator of the expression: £1,450 million pounds.

The numerator is the sum of the product of column one (q_0) in Table 10.8 and column two (p_1) in Table 10.9:

Table 10.10

	Value (p_1q_o) at Year 3 prices, *(£m)* Year 1 quantities
	324.9
	210.0
	487.8
	259.2
	100
Total	1,381.9

Index for Year 3 $= \frac{1{,}381.9}{1{,}450} \times 100$ = 95.3 to 1 dec. place

Therefore, there was an overall **decrease** in prices of 4.7% over the period Year 1 to Year 3.

You can check that using the price-relative method gives the same results. You will need a column for the price relatives and a column for the price relatives weighted with the base-year values:

Table 10.11

$\frac{p_1}{p_0} \times 100$	$\frac{p_1}{p_0}(p_0q_0) \times 100$
108.4	32,520
84.0	21,000
88.6	48,730
103.6	25,900
100.0	10,000
Total	138,150

Base-weighted price-relative index for Year 3 (Year 1 = 100)

$$= \frac{138{,}150}{1{,}450} =$$ 95.3 to 1 dec. place as before

You will see that this must be so by simplifying the base-weighted price-relative formula. There is not an equivalent rule for current-weighted indices, though.

You will see that in index number calculations you will have lots of multiplication and division to do. It is time-consuming to have to use logs at every stage, so if you do have a calculator, particularly one with a memory, it will be of great benefit.

G. CHANGING THE INDEX BASE-YEAR

To convert indices from an earlier to a later base year, divide all the indices by the index for the new base year. This is really a variation on the technique of chain-based indices except that we relate to one particular year rather than continuing to roll forward. We also multiply by 100 to regain percentage values.

The following indices have a base year of 1965 = 100:

Table 10.12

Year	1970	1974	1978	1982
Index	115	126	142	165

We will now convert to base year 1970 = 100 by dividing each index by 115 (1970 index) and multiplying by 100. You will immediately notice that the 1970 index becomes 100 as intended:

Table 10.13

Year	1970	1974	1978	1982
Index	100	$\frac{126}{115}\times 100 = 110$	$\frac{142}{115}\times 100 = 123$	$\frac{165}{115}\times 100 = 143$

Also, the 1965 index becomes $\frac{100}{115} \times 100 = 87$.

An Example

Table 10.14 shows the average weekly earnings of male workers (aged 21 and over) during the years 1970-78. Also shown is the value of the RPI for these years with 1962 as base period. Determine the real average weekly earnings over the period 1970-78.

Table 10.14

Year	1970	1971	1972	1973	1974	1975	1976	1977	1978
RPI (1962 = 100)	140.2	153.4	164.3	179.4	208.1	258.5	301.3	349.1	378.0
Earnings (£)	28.05	30.93	35.82	40.92	48.63	59.58	66.97	72.89	83.50

After calculation as above, we obtain:

Table 10.15

Year	1970	1971	1972	1973	1974	1975	1976	1977	1978
RPI (1970 = 100)	100	109.4	117.2	128.0	148.4	184.4	214.9	249.0	269.6
Real Earnings (£)	28.05	28.27	30.56	31.97	32.77	32.31	31.16	29.27	30.97

We thus see that, although from 1970 to 1978 the average weekly earnings had apparently jumped by £55.45, i.e. increased by almost 200%, the real purchasing power had increased by £2.92 or 10%.

With inflation, the public has in recent years become more aware of index numbers, e.g. index-linked pensions, savings, insurance premiums, etc. However, index numbers are of necessity imperfect measures, the values of which can be manipulated by changes in base year or in the weighting system. For pensions, the decision has to be made whether to link them with earnings or with prices, and if with earnings, the earnings of whom: manual workers, all workers, workers in the same industry? With house insurance premiums, is the index used to be based on the estimated market value of the house or on the cost of clearing the site and rebuilding the house? There is increasing discussion on these matters in the press so do be on the look-out for such articles and relate them to your own knowledge of how index numbers are constructed.

H. PRACTICAL PROBLEMS WITH INDEX NUMBERS

In the previous sections you learned about the basic principles involved in the calculation of index numbers. In the rest of this study unit we will deal with the problems which arise when we try to put these principles into practice.

An index number is nothing more than a specialised kind of average, and you must avoid treating it as anything more than that. Index numbers are very useful for indicating changes in the levels of economic activity. But remember that there is no **direct** measure of things such as “cost of living” or “industrial production”; what we have to do is to decide on the things that are relevant and calculate an index. For example, the correct description of the well-known index measuring changes in the cost of living is the Index of Retail Prices.

The main practical points which arise when constructing index numbers are:

- What items to include
- How to get data relating to the items
- What weighting factors to use
- What kind of averaging to use
- What period to choose as the base period

We will now deal with each of these in turn.

What Items to Include

A practical index number is, in fact, a particular sort of average of a number of values, such as prices, export values, etc. An average relates to a collection of items, and therefore we have to consider, in planning an index number, what items to use in the average. The main rule to be followed is that the

items **must be representative of the topic covered by the index**; in an index of prices intended to show the cost of living for middle-class town families, it would be silly to include the price of a Park Lane flat or the wages of a gamekeeper! Within the limits of this rule, every case must be considered on its own merits.

How many items to include is, of course, part of the problem. As many items as practicable should be included, although the labour of calculation each time the index is worked out should not be excessive. In this connection, the use of computers has enabled some indices to be based on hundreds of items and yet still be quickly and accurately calculated on a daily basis.

How to Obtain the Data

Taking an index of prices as our example, what data do we need to collect? We need data to arrive at:

- The figure to be used as the **price** of each item.
- The figure to be used as the **weight** of each item.

If butter is one of the items in an index of prices, then we know that the price at any time may vary between different places, different shops in the same place and different kinds of butter. If our index is being calculated monthly, then we must remember that the price of the items will also vary throughout the month. Clearly then, the price must be some form of average, and since it is impracticable to average all the prices of all the items in all the shops, some form of **sample survey** must be undertaken. We have covered the principles and techniques of sample surveys, and these must be applied when carrying out price surveys for the calculation of index numbers.

The weights used in calculating a price index are usually based on the relative **quantities** of commodities consumed over a certain period or the relative **values** expended on the commodities. Again, the quantities or values are usually determined by sample surveys; the actual weights are not necessarily the absolute results obtained, but some more convenient numbers which are proportional to the results. Remember – it is the **relative** sizes of the weights, not the actual sizes, that matter. Weights of 2, 7, 10 and 6 will give the same results in a weighted average as weights of 4, 14, 20 and 12 since each is **multiplied** by a constant factor, in this case 2, though it would not do to **add** the same number to each – try it and see!

What Weights to Use

We have to choose between base-period weighting and current-period weighting. A system where the weights are based on the quantities or values of commodities consumed in the base period is known as **base-period weighting**. If the weights are derived from the quantities or values consumed in the current period (for which the index is being calculated), the system is known as **current-period weighting**. As we have seen, the two systems also have names which come from those of two 19th-century economists:

- Base-period weighting: Laspeyres index
- Current-period weighting: Paasche index

Theoretical economists argue about the relative merits of these two kinds of index number. There is often not much difference between them, but with higher inflation the differences become larger. The main defect of a Laspeyres index is that the weights become out of date as the pattern of demand changes. The main drawback of a Paasche index is the additional survey and calculation work required each time the index is to be worked out.

Because people tend to buy more of a commodity when it is cheaper and less when it is dearer, prices and weights have an influence on each other. The effect, as you will remember, is that:

- A Laspeyres index tends to overstate the changes in prices because it takes no account of the fall in consumption when prices rise.
- A Paasche index tends to understate the changes in prices as people buy less of those commodities which have gone up most in price.

There are various possible refinements to overcome the difficulty of deciding on current- or base-period weighting. We can use "typical" period weighting using the quantities consumed in some representative period between the base period and the current period. Alternatively we can use the arithmetic mean of the base-period and current-period quantities as weights, giving the **Marshall-Edgeworth Index**. Another possible index, as we have seen, is **Fisher's Ideal Index** which is the geometric mean of the Laspeyres and Paasche indices.

The Kind of Average to Use

We have mentioned several times that an index number is no more than an average of a special kind. So far, we have taken the word "average" to mean "arithmetic mean". But, there are other kinds of average and, in particular, we could have used the **geometric mean**. At this point you might well go back and revise how to calculate geometric means. Then the following example, which uses data from earlier in the study unit, will be easier to follow.

We are going to calculate a base-period weighted geometric mean of price-relative index number using data as shown in Table 10.6.

Table 10.16: Price Relatives

	Year 10 Price Relative *(Yr 1 = 100)*	**Log Price Relative**	**Expenditure Weight (Base Year)** *(Pence)*	**Weight × Log**
Milk	420	2.6232	40	104.928
Butter	250	2.3979	40	95.916
Tea	333.3	2.5229	15	37.8435
Potatoes	350	2.5441	16	40.7056
			111	279.3931

$$\text{Log (Index number)} = \frac{279.3931}{111} = 2.5171$$

$$\text{Index number} = 328.9$$

Notice that this is **less** than the index number of 336.9 obtained earlier by using the weighted arithmetic mean. This is a general property of a geometric mean – it is always **less** than the arithmetic mean of the same figures. You should note that we have not calculated the geometric mean of the weighted price relatives directly, but instead we worked out a weighted arithmetic mean of the logs of the price relatives and then took the antilog of the answer. This is obviously more trouble to calculate than a weighted arithmetic mean of the price relatives.

An advantage claimed for the geometric mean is that it is less influenced by occasional extreme items than is the arithmetic mean. This is true, but the advantage is slight when large numbers of items are involved. The result is that, in practice, there is not much to choose between the two sorts of average, and the arithmetic mean is most often used. This does not deter examiners from setting questions

about geometric means, but do not calculate a geometric type index unless **specifically** requested. One special point to watch when using a weighted geometric mean is that it is applicable only to a price-relative type index and **not** an aggregative type index.

The one commonly known index that is calculated in this way is the Financial Times Industrial Ordinary Share Index. The index is based on the share prices of only 30 large companies. The effect of using a geometric mean is to stop a large fluctuation in the value of the index when only one of these shares shows a sudden rise or fall in price.

What Period to Use as Base

The main point about the base period is that it should, in some way, represent a standard or normal period. It is an economics problem to decide upon a standard period rather than a statistical one, and so we will not dwell too closely on it. There are, however, a few points about base periods that we do have to deal with.

Firstly, consider how long the base period should be. As with so many questions of this kind, there is no one correct answer. Sometimes you will see an index number the base period of which is one day (e.g. 14th July 1981 = 100), sometimes one month (e.g. July 1981 = 100) and sometimes one year (e.g. 1981 = 100). The general principle behind the decision about length of base period is that the period should be long enough to enable a good average to be taken, without including exceptional times such as booms or depressions. When a long base period is decided upon, it should not exceed a year, and it should preferably be exactly one year so as to include each season once, and once only. We had a similar consideration when dealing with moving averages.

One difficulty about base periods is that, by the very nature of the changing conditions that we are trying to measure, they eventually become out-of-date and unrepresentative. A base period in which wax candles were in wide use could hardly be representative of life in the age of electric light! There are three customary ways of counteracting (to some extent) this tendency to become out of date:

(a) Using a Laspeyres index, bring the base period up to date at regular intervals.

When the base period of an index number is changed, in order to make direct comparisons with historical values, we may need to splice two index series. For example, we may be given:

	1985	1986	1987	1988	1989	1990	1991	1992
Price index (1980 = 100)	123	135	147	160	182	193	-	-
Price index (1990 – 100)	-	-	-	-	-	100	110	115

We thus have two separate index series for the same commodities, but after 1990 the base year has been updated to 1990 from 1980. To enable us to compare all the price indices directly, we must recalculate the values for 1991 and 1992 based on 1980.

$$\text{Price index for 1991 (1980=100)} = \frac{\text{Price index for 1991 (1990=100)}}{\text{Price index for 1990 (1990=100)}} \times \text{Price index for 1990 (1980 = 100)}$$

$$= \frac{110}{100} \times 193 = 212.3$$

$$\text{Price index for 1992 (1980=100)} = \frac{\text{Price index for 1992 (1990=100)}}{\text{Price index for 1990 (1990=100)}} \times \text{Price index for 1990 (1980 = 100)}$$

$$= \frac{115}{100} \times 193 = 222.0$$

(b) Using a Paasche index, the weights are automatically brought up to date each time.

(c) Use what is called the **chain-base** method. Here the index for the current period is based on the last (i.e. the immediately preceding) period. For example, if we are calculating an index for Yr 3, we use Yr 2 as the base year; then, when we come to calculate the index for Yr 4, we use Yr 3 as the base year; and so on. This system has the advantage that it is always up to date and it is easy to introduce new items or delete old ones gradually without much upset to the reliability of the index. Its disadvantage is that it cannot be used for making comparisons over long periods of time, as we are simply comparing each year with the immediately preceding year.

If we do need to make long-term comparisons when a chain-base index number is in use, then it is necessary to convert the indices from a **chain base** to a **fixed base**. Such conversions are a favourite topic with some examiners. The method of working is shown in the following two examples.

Example 1

The indices for the years 9, 10, 11, (Yr 8 as base = 100) are:

Year 9	104
10	104
11	109

We are required to convert these to a chain-base set of indices. The Yr 8 index remains the same at 100; the Yr 9 index (based on Yr 8) is still 104; the Yr 10 index (based on Yr 9) is 100 because the two years have the same index; the Yr 11 index (based on Yr 10) is (109 × 100)/104 = 105.

Example 2

The following indices were arrived at by a chain-base method. Convert them to Yr 7 as a fixed base.

Year 7	100
8	106
9	110
10	95
11	100

The Yr 7 index remains at 100; the Yr 8 index (based on Yr 7) remains at 106; the Yr 9 index (based on Yr 8) is 110 and therefore the Yr 9 index will always be 110/100 of the Yr 8 index,

no matter what base is used. Now, the Yr 8 index (based on Yr 7) is 106, and so the Yr 9 index (based on Yr 7) is (110 × 106)/100 = 116.6. Similarly, the Yr 10 index will be 95/100 of the

Yr 9 index, no matter what base is used. And so the Yr 10 index (based on Yr 7) is (95 × 116.6)/100 = 110.8. The Yr 11 index (based on Yr 10) is 100 and therefore there is no change from Yr 10 to Yr 11; the Yr 11 index (based on Yr 7) is consequently the same as the Yr 10 index (based on Yr 7), namely 110.8.

I. CRITERIA FOR A GOOD INDEX

Easy to Understand

An ordinary person can appreciate the influence of the average change in the cost of a fixed bundle of goods on the cost of living or the cost of production, so this type of index is widely used. Thus the Laspeyres and Paasche indices and the means of price relatives are easily understood. However, the Fisher Ideal Index and others that use geometric means are of more theoretical interest and are more difficult concepts to grasp.

Reliable

Reliability is difficult to define in connection with index numbers as it need not imply a high level of accuracy. The ability to reflect realistically the type of change which is taking place is more important than numerical accuracy. Reliability depends mainly on the weighting used, and so is affected to some extent by the experience of the person constructing the index. In particular any change in the system of weighting should not generate a major change in the index. Thus a chain-based index is likely to be more reliable than a fixed-base index.

Cost-effective

It is not practicable to spend a considerable amount of time and money in constructing an index that will appeal to only a few specialists. An index must be produced for the minimum cost and appeal to the maximum number of people. In assessing the effectiveness of an index, you must ensure that the delay between the end of the relevant time period and the publication of the index is a minimum, i.e. the initial analysis of any new survey material must be kept to a minimum. In this context the Laspeyres index has the advantage over the Paasche index.

Base Year Up-to-date

Base-weighted indices are the most popular, so the bundle of goods in the base year must be similar to that of the current year. Changes in the bundle take place gradually under normal conditions, so a regular updating of the base year will allow for these changes. For example, indices formed from a Family Expenditure Survey need to be continually updated to take account of changing patterns of family expenditure.

Indices are used to compare conditions for the same bundles in different geographical areas or different bundles in the same area. These comparisons will be valid only if all the indices have the same base year, which should be a recent year. The method of updating an index was mentioned earlier but here is a further example. The following indices have a base year of 1975 = 100:

Year	1980	1984	1988	1992
Index	115	126	142	165

To convert to indices having base year 1980 = 100, divide each index by the index for 1980, giving:

Year	1980	1984	1988	1992
Index	100	$\frac{126}{115} \times 100 = 110$	$\frac{142}{115} \times 100 = 123$	$\frac{165}{115} \times 100 = 143$

(Note that after conversion the index for 1975 becomes

$$\frac{100}{115} \times 100 = 87 \quad (1980 = 100))$$

J. INDEX NUMBERS IN USE

There are several index numbers published and used in the UK. In this section we will look at three well-known ones that are of most interest to the public.

Retail Price Index

This is compiled by the Department of Education and Employment, and its purpose is to measure the relative change, every month, of a bundle of goods which represents the expenditure of an average family in that month. For each article or service included in the bundle, the current price is expressed as a relative of the price at the base date. Currently the base year is 1987.

Each article is weighted with weights derived from a government survey called the Family Expenditure Survey. The bundle includes such items as food, housing, fuel and light, household goods, clothing, transport and services. Prices are collected once a month from a number of retail outlets randomly selected throughout the country.

This index is widely used in settling wage claims, but because of the omission of certain items, notably income tax payments, it should not be interpreted as a cost of living index. In an effort to overcome this problem a new index called the Tax and Price Index was introduced in 1979.

Index of Wholesale Prices

This index, produced monthly, is designed to measure the changes in the costs of items to British industry and the prices of the items that are produced by the various sectors of industry.

The index shows the price movements of a large number of materials and products which are combined into a number of broad sectors as defined in the Standard Industrial Classification. In building up the index, weights which have been derived from the Census of Production are used. The index is an extremely useful source of market intelligence and very useful to industry.

Index of Industrial Production

This index is intended to provide a general measure of the monthly changes in the **volume** of industrial production in the United Kingdom. Industrial production can consist of mining, manufacturing and construction work as well as the production of such commodities as gas and

electricity. It includes the production of goods intended for the home and the export markets. The index is weighted by figures derived from the Census of Production.

K. CHOICE OF INDEX NUMBER

It is difficult to state any guidelines for this choice as there are no hard-and-fast rules. Your choice depends on experience, the data available and the purpose for which the index is needed. (In an examination question you will usually find that the examiner names the type of index number involved or asks for a choice between two or more stated types.)

In deciding which index to use, you should consider the following questions:

- Is current- or base-period weighting more appropriate?
- Is the necessary data readily available?
- What is the index to be used for?

The answers to these questions will usually lead you to make a choice between a Laspeyres or Paasche index.

Study Unit 11

Probability

Contents *Page*

(continued over)

A. WHAT IS PROBABILITY?

Chance and Uncertainty

"Probability" is one of those ideas about which we all have some notion but which are not very definite. Initially, we will not spend our time trying to get an exact definition, but will confine ourselves to the task of grasping the idea generally and seeing how it can be used. Other words which convey much the same idea are "chance" and "likelihood". Just as there is a scale of, say, temperature, because some things are hotter than others, so there is a scale of probability because some things are probable than others. Snow is **more** likely to fall in winter than in summer; a healthy person has **more chance** of surviving an attack of influenza than an unhealthy person.

There is, note, some uncertainty in these matters. Most things in real life are uncertain to some degree or other, and it is for this reason that the **theory of probability** is of great **practical** value. It is the branch of mathematics which deals specifically with matters of uncertainty. For the purpose of learning the theory, it is necessary to start with simple things like coin-tossing and dice-throwing, which may seem a bit remote from business and industrial life, but which will help you understand the more practical applications.

Choosing a Scale

First of all, let us see if we can introduce some precision into our vague ideas of probability. Some things are very unlikely to happen. We may say that they are very improbable, or that they have a very low probability of happening. Some things are so improbable that they are absolutely impossible. Their probability is so low that it can be considered to be zero. This immediately suggests to us that we can give a numerical value to at least one point on the scale of probabilities. For **impossible things**, such as pigs flying unaided, the **probability is zero**. By similarly considering things which are absolutely certain to occur, we can fix the top end of the scale at 100%. For absolutely certain things, such as dying, the probability is 100%. It is often more convenient to talk in terms of proportions than percentages, and so we say, for absolute certainty the probability is 1. In mathematical subjects we use symbols rather than words, so we indicate probability by the letter p and we say that the probability scale runs from $p = 0$ to $p = 1$. (Never greater than 1.)

Degrees of Probability

For things which may or may not happen, the probability obviously lies somewhere between 0 and 1.

First, consider tossing a coin. When it falls, it will show either "heads" or "tails". As the coin is a fairly symmetrical object and as we know no reason why it should fall one way rather than the other, then we feel intuitively that there is an equal chance (or, as we sometimes say, a 50/50 chance) that it will fall each way. For a situation of equal chance, the probability must lie exactly halfway along the scale, and so the probability that a coin will fall heads is 1/2 (or $p = 0.5$). For tails, p is also 0.5.

Next, consider rolling a 6-sided die as used in gambling games. Here again there is a fairly symmetrical object and we know of no special reason why one side should fall uppermost more than any other. In this case there is a 1 in 6 chance that any specified face will fall uppermost, since there are 6 faces on a cube. So the probability for any one face is 1/6 ($p = 0.167$).

As a third and final example, imagine a box containing 100 beads of which 23 are black and 77 are white. If we pick one bead out of the box at random (blindfold and with the box well shaken up)

what is the probability that we will draw a black bead? We have 23 chances out of 100, so the probability is $\frac{23}{100}$ (or p = 0.23).

Probabilities of this kind, where we can assess them from prior knowledge of the situation, are called **a priori** probabilities.

In many cases in real life it is not possible to assess a priori probabilities, and so we must look for some other method. What is the probability that a certain drug will cure a person of a specific disease? What is the probability that a bus will complete its journey without having picked up a specified number of passengers? These are the types of probabilities that cannot be assessed a priori. In such cases we have to resort to experiment. We count the **relative frequency** with which an event occurs, and we call that the probability. In the drug example, we count the number of cured patients as a proportion of the total number of **treated** patients. The probability of cure is then taken to be:

$$p = \frac{\text{Number of patients cured}}{\text{Number of patients treated}}$$

In a case like this, the value we get is only an **estimate**. If we have more patients, we get a better estimate. This means that there is the problem of how many events to count before the probability can be estimated accurately. The problem is the same as that faced in sample surveys. Probabilities assessed in this way, as observed proportions or relative frequencies, are called **empirical probabilities**.

B. TWO LAWS OF PROBABILITY

Addition Law for Mutually Exclusive Events

If a coin is tossed, the probability that it will fall heads is 0.5. The probability that it will fall tails is also 0.5. It is certain to fall on one side or the other, so the probability that it will fall either heads or tails is 1. This is, of course, the **sum** of the two separate probabilities of 0.5. This is an example of the **Addition Law** of probability. We state the addition law as:

> "The probability that one or other of several mutually exclusive events will occur, is the sum of the probabilities of the several separate events."

Note the expression "mutually exclusive". This law of probability applies only in cases where the occurrence of one event **excludes** the possibility of any of the others. We shall see later how to modify the addition law when events are not mutually exclusive.

Heads automatically excludes the possibility of tails. On the throw of a die, a six excludes all other possibilities. In fact, all the sides of a die are mutually exclusive; the occurrence of any one of them as the top face necessarily excludes all the others.

Example

What is the probability that when a die is thrown, the uppermost face will be either a two or a three?

The probability that it will be two is $1/6$

The probability that it will be three is $1/6$.

Because the two, the three, and all the other faces are mutually exclusive, we can use the addition law to get the answer, which is $2/6$, i.e. $1/6 + 1/6$, or $1/3$.

You may find it helpful to remember that we use the addition law when we are asking for a probability in an either/or situation.

Complementary Events

An event either occurs or does not occur, i.e. we are **certain** that one or other of these two situations holds. Thus the probability of an event occurring plus the probability of the event not occurring must add up to one, i.e.:

$$p(A) + p(\text{not } A) = \ldots\ldots \qquad (a)$$

where p(A) stands for the probability of event A occurring.

A^1 or $\overline{A}$ is often used as a symbol for "not A". "A" and "not A" are referred to as complementary events. The relationship (a) is very useful as it is often easier to find the probability of an event not occurring than to find the probability that it does occur. Using (a) we can always find p(A) by subtracting p(not A) from one.

Example

What is the probability of a score greater than one when one dice is thrown once?

Method 1

Probability of a score greater than 1 = p(score > 1)

= p(score not = 1)

= 1 – p(score = 1), using (a)

$$= 1 - \frac{1}{6} = \frac{5}{6}$$

Method 2

Probability of a score greater than 1 = p(score > 1)

= p(2 or 3 or 4 or 5 or 6)

= p(2) + p(3) + p(4) + p(5) + p(6) using addition law

$$= \frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6} = \frac{5}{6} \text{ as before}$$

Multiplication Law for Independent Events

Suppose, now, that the two coins are tossed. The probability that the first coin will show heads is 0.5, and the probability that the second coin will show heads is also 0.5. But what is the probability that **both** coins will show heads? We cannot use the addition law because the two events are not mutually exclusive – a first coin landing heads does not prevent a second coin landing heads. If you did try to apply the addition law, you would not get 0.5 + 0.5, which is 1.0, meaning that two heads are bound to result – and that is nonsense! We can get the correct answer by listing all the possible results of tossing two coins:

First Coin	Second Coin
Heads	Heads
Heads	Tails
Tails	Heads
Tails	Tails

There are four possible results, one of which is the result we are seeking. We have no reason to suppose that there is anything to favour any particular result, and so the probability that both coins will show heads is 1 in 4 (0.25), i.e. p = 0.25. This is an example of the **multiplication law** of probability, which states:

> "The probability of the **combined occurrence** of two or more independent events is the product of the probabilities of the separate events."

Note the word "independent". Events are said to be **independent** when the probability of either of them is not affected by the occurrence or non-occurrence of the other. In our coin example, the way one coin falls has absolutely no effect on the way the other one will fall, and so the two events are independent. We may therefore apply the multiplication law, which gives the probability that both heads will occur as 0.5×0.5, i.e. 0.25.

Example

Two dice are thrown separately. What is the probability that both dice will show a five uppermost?

The events in this case are the showing of a five. The dice have no effect on one another, and so the events are independent.

The probability that the first dice shows five is 1/6. The probability that the second dice shows five is 1/6.

The probability that both dice show five is $1/6 \times 1/6$, i.e. 1/36.

More Examples

Try these few examples without looking at the answers which follow.

(a) In a box of 200 items taken from a factory production line, there are 25 faulty items. An inspector picks out an item at random. What is the probability that the selected item is not faulty?

(b) In a second box, there are 1,000 items of which 100 are faulty. The inspector picks out an item at random.

  (i) What is the probability that this item is faulty?

  (ii) What is the probability that both items (i.e. one from each box) are not faulty?

(c) From a pack of 52 ordinary playing cards, a card is drawn at random. What is the probability that it is **either** a two **or** a seven?

(d) If 3 coins are thrown, what is the probability that all 3 will show tails?

Answers to Examples

(a) Among the 200 items, 25 are faulty and therefore 175 items must be not faulty. In a random selection of one item, the probability that the item is not faulty is 175/200, which is 7/8.

(b) (i) The probability that the item taken from the second box is faulty is 100/1,000 = 0.1.

  (ii) The probability that the second item is not faulty is:

$$\frac{1{,}000-100}{1{,}000}=\frac{900}{1{,}000}=\frac{9}{10}.$$

The events are independent, so, by multiplication law, the probability is:

$$\frac{7}{8}\times\frac{9}{10}=\frac{63}{80}.$$

(c) In the pack of 52 cards, there are 4 twos and 4 sevens, so altogether there are 8 cards favourable to the event we are considering. The probability is therefore 8/52 (or 2/13). Alternatively you could say that picking a two excludes the possibility of picking a seven, so, by the addition law, the choice is:

$$\frac{4+4}{52}=\frac{8}{52}=\frac{2}{13}$$

(d) The probability of tails is 1/2 for each coin. The throws are independent, and so the probability that all 3 will show tails is, by the multiplication law:

$\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}$ which is $\frac{1}{8}$.

Distinguishing the Laws

Although the above laws of probability are not complicated, you must think carefully and clearly when using them. Remember that events must be **mutually exclusive** before you can use the **addition law**, and they must be **independent** before you can use the **multiplication law.** Another matter about which you must be careful is the listing of equally likely outcomes. Be sure that you list all of them. Earlier we listed the possible results of tossing two coins:

First Coin	Second Coin
Heads	Heads
Tails	Heads
Heads	Tails
Tails	Tails

There are 4 equally likely outcomes. Do not make the mistake of saying, for example, that there are only 2 outcomes (both heads or not both heads), you must list all the possible outcomes. (In this case "not both heads" can result in 3 different ways, so the probability of this result will be higher than "both heads".)

In this example the probability that there will be one heads and one tails (heads – tails or tails – heads) is 0.5. This is a case of the addition law at work, the probability of heads – tails (¼) **plus** the probability of tails – heads (¼). Putting it another way, the probability of different faces is equal to the probability of the same faces – in both cases ½.

C. PERMUTATIONS

Listing Possible Results

When we deal with simple things like throwing a few coins or dice, it is easy to make a list of all the possible results. In more complicated cases, it would be impracticable to write out the whole list. Fortunately there are some simple mathematical methods for calculating the number of possible results. These methods are referred to as **permutations** and **combinations**.

What is a Permutation?

The word permutation means a particular sequence or order of arrangement of things. For example, BAC and CBA are both permutations of the first 3 letters of the alphabet. Permutation problems are concerned with the number of possible sequences into which things can be arranged. There is a basic principle governing such problems:

> "If one operation can be done in m ways, and if a second operation can be done in n ways, then the two operations can be done in succession in m times n different ways."

For the purposes of this course we do not need to prove it but only to know and understand it thoroughly. The principle can be extended to any number of operations greater than 2.

Example

There are 3 different coloured buses (red, yellow and green) which run between 2 places. If I want to use a different coloured bus for each direction, in how many different ways can I make the double journey?

Applying the basic principle, we see that the first part of the trip can be done in 3 ways (red, yellow and green), while the second part of the trip can be done in only 2 ways (excluding the colour already used). Thus the total number of different possible ways is $3 \times 2 = 6$.

It would be a good idea at this stage for you to try to write out the list of the 6 possible alternatives.

The principle can be applied to the coin example. The first coin can fall in 2 possible ways (heads or tails) and the second coin can fall in 2 possible ways (heads or tails) and so the total number of different possible ways is $2 \times 2 = 4$.

Permutations of Different Items

If we have a group of **different** items, we can now calculate the total number of permutations quite easily. Suppose there are 4 different things which we arrange in a row. Any one can be put first, so the first place can be filled in 4 different ways. After the first place has been filled, only 3 things remain, and so the second place can be filled in 3 possible ways. Then the third place can be filled in 2 possible ways and the fourth place in only 1 way. The basic principle tells us that the total number of different arrangements is therefore $4 \times 3 \times 2 \times 1 = 24$. If there had been 5 items, the total number of permutations would have been $5 \times 4 \times 3 \times 2 \times 1 = 120$. You can see the pattern – you simply multiply all the numbers down to 1. There is a special name for this continued product from a given number down to 1; it is called a **factorial**. Thus:

2×1 is factorial 2, and it is equal to 2

$3 \times 2 \times 1$ is factorial 3, and it is equal to 6

$4 \times 3 \times 2 \times 1$ is factorial 4, and it is equal to 24

and so on.

A special sign is used for a factorial, an exclamation mark, so:

$2! = 2 \times 1$

$3! = 3 \times 2 \times 1$

$7! = 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$

and so on.

We can now say that the rule for calculating the total number of permutations of n things is n!.

Permutations of Some Items Only

Sometimes we need permutations of only **some** of the items at our disposal. The calculation is very similar, but we stop after the appropriate number of factors instead of taking the multiplication right down to 1. Thus, if we wish we know the number of possible arrangements of **any** 3 things taken

from a group of 9 things, we calculate (i.e. 504). That is just like 9! except that we stop after 3 factors. We use another symbol for this. The number of permutations of 9 things taken only 3 at a time is written as ${}_9P_3$ or ${}_9P3$, and read as nine, P, three.

The rule is quite general, and we speak of the number of permutations of n things taken r at a time, as ${}_nP_r$. To calculate the value of ${}_nP_r$ we start to work out n! but stop when we have done r factors. Thus ${}_9P_4 = 9 \times 8 \times 7 \times 9$ which is 3,024, and ${}_{100}P_2 = 100 \times 99$ which is 9,900. An alternative method can be seen by taking the ${}_9P_3$ example again and writing:

$$ {}_9P_3 = 9 \times 8 \times 7 = \frac{9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{6 \times 5 \times 4 \times 3 \times 2 \times 1} = \frac{9!}{6!} $$

You see that the 6 is 9 minus 3, so, for all values of n and r, we can put:

$$ {}_nP_r = \frac{n!}{(n - r)!} $$

Note that 0! is defined as equal to 1, **not** zero:

$$ 0! = 1 $$

The formula for ${}_nP_n$ is then $= \frac{n!}{(n - n)!} = \frac{n!}{0!} = n!$, just as we would expect.

Permutations Including Similar Items

So far we have assumed that the n items from which the permutations are taken are all different. But what if some items are **identical**? Take, for example, the word ACCOUNT. There are 7 letters and therefore 7! possible permutations of those letters. But some of these permutations will look the same because the two letters C are merely interchanged. So there will be **fewer than** 7! distinguishable permutations. The formula for calculating the number of distinguishable permutations is quite simple, and for this course we do not need to prove it but merely to know it. In the example just given, there are 2 identical letters C, and the total number of distinguishable permutations is:

$$ \frac{7!}{2!}. $$

A more complicated example would be the word STATISTIC, where there are 9 letters altogether, but S occurs twice, T occurs 3 times and I occurs twice. Here the total number of distinguishable permutations is:

$$ \frac{9!}{2! \times 3! \times 2!} $$

From these two examples you should be able to see the pattern of the general rule. The rule says that if we have n items, of which p are alike of one kind, q are alike of another kind and r are alike of yet another kind, then the total number of distinguishable permutations of the n items is:

$$ \frac{n!}{p! \times q! \times r!} $$

and so on if there are more than 3 groups.

Examples

Before we go on to combinations, you should make yourself quite familiar with permutations so that you do not become confused. Make sure you know what they are – not merely how to calculate them. Attempt the following examples:

(a) Find the value of

(i) $\frac{8!}{5!}$

(ii) $\frac{73!}{72!}$

(b) In how many different orders can 6 objects be arranged for checking by an inspector?

(c) In how many different ways can 3 ledgers be submitted to 5 auditors if:

(i) No auditor deals with more than 1 ledger?

(ii) Any auditor may deal with any number of ledgers?

(d) Express in factorials only;

(i) $9 \times 8 \times 7 \times 6$

(ii) $5 \times 6 \times 7$

(e) Write down the formulae for:

(i) 4Pn

(ii) nP4

(iii) 2nPn

(f) In how many ways can 3 dots and 6 commas be arranged in a line?

Answers to Examples

(a) (i) $\frac{8!}{5!} = \frac{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{5 \times 4 \times 3 \times 2 \times 1} = 8 \times 7 \times 6 = 336$

(ii) $\frac{73!}{72!} = \frac{73 \times 72 \times \text{etc. down to} \times 1}{72 \times 71 \times \text{etc. down to} \times 1} = 73$

(b) 6! which is 720.

(c) (i) Any auditor may be chosen to deal with the first ledger, then any one of the remaining 4 may be chosen for the second ledger and so on. The total number of ways is therefore $5 \times 4 \times 3$, which is 60.

(ii) Any 1 of the 5 auditors may deal with the first ledger, any one of the 5 with the second ledger and any 1 of the 5 with the third ledger. The total number of possible allocations of the jobs is therefore $5 \times 5 \times 5$=125.

(d) (i) $9 \times 8 \times 7 \times 6 = \frac{9!}{5!}$

(ii) $5 \times 6 \times 7 = \frac{7!}{4!}$

Notice that any product which is a run of consecutive whole numbers can always be written as the quotient of two factorials.

(e) (i) $_4P_n = \dfrac{4!}{(4-n)!}$

(ii) $_nP_4 = \dfrac{n!}{(n-4)!}$

(iii) $_{2n}P_n = \dfrac{(2n)!}{(2n-n)!} = \dfrac{(2n)!}{n!}$

(f) Because dots all look alike and commas all look alike, some of the 9! possible arrangements of the 9 symbols will not be distinguishable. The number of distinguishable arrangements is:

$$\frac{9!}{3!6!} = \frac{9\times8\times7}{3\times2\times1} = 84$$

D. COMBINATIONS

What is a Combination?

When dealing with permutations, we are concerned principally with the order or **sequence** in which things occur. Other problems occur in which we need to calculate the number of groups of a certain size **irrespective of their sequence**.

For example, consider the question of how many possible ways there are of choosing a football team of 11 men from a group of 15 club members. Here there is no question of putting the team in a particular sequence, but merely the problem of finding the number of possible different teams. The problem is one of finding the number of **combinations** of 15 things taken 11 at a time. The symbol for this is $_{15}C_{11}$ or $^{15}C_{11}$.

Sometimes you may come across a different symbol, $\begin{bmatrix}15\\11\end{bmatrix}$.

However, this means the same thing and they are all read as fifteen, C, eleven.

The number of combinations of n things taken r at a time is obviously less than the number of permutations, **because each combination can have r! different arrangements**. That gives us the clue to finding a formula for $_nC_r$. There must be r! times as many permutations as there are combinations and so:

$$\begin{bmatrix}n\\r\end{bmatrix} = {}_nP_r \div r! = \frac{n!}{r!(n-r)!}$$

Example

A factory has 5 identical production lines. During a certain period, there are only enough orders to keep 3 lines working. How many different ways of choosing the 3 lines to be worked are available to the factory manager?

The manager's problem, expressed mathematically, is to find the number of combinations of 5 things taken 3 at a time, i.e.:

$$\begin{bmatrix}5\\3\end{bmatrix} = \frac{5!}{3!(5-3)!} = \frac{5\times4\times3\times2\times1}{3\times2\times1\times2\times1}$$

Cancelling out as many factors as possible gives:

$$\begin{bmatrix}5\\3\end{bmatrix} = 5 \times 2 = 10$$

As in previous examples, it would be a good idea for you to verify this result by writing out all the 10 possible combinations.

Equivalent Combinations

If we select 4 items from 7 items, we leave 3 items behind. For every set of 4 that we choose there must be a set of 3 that we do **not** choose. It follows that:

the number of combinations we can choose, $\begin{bmatrix}7\\4\end{bmatrix}$, must equal

the number of combinations that we do not choose, $\begin{bmatrix}7\\3\end{bmatrix}$.

By applying the reasoning to the general case, we get the important fact that:

$$\begin{bmatrix}n\\r\end{bmatrix} = \begin{bmatrix}n\\n-r\end{bmatrix}.$$

This is very useful sometimes when one of the calculations may be much easier than the other. All that this means is that you should cancel out the **larger** of the two factors in the denominator, as an example will show:

Example

(a) Find the value of $\begin{bmatrix}8\\5\end{bmatrix}$.

$$\begin{bmatrix}8\\5\end{bmatrix} = \frac{8!}{5!\ 3!}$$

Cancel out the 5! and we get

$$\frac{8\times7\times6}{3\times2\times1} = 56$$

(b) Find the value of $\begin{bmatrix}8\\3\end{bmatrix}$.

$$\begin{bmatrix}8\\3\end{bmatrix} = \frac{8!}{3!\ 5!} = 56$$

which is mathematically the same as above.

Note

$\begin{bmatrix} n \\ r \end{bmatrix}$ is also termed a **binomial coefficient**.

Complications and Restrictions

In some practical problems it is necessary to calculate the number of combinations, subject to certain restrictions. There are no general rules to cover such cases, and so each one must be thought out carefully. The arithmetic is no more difficult than that which we have already done. Here are some examples, which you should study carefully:

(a) There are 7 clerks in an office. A team of 4 clerks is needed for a special checking job. In how many ways can the team be made up if the longest-serving clerk **must** be included?

If the longest-serving clerk is put in the team, that leaves us with the problem of finding another 3 clerks out of 6. So the answer will be:

$\begin{bmatrix} 6 \\ 3 \end{bmatrix}$, which is 20.

(b) In how many ways can the team be made up if the only restriction is that one particular clerk is not available for this job?

In this case, we have to find the team of 4 from only 6 clerks. The answer is therefore

$\begin{bmatrix} 6 \\ 4 \end{bmatrix}$, which is 15.

Notice that a careful consideration of the restrictions usually enables you to formulate a slightly different question which can then be answered by the usual kind of calculation.

Examples

Now try these examples before looking at the answers.

(a) In how many ways can a committee of 5 be chosen from 9 candidates so as to include both the youngest and the oldest candidate?

(b) In how many ways can the committee be formed if the youngest candidate is excluded?

(c) In a bin of 50 items from a cutting machine there are 17 defective items. Calculate the number of possible different samples of 5 which contain no defectives.

(d) Calculate, using the data of the above example, the number of possible samples of 5 items from the cutting machine which contain exactly 1 defective item.

Answers to Examples

(a) If the youngest and the oldest must both be included, then we are left with the problem of choosing 3 out of 7 which is:

$$\begin{bmatrix} 7 \\ 3 \end{bmatrix} = \frac{7!}{3!\ 4!} = \frac{7 \times 6 \times 5}{3 \times 2} = 35$$

(b) If the youngest is excluded, we have the problem of choosing 5 from 8:

$$\begin{bmatrix} 8 \\ 5 \end{bmatrix} = \frac{8!}{5!\ 3!} = \frac{8 \times 7 \times 6}{3 \times 2} = 56$$

(c) If there are 17 defective items, then there must be 33 good items. The number of possible samples of 5 items (all good) is therefore:

$$\begin{bmatrix} 33 \\ 5 \end{bmatrix} = \frac{33!}{5!\ 28!} = \frac{33\times 32\times 31\times 30\times 29}{5\times 4\times 3\times 2\times 1} = 237{,}336$$

(d) There are 17 possible ways in which the 1 defective item may be chosen. For the remainder of the sample there are 33C4 ways of choosing the 4 good items. The total number of possible samples is therefore:

$17 \times \begin{bmatrix} 33 \\ 4 \end{bmatrix}$, which is 695,640.

E. CONDITIONAL PROBABILITY

When we dealt with the multiplication law of probability, we insisted that the events must be independent. Now we can apply the same law to non-independent events provided that we make due allowance for the dependence. The method is best explained by first working an example.

A box contains 13 red beads and 7 white beads. If we make a random selection of 1 bead from the box, then the probability that it will be a white bead is 7/20. Now suppose that this first bead turns out to be a white one. We put it on one side and make another random selection of 1 bead. What is the probability that this second bead will be white? It is not 7/20 this time, because there are only 6 white beads left in the box. There are still 13 red ones, making 19 altogether, so the probability that the second one is white is 6/19.

We can now apply the multiplication law to find out the probability of both the first and second beads being white. It is:

$$\left[\frac{7}{20}\right]\times\left[\frac{6}{19}\right]=\left[\frac{42}{380}\right]=\left[\frac{21}{190}\right]$$

The probability of 6/19 for the second selection is called the **conditional probability** for the second event, on the assumption that the first event has happened. The more general form of the multiplication law is therefore:

> "The probability of the combined occurrence of two events, A and B, is the product of the probability of A and the **conditional probability** of B on the assumption that A has happened."

The law can be applied to a series of more than two events if care is taken to assess the successive conditional probabilities accurately.

Examples

(a) With the same box of beads as used in the previous paragraphs, what is the probability that the first 2 beads drawn from the box will be red?

(b) With the same box again, what is the probability that the first 2 beads will be a red followed by a white?

(c) A bin contains 100 snoggle pins made on a new machine. Of these 100 items, 20 are defective. An inspector draws 5 items at random from the bin. What is the probability that all 5 are not defective? (Don't work out all the arithmetic – just show the numbers and the calculations to be done.)

(d) Show how to do (c), using combinations only, and verify that this gives the same result as before.

Answers to Examples

(a) The probability that the first bead is red is 13/20. If the first bead is red then the probability that the second bead is red is 12/19. The probability that **both** beads will be red is therefore:

$$\frac{13}{20}\times\frac{12}{19}=\frac{156}{380}=\frac{39}{95}$$

(b) The probability that the first bead will be red is 13/20. If it is red, then the probability that the second bead will be **white** is 7/19. The total probability is then:

$$\frac{13}{20}\times\frac{7}{19}=\frac{91}{380}$$

(c) The probability that the first one is not defective is 80/100. If the first one is good, then the probability that the second one will be good is 79/99. Continue the reasoning and you get the answer:

$$\frac{80}{100}\times\frac{79}{99}\times\frac{78}{98}\times\frac{77}{97}\times\frac{76}{96}$$

(d) The total possible number of samples of 5 is 100C5. Out of these possible samples, many will be all non-defective. The possible number of all non-defective samples is 80C5. So the probability of an all-good sample is:

$$\frac{\begin{bmatrix}80\\5\end{bmatrix}}{\begin{bmatrix}100\\5\end{bmatrix}}=\frac{80!}{5!\ 75!}\div\frac{100!}{5!\ 95!}=\frac{80\times79\times78\times77\times76}{100\times99\times98\times97\times96}$$

F. SAMPLE SPACE

You need a clear head to perform probability calculations successfully. It helps to have some diagrammatic form of representation, and this is our concern in the final sections of this study unit. First, however, we must introduce more terminology. When we, say, toss a coin three times and note the outcome, we are performing a **statistical experiment**. If we make a list of all possible outcomes of our experiment, we call this a **sample space**. (This is a similar idea to a sampling frame, i.e. a list of a population, mentioned earlier in our discussion of practical sampling methods.) The sample space in the above coin-tossing experiment is:

H H H	H T T
T H H	T H T
H T H	T T H
H H T	T T T

where, for example, T H H means that on the first toss we obtained a tail, on the second a head, and on the third a head.

Consider another example. Suppose we have 5 people A, B, C, D, E and we wish to select for interview a random sample of 2, i.e. each of A, B, C, D, E must have the same chance of being chosen. What is the sample space, i.e. the list of all possible different samples? The sample space is:

A B	B C	C D	D E
A C	B D	C E	
A D	B E		
A E			

In this example the order of the sample, i.e. whether we choose A followed by B or B followed by A, is not relevant as we would still interview both A and B.

Having written down our sample space, we might be interested in a particular section of this sample space. For instance, in our first example we might want to see in how many of the outcomes we obtained only one head. We call this collection of outcomes an **event**, i.e. we are interested in the event: obtaining exactly one head. We often find it convenient to label an event by a capital letter such as A, B, etc., i.e. we could say event A is obtaining exactly one head. Looking back at the sample space, we see that there are three outcomes making up this event. If we have a fair coin, the probability of obtaining a head or a tail at any toss is ½ and all the outcomes in the sample space are equally likely. There are eight outcomes in all, so we can now deduce that:

Probability of obtaining exactly one head in 3 tosses of a fair coin

$$= \frac{\text{No. of outcomes in sample space with one head}}{\text{Total no. of outcomes in sample space}} = \frac{3}{8}$$

or, alternatively, we could write:

$$\text{p (Event A)} = \frac{\text{No. of outcomes in A}}{\text{Total no. of outcomes}} = \frac{3}{8}$$

Note:

p(Event A) is usually written as p(A), P(A), Pr(A) or Prob. (A).

Try writing out the details of the following example yourself before looking at the answer which follows.

Example

Experiment: rolling one fair die. Sample space is:

Event A	**p(A)**
1. Obtaining a score greater than 3.	
2. Obtaining an odd number.	
3. Obtaining both a score greater than 3 and an odd number.	

Answer to Example

Sample space is (1, 2, 3, 4, 5, 6).

Event A	p(A)
1	$\frac{3}{6}=\frac{1}{2}$
2	$\frac{3}{6}=\frac{1}{2}$
3	$\frac{1}{6}$

G. VENN DIAGRAMS

In a Venn diagram, the sample space, S, is represented by a rectangle, and events in the sample space are denoted by areas within the rectangle (Figure 11.1):

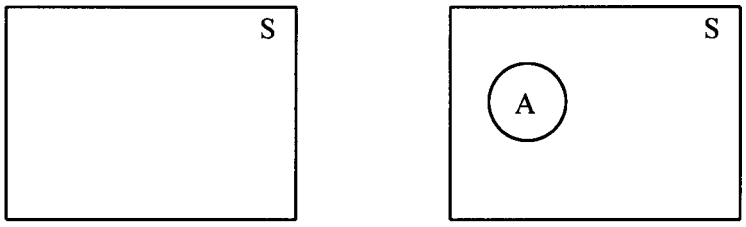


Figure 11.1

If all the outcomes listed in S are equally likely, then the probability of event A occurring is given by:

$$p(A) = \frac{\text{Number of outcomes in A}}{\text{Number of outcomes in S}}$$

$$= \frac{n(A)}{n(S)}$$ where n(A) is shorthand for the number of outcomes in event A.

General Addition Law of Probabilities

If we are interested in two events, A and B, then we have two areas representing A and B inside rectangle S and these **may** overlap (Figure 11.2):

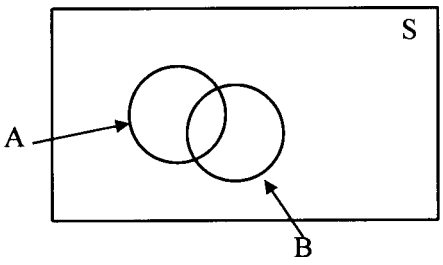


Figure 11.2

Consider the three tosses of a coin example which we introduced before.

- Let event A be "obtaining exactly one head",

  therefore, event A contains the outcomes (TTH, HTT, THT).

- Let event B be "obtaining a tail on the first toss",

  therefore, event B contains the outcomes (TTT, TTH, THT, THH).

In this case A and B overlap because the outcomes THT and TTH are common to both.

We call this overlap "A intersection B" denoted by $A \cap B$, and this is where **both A and B** occur. Thus to evaluate the probability of both A and B occurring together, we need:

$$p(A \text{ and } B) = \frac{\text{No. of outcomes in } A \cap B}{\text{No. of outcomes in S}}$$

$$= \frac{n(A \cap B)}{n(S)} = \frac{2}{8} \text{ for our example} = \frac{1}{4}$$

i.e. when we toss a coin three times, the probability of obtaining exactly one head and obtaining a tail on the first toss is ¼.

If we now look at the combined area covered by A and B, we see that within this region we have either event A occurring **or** event B occurring or both events occurring. We call this area "A union B" denoted by $A \cup B$. Thus to evaluate the probability of A or B or both occurring we need:

$$p(A \text{ or } B \text{ or both}) = \frac{\text{No. of outcomes in } A \cup B}{\text{No. of outcomes in S}}$$

$$= \frac{n(A \cup B)}{n(S)} = \frac{5}{8}$$

The 5 events that are in $A \cup B$ are (TTH, HTT, THT, TTT, THH). We have to be careful not to count THT and TTH twice. TTH and THT are the events that belong to $A \cap B$. We thus have the result which holds in general that:

$$A \cup B = A + B - A \cap B$$

$$\text{Thus, p(A or B or both)} = \frac{\text{No. of outcomes in (A+B} - \text{A} \cap \text{B)}}{\text{No. of outcomes in S}}$$

$$= \frac{\text{No. of outcomes in A + No. in B} - \text{No. in A} \cap \text{B}}{\text{No. in S}}$$

$$= \frac{n(A) + n(B) - n(A \cap B)}{n(S)}$$

$$= \frac{n(A)}{n(S)} + \frac{n(B)}{n(S)} - \frac{n(A \cap B)}{n(S)}, \text{ dividing each term in turn by n(S)}$$

$$= p(A) + p(B) - p(A \cap B) \text{ by definition of probability}$$

We thus have the general law of probabilities for **any** two events A and B:

$$p(A \text{ or } B \text{ or both}) = p(A) + p(B) - p(A \cap B)$$

i.e. $p(A \cup B) = p(A) + p(B) - p(A \cap B)$

Example 1

If one card is drawn from a pack of 52 playing cards, what is the probability (a) that it is either a spade or an ace; (b) that it is either a spade or the ace of diamonds?

(a) Let event B be "the card is a spade"

Let event A be "the card is an ace".

We require p(spade or ace [or both]) = p(A or B)

$$= p(A) + p(B) - p(A \cap B)$$

$$p(A) = \frac{\text{No.of aces}}{\text{No. in pack}} = \frac{4}{52}$$

$$p(B) = \frac{\text{No. of spades}}{\text{No. in pack}} = \frac{13}{52}$$

$$p(A \cap B) = \frac{\text{No. of aces of spades}}{\text{No. in pack}} = \frac{1}{52}$$

Therefore, p (spade or ace) $= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$

(b) Let event B be "the card is a spade".

Let event A be "the card is the ace of diamonds".

We require p (spade or ace of diamonds) = p(A or B)

$$= p(A) + p(B) - p(A \cap B)$$

$$p(A) = \frac{\text{No. of aces of diamonds}}{\text{No. in pack}} = \frac{1}{52}$$

$$p(B) = \frac{\text{No. of spades}}{\text{No. in pack}} = \frac{13}{52}$$

$$p(A \cap B) = \frac{\text{No. of spades which are also aces of diamonds}}{\text{No. in pack}} = 0$$

Therefore, p(spade or ace of diamonds) = $\frac{1}{52} + \frac{13}{52} = \frac{14}{52} = \frac{7}{26}$

Example 2

At a local shop 50% of customers buy unwrapped bread and 60% buy wrapped bread. What proportion of customers buy at least one kind of bread if 20% buy both wrapped and unwrapped bread?

- Let S represent all the customers.
- Let T represent those customers buying unwrapped bread.
- Let W represent those customers buying wrapped bread.

The Venn diagram is as shown in Figure 11.3:

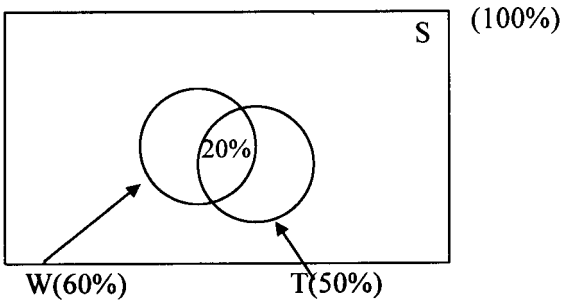


Figure 11.3

p(buy at least one kind of bread) = p(buy wrapped or unwrapped or both)

= p(T or W)

= p(T) + p(W) – p(T ∩ W)

= 0.5 + 0.6 – 0.2 = 0.9

i.e. nine-tenths of the customers buy at least one sort of bread.

The addition law can be extended **to cover more events**, e.g. for three events A, B and C:

p(at least one of A or B or C occurring) = p(A ∪ B ∪ C) and we find:

p(A ∪ B ∪ C) = p(A) + p(B) + p(C) – p(A ∩ B) – p(A ∩ C) – p (B ∩ C) + p(A ∩ B ∩ C)

Unless you like learning formulae, do not bother to remember this as you can always solve problems involving three events more easily using the Venn diagram.

Example 3

Three magazines A, B and C are published in Townsville. Of the adult population of Townsville 65% read A, 40% read B, 25% read C, 20% read both A and B, 20% read both A and C, 10% read both B and C and 5% read all three. What is the probability that an adult selected at random reads at least one of the magazines? Figure 11.4 shows this example.

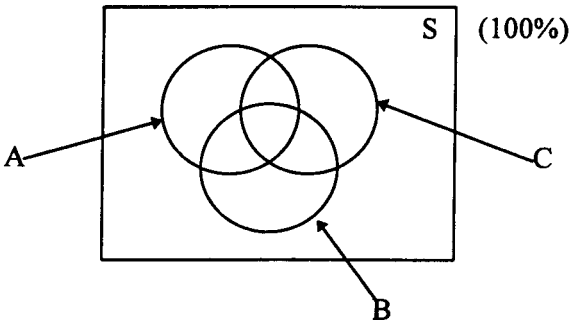


Figure 11.4

We want p(A or B or C), i.e. we require the proportion of S that is contained within the overlapping circles A, B and C. We take the information given in the question and insert the percentage lying within each section of the diagram. We start at the centre. We know 5% read all three magazines, so we insert 5% where all the three circles overlap. We are told that 20% read both A and B but 5% read C as well so that leaves 15% reading A and B but not C, so we insert 15% where just A and B overlap. Similarly for (A and C) and (B and C). Our diagram is now as shown in Figure 11.5.

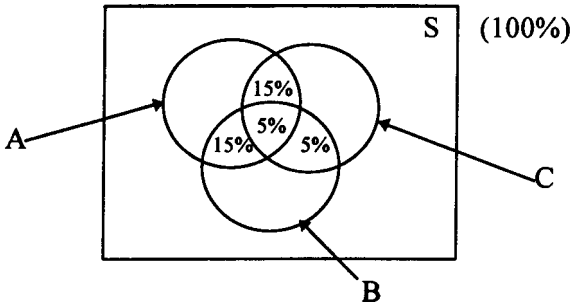


Figure 11.5

We are told 65% read A, but looking at Figure 11.5 you will see that 5% + 15% + 15%, i.e. 35% read A together with at least one other magazine, leaving 30% who read A only. Similarly, 15% read B only and no one reads C only. The completed diagram is shown in Figure 11.6.

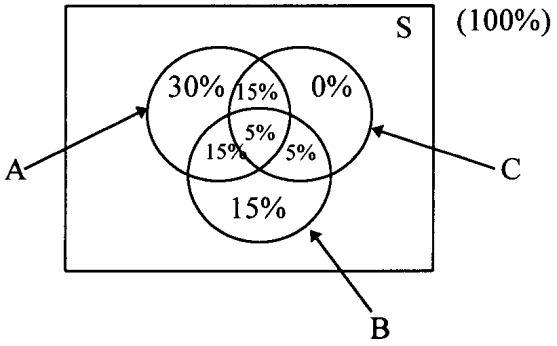


Figure 11.6

The total percentage within A, B and C is 30 + 15 + 15 + 5 + 5 + 15, i.e. 85%. Thus the probability that an adult selected at random reads at least one of the magazines is 0.85.

Now try the following further examples yourself.

Further Examples

(a) In a group of 25 trainees, 16 are males, 12 are university graduates and 10 are male graduates. What is the number of female non-graduates?

(b) The probability that a manager reads the Daily Telegraph is 0.7. The probability that he reads the Daily Telegraph but not the Financial Times is 0.6. The probability that he reads neither is 0.2. Find the probability that he reads the Financial Times only.

(c) Employees have the choice of one of three schemes, A, B or C. They must vote for one but, if they have no preference, can vote for all three or, if against one scheme, they can vote for the two they prefer.

A sample poll of 200 voters revealed the following information:

15 would vote for A and C but not B

65 would vote for B only

51 would vote for C only

15 would vote for both A and B

117 would vote for either A or B, or both A and B, but not C

128 would vote for either B or C, or both B and C, but not A

How many would vote for:

(i) All three schemes?

(ii) Only one scheme?

(iii) A irrespective of B or C?

(iv) A only?

(v) A and B but not C?

Answers to Further Examples

(a)

Male G 10	Male Non-G 6
Female G 2	Female Non-G

S(25) Fill in the number of male graduates (10) first.

Then fill in the number of male non-graduates (16 – 10)

Next fill in the number of female graduates (12 – 10)

There are 25 trainees altogether, so the number of female non-graduates is
25 – (10 + 6 + 2) = 7.

(b) First mark in 0.6 for those reading the Telegraph but not the Financial Times. Then fill in the overlap for those who read both (0.7 – 0.6). 0.2 of the managers are outside both circles. All

the probabilities must add up to 1, so the probability that he reads only the Financial Times is $1 – (0.1 + 0.6 + 0.2) = 0.1$ (see Figure 11.7).

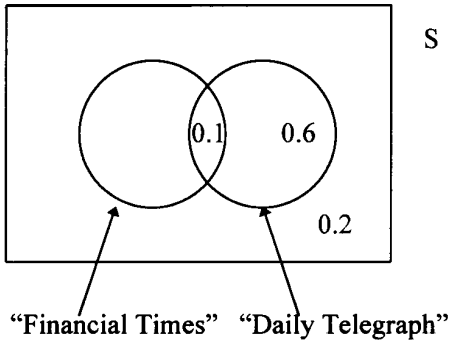


Figure 11.7

(c) Let x be the number voting for all three schemes. We then fill in the numbers in each section in terms of x and use the fact that the sum of the numbers in all the sections must be 200. Figure 11.8 contains these values.

Note: We are told that 15 would vote for both A and B and we must assume that this 15 includes those who might vote for C as well, so there are only 15 – x who vote for A and B but not C.

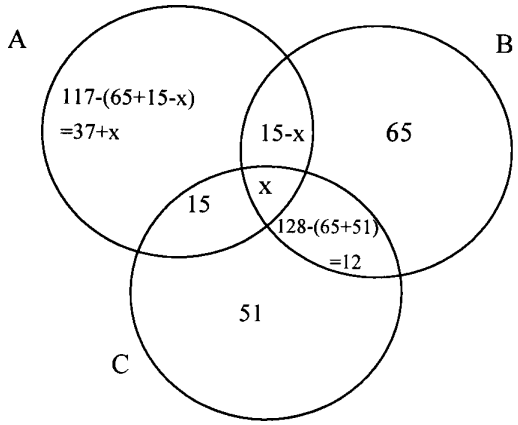


Figure 11.8

Therefore, $37 + x + 15 – x + 15 + x + 65 + 12 + 51 = 200$

i.e. $195 + x = 200$

$x = 5$

(i) 5 would vote for all three schemes.

(ii) $42 + 65 + 51 = 158$ would vote for only one scheme.

(iii) $42 + 15 – x + x + 15 = 72$ would vote for A.

(iv) 42 would vote for A only.

(v) 10 would vote for A and B but not C.

Mutually Exclusive Events

In an earlier section, we defined the term "mutually exclusive events" and said that if two events were mutually exclusive then if one occurred the other could **not**. In a Venn diagram, if two events A and B are mutually exclusive then the areas corresponding to A and B will not overlap. If we return again to the example where we tossed a coin three times, let us consider:

- Event A = (obtaining exactly one head)
- Event B = (obtaining three heads)

These events are mutually exclusive so the Venn diagram is as shown in Figure 11.9:

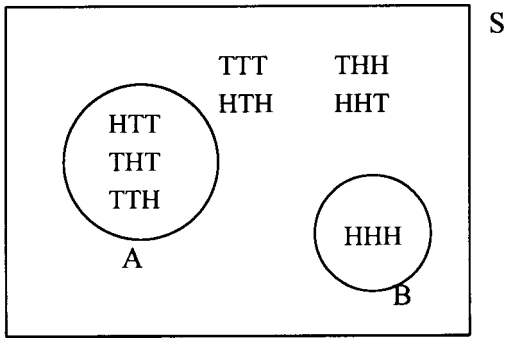


Figure 11.9

There is no overlap between Events A and B so $p(A \cap B) = 0$ as we cannot obtain exactly one head and three heads at the same time.

Thus the general addition law simplifies to become:

$$p(A \text{ or } B \text{ or both}) = p(A) + p(B)$$

This is the simple addition law which we stated previously for mutually exclusive events and it can be generalised for any number of mutually exclusive events (Figure 11.10):

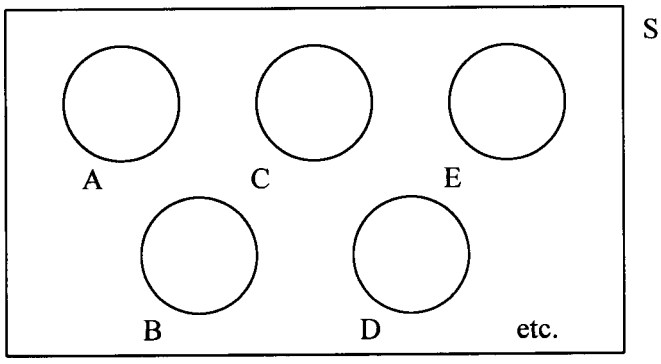


Figure 11.10

$$p(A \text{ or } B \text{ or } C \text{ or } D \text{ or } E \text{ or } ...) = p(A) + p(B) + p(C) + p(D) + p(E) + ...$$

General Multiplication Law of Probability

We have seen before that sometimes we need to work out the probability of an event A occurring, given that event B has already occurred. We called this the **conditional probability** of A given B has occurred. Let us now consider how we can work out such probabilities from the Venn diagram. Considering once more the example when we toss a coin three times, we might want to know the probability of obtaining 3 heads given that the first toss is known to be a head.

Let A = (obtaining 3 heads) = (HHH)

$$p(A) = \frac{1}{8}$$

B = (first toss a head) = (HTT, HTH, HHT, HHH)

$$p(B) = \frac{4}{8}$$

A and B can be represented on the Venn diagram, Figure 11.11:

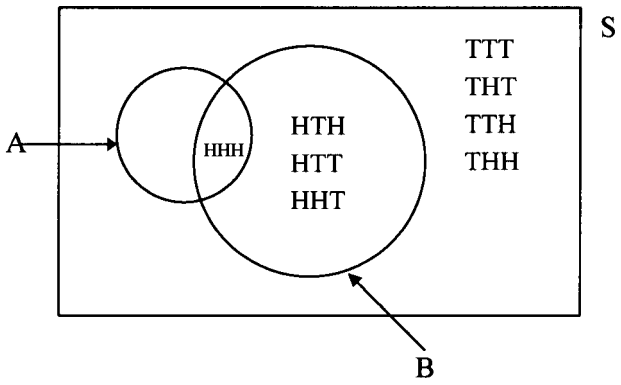


Figure 11.11

The conditional probability p(obtaining 3 heads given first toss a head) is written as $p(A|B)$ and read as probability of A given B. Thus required probability

$$= \frac{\text{p(obtaining 3 heads } \textbf{and} \text{ first toss head)}}{\text{p(first toss is a head)}} = \frac{\left(\frac{1}{8}\right)}{\left(\frac{4}{8}\right)} = \frac{1}{4}$$

Thus what we are working out is the number of outcomes in the overlap as a fraction of the number of outcomes in B, as we know that B has to have occurred already

$$\text{i.e. } p(A|B) = \frac{n(A \cap B)}{n(B)} = \frac{1}{4}$$

$$\text{Thus } p(A|B) = \frac{p(A \cap B)}{p(B)} \quad \text{.....} \qquad \textit{(b)}$$

Example 1

If two coins are tossed, what is the probability that both are heads, given that at least one is a head?

Let A = (both are heads) = (HH)

and B = (at least one is a head) = (HH, TH, HT)

S = (HH, TH, HT, TT)

The Venn diagram is shown in Figure 11.12:

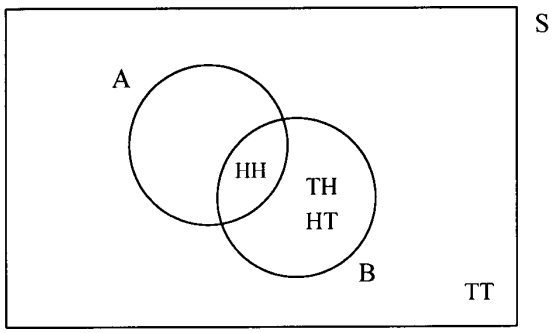


Figure 11.12

$$p(A) = \frac{1}{4} \qquad p(B) = \frac{3}{4} \qquad p(A \cap B) = \frac{1}{4}$$

$$\text{Required probability} = p(A|B) = \frac{p(A \cap B)}{p(B)} = \frac{\left(\frac{1}{4}\right)}{\left(\frac{3}{4}\right)} = \frac{1}{3}$$

If we rewrite equation (b) we obtain:

$$p(A \cap B) = p(B)p(A|B)$$

i.e. $p(A \textbf{ and } B) = p(B)p(A|B)$

Thus the probability of both events A and B occurring is the probability of B occurring **times** the probability of A occurring, given that B has already occurred. The above law is the general multiplication law for conditional probabilities stated in Section E.

Example 2

A bag contains 5 red and 3 white billiard balls. If two are selected at random without replacement, what is the probability that one of each colour is drawn?

- Let R = (drawing a red ball)
- Let W = (drawing a white ball)

As the sampling is without replacement, after selecting one ball we do not return it to the bag before choosing the second ball.

$$p(R \text{ and } W) = p(R \text{ then } W) \text{ or } (W \text{ then } R)$$

$$= p(R \text{ then } W) + p(W \text{ then } R) \text{ using addition law for mutually exclusive events}$$

$$= p(R)p(W|R) + p(W)p(R|W)$$

$$= \left[\frac{5}{8} \times \frac{3}{7}\right] + \left[\frac{3}{8} \times \frac{5}{7}\right] = \frac{15}{56} + \frac{15}{56} = \frac{15}{28}$$

Therefore, probability that one of each colour is drawn is $\frac{15}{28}$.

Independent and Dependent Events

For independent events the probability of A occurring does not depend on whether B has already occurred. Thus

$$p(A|B) = p(A)$$

and $p(A \cap B) = p(A)p(B)$

which is the multiplication law for probabilities of independent events which we used earlier in the Study Unit.

Example

Two cards are drawn at random from a pack of 52 cards. Find the probability of drawing two aces

(a) if the sampling is with replacement; and

(b) if the sampling is without replacement.

(a) If the sampling is with replacement, this implies that after we have drawn the first card and looked at it, we put it back in the pack before drawing the second card. Thus the probability of an ace at either draw is

$$\frac{4}{52} = \frac{1}{13}.$$

By the multiplication law for independent events, the probability of drawing two aces is

$$\frac{1}{13} \times \frac{1}{13} = \frac{1}{169} = 0.0059.$$

(b) If the sampling is without replacement, after we have drawn the first card we do not put it back in the pack before taking the second card.

Thus the probability of an ace at the first draw is $\frac{4}{52}$

However, the probability of an ace at the second draw depends on whether or not we had an ace first time. We thus need to use the multiplication law for conditional probabilities:

Probability of drawing two aces = p (drawing an ace first time) × p (drawing a second ace, given that the first was an ace)

$$= \frac{4}{52} \times \frac{3}{51} = 0.0045$$

i.e. the probability this time is smaller, as was to be expected.

H. SUMMARY

You must learn the definitions and notation in this study unit so that you are quite sure of the meaning of any question on probability. In particular, you should learn the following formulae:

(1) Relative frequency $= \dfrac{\text{Frequency in a particular class}}{\text{Total frequency}}$

$$= \frac{n(A)}{n(S)} = P(A)$$

(2) $P(A) + P(\overline{A}) = 1 \Rightarrow P(\overline{A}) = 1 - P(A)$

(3) The multiplication law:

(i) $P(A \cap B) = P(A)P(B/A)$

(ii) $P(A \cap B) = P(A)P(B)$ if A and B are independent.

(4) The addition law

(i) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(ii) $P(A \cup B) = P(A) + P(B)$ if A and B are mutually exclusive.

Practice in drawing Venn diagrams will help in solving problems. If you do not see at once how to solve a problem, tackle it in stages:

(a) Define the sample space.

(b) Define the events of interest and make sure that they belong to the sample space.

(c) Decide what type of events you are dealing with.

(d) Try to draw a Venn diagram to illustrate the problem. (Remember this may not always be possible.)

(e) Decide which probability law is required to solve the problem.

Make sure that you explain the method you use.

Study Unit 12

Frequency Distributions

Contents ***Page***

A. INTRODUCTION

Earlier in the course we discussed the measures (called statistics) used to give a numerical description of the location and variability or dispersion of a set of data. Unless these sets of data are very small, these statistics are calculated from frequency or grouped frequency distributions which are illustrated graphically by histograms or frequency polygons. If the number of classes in a frequency distribution is very large, these polygons can be drawn as smooth frequency curves.

In this study unit we will use these curves, and also some of the concepts of probability from the previous two study units, to introduce the idea of theoretical frequency distributions. A theoretical frequency distribution is constructed from a trial and its sampling space; its shape can be deduced from its measures of location and dispersion.

B. THEORETICAL FREQUENCY CURVES

Discrete Variables

Earlier we found the probability of occurrence of the values of the discrete variable (result of trial) obtained when a coin is tossed, a die is thrown or a card is picked. From the sample space we can work out these probabilities and, since they are relative frequencies, we can construct a theoretical relative frequency distribution, e.g. if a true die is thrown once, the theoretical relative frequency distribution is:

Score	1	2	3	4	5	6
Relative Frequency	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Now suppose you throw a die a hundred times and obtain the observed frequencies shown in Table 12.1. Since all the throws are independent, this is a random sample of scores taken from all the possible scores.

Table 12.1: Observed Frequencies of a Die

Score	Observed Frequency	Theoretical Frequency
1	18	16.6....
2	14	16.6....
3	22	16.6....
4	14	16.6....
5	17	16.6....
6	15	16.6.....
Total	100	100.0

The third column in this table is calculated from the relative frequency table, as in theory the actual frequency is equal to the relative frequency multiplied by the total frequency.

There are several important features that you should notice about this simple frequency distribution:

(a) We have had to make an **assumption** in order to calculate the theoretical frequency. In this case the assumption was that we were throwing a true or unbiased die.

(b) As the observed frequencies are obtained from a small sample, they are **not** exactly the same for each score. If the assumption we have made is correct, they will approach equality as the sample size increases.

(c) The theoretical frequencies are not whole numbers and so, in practice, they can never occur. In spite of this they are **useful** in statistical inference as they help us to make decisions about the distribution.

Suppose you throw another die 100 times and obtain the frequencies shown in Table 12.2:

Table 12.2: Further Frequencies

Score	Observed Frequency
1	25
2	10
3	15
4	5
5	10
6	35
Total	100

Compare these frequencies with the theoretical frequencies and observed frequencies in Table 12.1. The frequencies for 1 and 6 suggest that the die is loaded towards these scores, i.e. that the assumption made is wrong and you have a biased die.

This use of theoretical frequency distributions can be applied to more practical situations.

Example

Suppose you were in charge of a premium bond type scheme in which there were 1,000 numbered bond-holders and 100 £1 prizes, so that everybody had a theoretical chance of 1/10 (100 ÷ 1,000) of winning a prize. You would suspect that there was something wrong with the selection system if one bond-holder won £25 or ¼ of the prizes. You might even be doubtful of the system if one bond-holder won as much as £5.

The decision as to when the difference between observed and theoretical values is large enough for you to initiate some checking action is made by performing **hypothesis tests**. We will cover these in a later study unit.

Continuous Variables

The calculation of the probability of occurrence of values of continuous variables must be tackled in a slightly different way from that used for discrete variables. Since a continuous variable can take

any value on a continuous scale, it never takes an exact value, so we find the probability with which it lies in a given interval.

To find out how to construct a theoretical frequency distribution for a continuous variable, suppose we look at the probability of occurrence of a road accident during the course of a day. Accidents will occur throughout the day with generally a peak during rush hours. It would be impossible to group accidents by each second of the day, so usually they are grouped into hours.

As the intervals chosen are quite small, we have an observed frequency distribution with large frequencies in some intervals and small frequencies in others. From such a frequency distribution, a frequency curve may be plotted.

From the shape of this curve and the value of the statistics calculated from the sample observations taken, we look for a curve the shape and parameters of which appear to fit the population from which the sample has come. This curve is the theoretical frequency distribution of the continuous variable we have observed, i.e. the time at which accidents are likely to occur with a particular probability.

This method for finding a theoretical frequency distribution is also used for discrete variables when the number of values the variable can take is so large that it is best to group them together in some form.

C. SHAPES OF DIFFERENT DISTRIBUTIONS

Rectangular Distribution

Figure 12.1 is a typical example of a rectangular distribution. This is a distribution where each value of the variable, continuous or discrete, occurs with the same frequency, i.e. it is equally likely or occurs with the same probability.

It is not hard to see why it is called a rectangular distribution because the **frequency polygon** is formed by a straight line parallel to the variable axis, with the sides of the rectangle being formed by the lowest and highest class boundaries. Note that the left-hand side is the frequency axis only if the lowest class boundary is zero.

Examples of this distribution are many; throwing a die, picking a card or drawing a lottery ticket are a few of them.

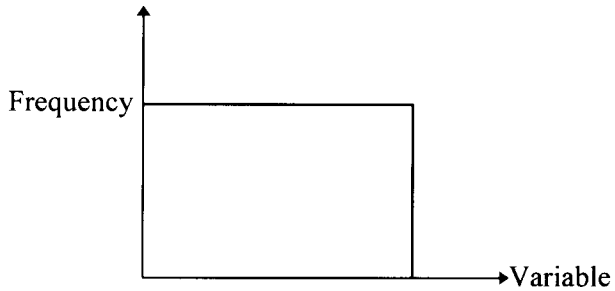


Figure 12.1: Rectangular Distribution

J-Shaped Distribution

This has a frequency curve that has no mode and is extremely skewed either positively or negatively, as shown in Figure 12.2.

If the shape of the curve on one side of the mean is exactly the same as its shape on the other side, then the distribution is called **symmetrical**. Remember that if the shape of the curve is different on either side of the mean then the distribution is called **skew** or **asymmetrical**. A distribution is **positively skewed** if the mode is to the left with a long tail to the right; it is **negatively skewed** if the mode is to the right and has a long tail to the left.

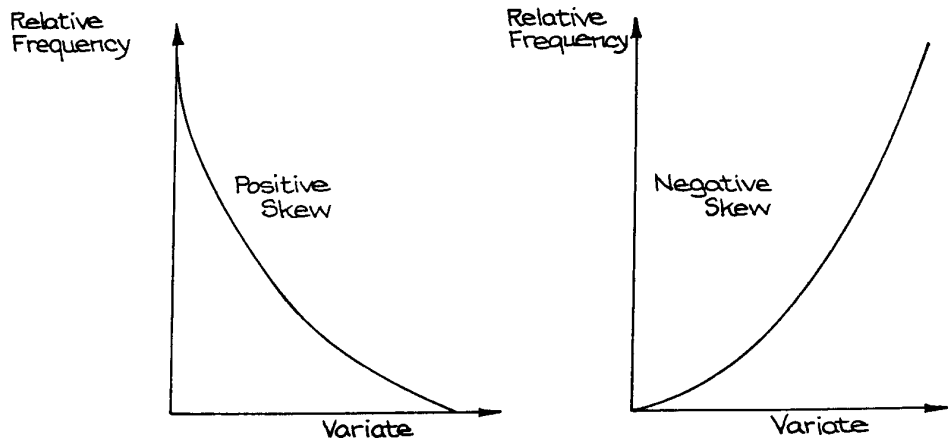


Figure 12.2: J Distribution

The amount of money in building society members' accounts or the length of time of telephone calls are examples of positively skewed J-distributions. Negatively skewed J-distributions are rare; one is the number of adult deaths from a specified disease.

U-Shaped Distribution

A distribution of this type is shown in Figure 12.3. The smallest frequency is in the middle range of values of the variable, with the largest frequencies at the end.

Sometimes this distribution can be formed by combining a positive and a negative J-shaped distribution. Amounts of saving at different ages sometimes exhibit this curve, because when you are young you often save considerably more than at middle age, and then your savings often creep up again in old age.

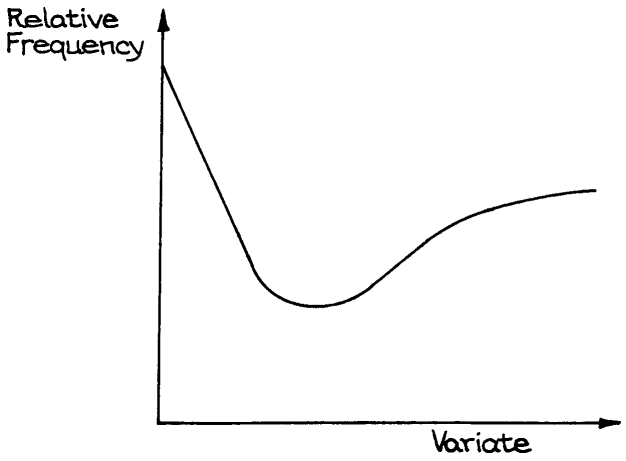


Figure 12.3: U-Shaped Distribution

D. THE NORMAL DISTRIBUTION

Definition

The most **important** frequency distribution is one which is unimodal, with more frequencies near the centre and fewer frequencies in the tails. This type occurs very often. Consider a coin-tossing trial, but instead of constructing the rectangular frequency distribution from the number of named outcomes as earlier, take the discrete variable, x, as the number of heads occurring at each toss. If the coin is tossed once, x takes the values 0 or 1; if it is tossed twice, x can be 0, 1 or 2; if it is tossed three times, x can be 0, 1, 2 or 3, and so on.

Figure 12.4 shows the relative frequency histogram and polygon for ten tosses of an unbiased coin. Both the histogram and the polygon are symmetrical about x = 5, which is thus the value of the mode, median and mean; the relative frequencies decrease symmetrically on each side of x = 5 and the value for x = 0 and x = 10 is small.

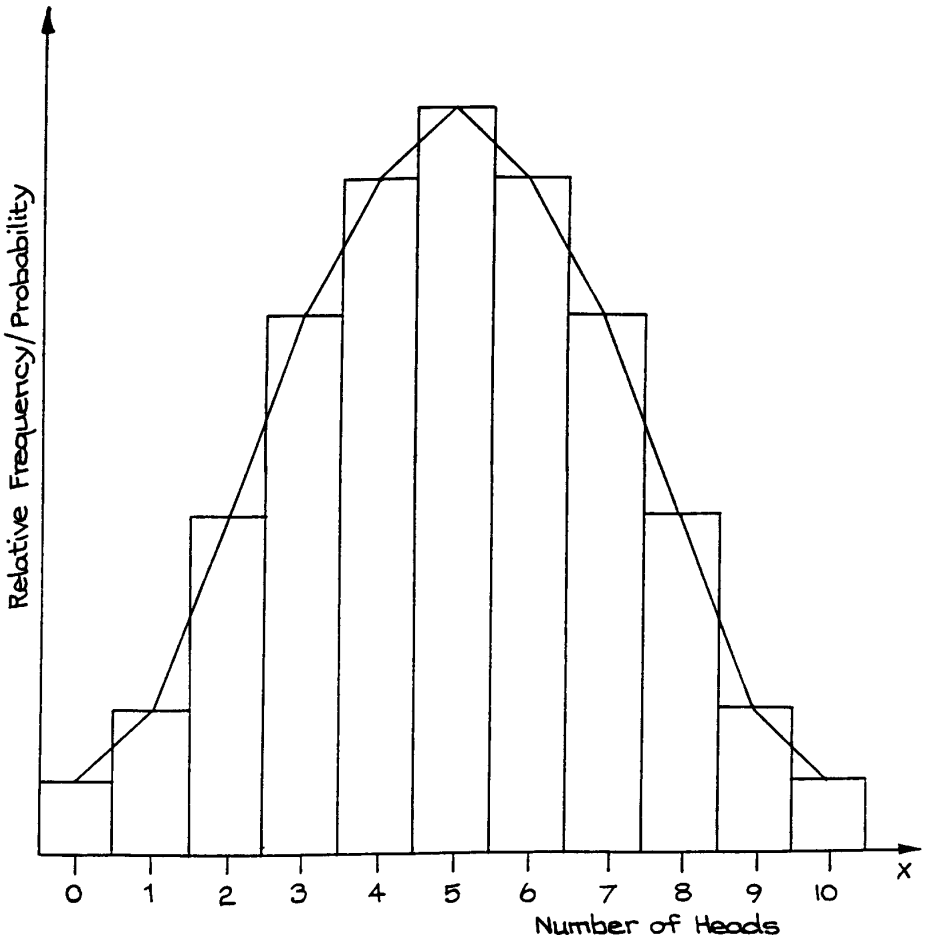


Figure 12.4: Histogram and Polygon of a Symmetrical Distribution

For your own satisfaction you should construct the relative frequency histograms and polygons for other numbers of tosses to see that the same shape relative frequency polygon is obtained each time. The same shape is obtained if, for example, two dice are thrown and the variable is the total score showing each time.

Examine Figure 12.4 carefully and notice that, if we take the width of each bar of the histogram as one unit, the area of the bar for each value is the relative frequency of that value, and the sum of all the relative frequencies is one unit.

This relative frequency distribution is a probability distribution and the probability of occurrence of any value of x is the area of the bar above the value. If you compare the areas of the histogram bars with the area between the boundaries of the bars, the x-axis and the relative frequency polygon, you can see that they are approximately the same size. So you can find the probability of occurrence of any value or values of x either from the histogram or from the polygon. As the number of possible values of x increases, the relative frequency polygon becomes a smooth relative frequency curve as shown in Figure 12.5.

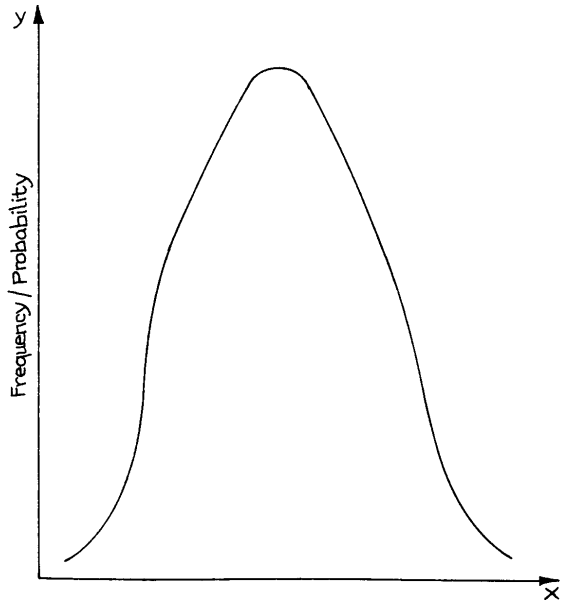


Figure 12.5: Relative Frequency Curve of a Symmetrical Distribution

This relative frequency curve includes the relative frequency of all the values of the variable, x, $-\infty$ to $+\infty$, and the area between the curve and the x-axis is one unit so it is a theoretical relative frequency distribution or a probability distribution.

This unimodal symmetrical bell-shaped curve is of great theoretical and practical importance in statistical work. It is called the **normal distribution** or sometimes the **Gaussian distribution** after the scientist who developed its use for examining random errors of observation in experimental work.

When we consider the relative frequency curves of continuous variables, we discover a similar pattern in the measurements of a great many natural phenomena. For example, the frequency curve obtained from the set of heights of 80 employees, used earlier in the course, is unimodal with small frequencies in the tails. Later, we calculated the mean, standard deviation and median of this set of data. Since we were dealing with a comparatively small sample, the values of these measures were empirical, and it is reasonable to assume that the theoretical relative frequency distribution (the probability distribution) deduced from the data would be symmetrical and normal.

The same type of frequency distribution is found in the populations of dimensions of items from a factory production line and errors of observation in scientific experiments. This wide range of application of the normal distribution accounts for its importance in statistical inference.

Properties

If y is the ordinate of the probability curve, then the normal probability curve can be plotted from the equation $y = \phi(x)$ where x is a continuous variable which can take all values from $-\infty$ to $+\infty$ and $\phi(x)$ is a function of x which is completely defined by two parameters. (For this course you do not need to know the actual equation, which is rather complicated.) We shall denote these two parameters by the Greek letters μ and σ, in other words the mean and standard deviation of the distribution.

By giving these parameters different values, we can obtain a set of curves, each of which has the following properties:

(a) The curve is symmetrical, unimodal and bell-shaped.

(b) All the values of y are greater than zero and approach zero as x approaches $\pm\infty$.

(c) It can be proved that the area between the curve and the x-axis is one unit.

(d) It can be proved that:

  (i) The mean, mode and median are all equal to the parameter μ.

  (ii) The standard deviation is equal to the parameter σ.

(e) $P(x_1 < x < x_2)$ = Area under the curve between the ordinates $\phi(x_1)$ and $\phi(x_2)$.

(f) If $z = \dfrac{x-\mu}{\sigma}$, it can be proved that z has the same normal distribution for **every** pair of values of the parameters μ and σ.

This distribution is called the **standard normal distribution**, and z is called the **z-score** or the **standardised value** of the variable.

Standard Normal Distribution

Since z is normally distributed, the range of z is $-\infty$ to $+\infty$ and the distribution has the properties (a), (b), (c), (d) and (e) listed above. In addition, it can be proved that $\mu = 0$ and $\sigma = 1$.

Figure 12.6 shows the shape and location of the standard normal distribution.

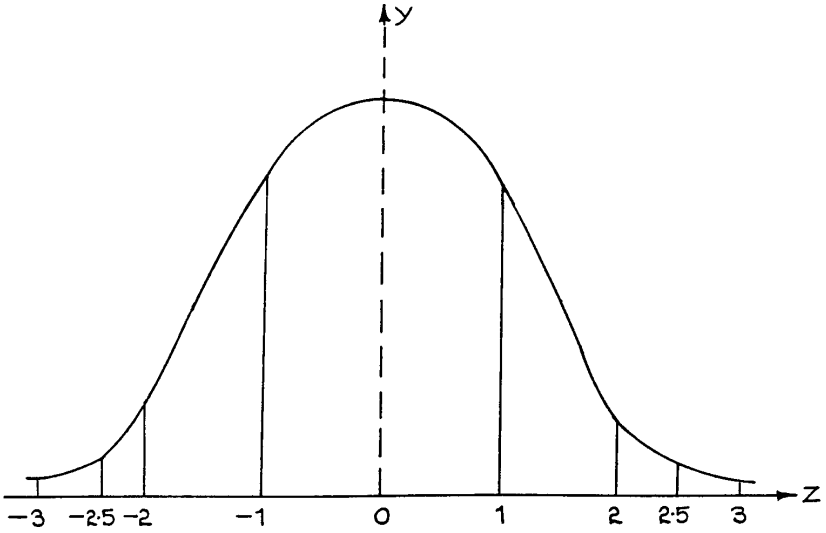


Figure 12.6: Standard Normal Distribution

Since $\sigma = 1$, the ordinates drawn in Figure 12.6 are 1, 2 and 2.5 standard deviations on each side of the mean. You can see that values of z more than 3 standard deviations from the mean are very unlikely to occur and that 50% of the values of z lie below zero (the mean), and 50% above zero.

It can be calculated that:

- About 68% of the distribution lies within 1 standard deviation of the mean.
- About 95% of the distribution lies within 2 standard deviations of the mean.
- About 99% of the distribution lies within 2.5 standard deviations of the mean.

Since z is a continuous variable we cannot find the probability that z takes any exact value but only the probability that it lies in a given range of values, i.e. the probability that the value of z lies between any two given values is equal to the area under the curve between the ordinates at these two values.

Figure 12.7 shows the standard normal distribution divided into several areas by the ordinates at z_1, z_2 and z_3.

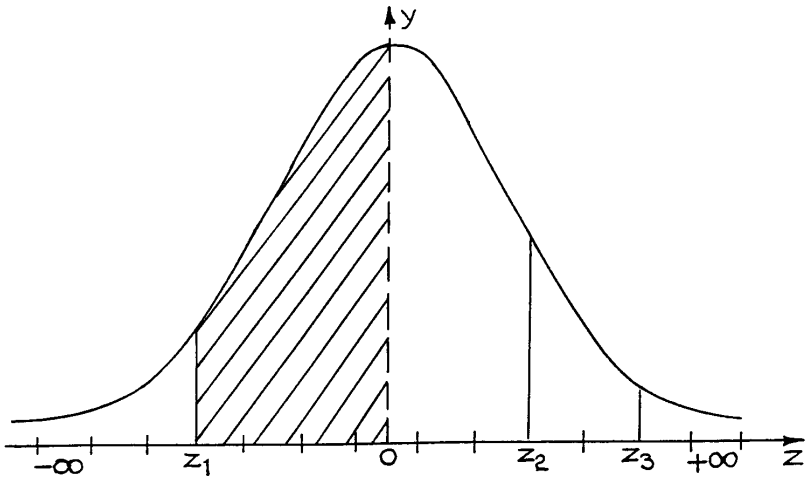


Figure 12.7: Different Areas of the Standard Normal Distribution

Suppose we let $\Phi(z_1)$ = Area under curve between the ordinate at z_1 and mean, in this case 0 (i.e. the shaded area).

Then: $P(z < z_1) = 0.5 - \Phi(z_1)$

Similarly: $P(z < z_2) = 0.5 + \Phi(z_2)$

$P(z < z_3) = 0.5 + \Phi(z_3)$

$P(z_1 < z < z_2)$ = Area under curve between the ordinates at z_1 and z_2

$= \Phi(z_2) + \Phi(z_1)$

Similarly, $P(z_1 < z < z_3) = \Phi(z_3) + \Phi(z_1)$, but, $P(z_2 < z < z_3) = \Phi(z_3) - \Phi(z_2)$

$P(z > z_3)$ = Area under curve above the ordinate at z_3

= Area under whole curve from mean to $+\infty$ – area under curve below the ordinate at z_3

$= 0.5 - \Phi(z_3)$ (since area under whole curve from mean to $\infty = 0.5$ of a unit)

Similarly, $P(z > z_2) = 0.5 - \Phi(z_2)$, but $P(z > z_1) = 0.5 + \Phi(z_1)$

You will see that in each case we have expressed the probability we require in terms of the areas between values of z.

The reason for doing this is that we do not calculate each individual standard normal probability; we use a table of standard normal probabilities and this table is compiled in the form of areas between mean and values of z. In the next section we will examine this table and work at a number of examples with its help.

E. USE OF THE STANDARD NORMAL TABLE

The Appendix to this unit sets out a Table showing the standard normal probabilities.

This gives the area under the standard normal curve between the mean, μ, in this case 0, and a point x standard deviations above the mean. These areas we shall refer to using the previous notation of Φ(z). The values of z range from 0 to 3.5. You will notice that as the value of z increases above 3.0, the number of decimal places increases from 4 to 5, because the curve is approaching the axis and flattening out, so the probabilities are changing slowly and for values above 3.5 the probability is effectively 0.5. Table 12.3 shows a selection of z values taken from the Table in the Appendix.

Table 12.3

z	Φ(z)	z	Φ(z)	z	Φ(z)
0.0	0.0	0.5	0.1915	1.0	0.3413
0.01	0.0040	0.51	0.1950	1.01	0.3438
0.02	0.0080	0.52	0.1985	1.02	0.3461
0.25	0.0987	0.75	0.2734	2.0	0.4772

Remember that these values are the probabilities from μ to z. In order to calculate the values of probabilities from –∞ to z, 0.5 must be added to the values in the table.

Some tables are in fact drawn up with this value of 0.5 added in already, and therefore show Φ(z) values from –∞ to z. Table 12.4 shows several examples:

Table 12.4: Values of Φ(z) from –∞ to z

z	Φ(z)	z	Φ(z)	z	Φ(z)
0.0	0.5	0.5	0.6915	1.0	0.8413
0.02	0.5040	0.75	0.7734	2.0	0.9772

In all our examples we will use the method shown in Table 12.3.

To use the table of standard normal probabilities, as well as expressing the probability we require in the form shown at the end of the previous section, we have to see how to use the symmetry of the standard normal curve for values of z between –∞ and 0 and also for values of z with 3 decimal places. But first let us practise using the table for positive values of z with 2 decimal places.

Example 1

From Table 12.3 and with the help of Figure 12.7 find:

(a) $P(z < 0.01)$ (b) $P(z < 0.52)$ (c) $P(z < 1.00)$

(d) $P(0.02 < z < 1.02)$ (e) $P(z > 0.51)$

Answers

(a) $P(z < 0.01) = 0.5 + \Phi(0.01) = 0.5040$

(b) $P(z < 0.52) = 0.5 + \Phi(0.52) = 0.6985$

(c) $P(z < 1.00) = 0.5 + \Phi(1.00) = 0.8413$

(d) $P(0.02 < z < 1.02) = \Phi(1.02) - \Phi(0.02) = 0.3461 - 0.0080 = 0.3381$

(e) $P(z > 0.51) = 0.5 - \Phi(0.51) = 0.5 - 0.1950 = 0.3050$

Example 2

Find:

(a) $P(z < 0.512)$ (b) $P(z < 0.006)$ (c) $P(z < 1.017)$

Answers

(a) Since 0.512 lies between 0.51 and 0.52, $\Phi(0.512)$ will lie between $\Phi(0.51)$ and $\Phi(0.52)$ and we have to interpolate between these two values. The two values of z are close together so we can get a sufficiently accurate value for $\Phi(0.512)$ by assuming that the part of the distribution curve between $\Phi(0.51)$ and $\Phi(0.52)$ is a straight line. Thus, by simple proportion:

$$\Phi(0.512) = 0.5 + \Phi(0.51) + \frac{2}{10} \text{ of } [\Phi(0.52) - \Phi(0.51)]$$

$$= 0.6950 + \frac{2}{10} \times 0.0035$$

$$= 0.6950 + 0.0007 = 0.6957$$

Therefore, $P(z < 0.512) = 0.6957$

(b) To find this probability, use the first and second rows of the first column headed $\Phi(z)$. The difference we require is 0.0040, and 0.006 is 0.6 of the distance between 0.00 and 0.01, giving

$$P(z < 0.006) = 0.5000 + \frac{6}{10} \times 0.0040 = 0.5024$$

(c) Using the second and third rows of the third column headed $\Phi(z)$:

$$P(z < 1.017) = 0.5 + 0.3438 + \frac{7}{10} \times 0.0023$$

$$= 0.8438 + 0.00161$$

$$= 0.8454 \text{ to 4 dp}$$

Example 3

Find:

(a) $P(z < -1.02)$ (b) $P(z > -0.50)$ (c) $P(-1.01 < z < -0.52)$

(d) $P(-0.51 < z < 1.00)$

Answers

(a) Look at Figure 12.7 and let $z_1 = -1.02$. Then by symmetry the area between the ordinate at –1.02 and μ is equal to the area between the ordinate at 1.02 and μ.

This implies that $P(z < -1.02) = 0.5 - \Phi(1.02)$

$= 0.5 - 0.3461$

$= 0.1539$

(b) This time let z_1 be –0.50, then by symmetry the area between the ordinate at –0.50 and μ is equal to the area between the ordinate at 0.50 and μ. This implies that:

$$P(z > -0.50) = P(z < 0.50) + 0.5 = \Phi(0.50) + 0.5 = 0.6915$$

(c) $P(-1.01 < z < -0.52) = \Phi(1.01) - \Phi(0.52)$

$= 0.3438 - 0.1985$

$= 0.1453$

(d) $P(-0.51 < z < 1.00) = \Phi(0.51) + \Phi(1.0)$

$= 0.1950 + 0.3413$

$= 0.5363$

Example 4

Now use the table in the Appendix to find:

(a) $P(z < 2.86)$ (b) $P(z > 1.58)$ (c) $P(z < 2.224)$

(d) $P(z < -1.83)$ (e) $P(z > -2.49)$ (f) $P(-1.6 < z < 1.34)$

Answers

(a) $P(z < 2.86) = \Phi(2.86) + 0.5 = 0.9979$

(b) $P(z > 1.58) = 0.5 - \Phi(1.58) = 0.5 - 0.4430 = 0.0570$

(c) $P(z < 2.224) = 0.5 + \Phi(2.22) + \frac{4}{10}[\Phi(2.23) - \Phi(2.22)]$

$= 0.9868 + \frac{4}{10} \times 0.0003$

$= 0.9868 + 0.00012 = 0.98692$ to 5 dec. pl.

(d) $P(z < -1.83) = 0.5 - \Phi(1.83) = 0.5 - 0.4664 = 0.0336$

(e) $P(z > -2.49) = 0.5 + \Phi(2.49) = 0.9936$

(f) $P(-1.6 < z < 1.34) = \Phi(1.6) + \Phi(1.34) = 0.8551$

When you are confident in the use of this table there is no need to write in the lines of working which give the probabilities in terms of Φ(z).

Now that we know how to use the standard normal tables to find the probability of occurrence of any required range of values of z, we must look at the application of this theory to practical problems where we are dealing with variables which have normal distributions with means not equal to zero and standard deviations not equal to one. In the next section we will establish the connection between standard normal probabilities and other practical probabilities.

F. GENERAL NORMAL PROBABILITIES

Let the continuous variable x have a normal distribution with mean μ and standard deviation σ. The probability curve for this distribution will have all the properties listed earlier; in particular, it will be symmetrical about the ordinate at μ.

Then $P(x_1 < x < x_2)$ = Area under the normal curve between the ordinates at x_1 and x_2.

The standardised values of x_1 and x_2 are:

$$z_1 = \frac{x_1 - \mu}{\sigma} \text{ and } z_2 = \frac{x_2 - \mu}{\sigma}$$

So, if $x_1 < x < x_2$, $z_1 < z < z_2$ where $z = \frac{x - \mu}{\sigma}$,:

- z will be a standard normal variable, and
- $P(z_1 < z < z_2)$ = Area under the standard normal curve between the ordinates at z_1 and z_2

Since z_1 and z_2 are linear functions of x_1 and x_2, the ratios of these two areas to the total areas under the two curves are equal, i.e.

$$\frac{\text{Area under curve between } z_1 \text{ and } z_2}{\text{Total area under standard normal curve}} = \frac{\text{Area under curve between } x_1 \text{ and } x_2}{\text{Total area under normal curve}}$$

But the total area under each of these curves is one unit, so the numerators of these two curves must also be equal.

This implies that $P(x_1 < x < x_2) = P(z_1 < z < z_2)$.

Therefore, in dealing with any practical problem, begin by working out the standardised values of the boundaries of the ranges in the problem. Then, using these standardised values and the standard normal table, find the required probability as shown in our earlier examples.

Example 1

A firm of stockbrokers will on average handle 2,500 shares a day with a standard deviation of 250 shares. If the number of shares sold is normally distributed, find the answers to the following questions:

(a) What is the probability that more than 2,700 shares will be sold in one day?

(b) What is the probability that less than 1,900 shares are sold on any one day?

(c) What is the probability that the stockbrokers will sell between 2,300 and 2,550 shares a day?

(d) What is the probability that they will sell either more than 3,125 or less than 2,000 shares in a day?

Answers

In all the questions, let x = number of shares sold in a day and since μ = 2,500 and σ = 250, then

$z = \frac{x - 2{,}500}{250}$.

(a) $P(x > 2{,}700) = P\left(z > \frac{2{,}700 - 2{,}500}{250}\right)$

$= P(z > 0.8) = 0.5 - \Phi(0.8)$

$= 0.5 - 0.2881 = 0.2119$

(b) $P(x < 1{,}900) = P\left(z < \frac{1{,}900 - 2{,}500}{250}\right)$

$= P(z < -2.4) = P(z > 2.4)$

$= 0.5 - \Phi(2.4) = 0.5 - 0.4918 = 0.0082$

(c) $P(2{,}300 < x < 2{,}550) = P\left(\frac{2{,}300 - 2{,}500}{250} < z < \frac{2{,}550 - 2{,}500}{250}\right)$

$= P(-0.8 < x < 0.2) = \Phi(0.8) + \Phi(0.2)$

$= 0.2881 + 0.0793 = 0.3674$

(d) $P(x > 3{,}125) = P\left(z > \frac{3{,}125 - 2{,}500}{250}\right)$

$= P(z > 2.5) = 0.5 - \Phi(2.5) = 0.5 - 0.4938$

$= 0.0062$

$P(x < 2{,}000) = P\left(z < \frac{2{,}000 - 2{,}500}{250}\right)$

$= P(z < -2) = 0.5 - \Phi(2)$

$= 0.5 - 0.4772 = 0.0228$

$P(x < 2{,}000 \text{ or } x > 3{,}125) = P(x < 2{,}000) + P(x > 3{,}125)$ (mutually exclusive events)

$= 0.0228 + 0.0062 = 0.029$

Example 2

Computers consist of a number of components including what is called a memory. These memories, produced by an automatic process, have life length which is normally distributed with a mean of 500 hours and a standard deviation of 30 hours. If one thousand of these memories are selected at random from the production line, answer the following questions.

(a) How many of the memories would you expect to last for longer than 550 hours?

(b) How many memories would you expect to have a life of between 480 and 510 hours?

(c) How many memories would you expect to have a life of more than 560 hours or less than 440 hours?

Answers

In all these questions, let x = length of life in hours of a memory and, since $\mu = 500$ and $\sigma = 30$,

$z = \frac{x - 500}{30}$.

(a) $P(x > 550) = P\left(z > \frac{550 - 500}{30}\right)$

$= P(z > 1.667) = 0.5 - \Phi(1.67)$

$= 0.5 - 0.4525 = 0.0475$

i.e. the probability that one memory lasts longer than 550 hours is 0.0475.

Therefore, in a random sample of 1,000 memories you would expect 0.0475 × 1,000 to last longer than 550 hours, i.e. 47½.

(b) $P(480 < x < 510) = P\left(\frac{480 - 500}{30} < z < \frac{510 - 500}{30}\right)$

$= P(-0.667 < z < 0.333)$

$= \Phi(0.33) + \Phi(0.67)$

$= 0.1293 + 0.2486 = 0.3779$

i.e. the probability that one memory will last between 480 and 510 hours is 0.3779.

In a random sample of 1,000 memories you would expect 0.3779 × 1,000 of them to last between 480 and 510 hours, i.e. 377.9.

(c) $P(x > 560) = P\left(z > \frac{560 - 500}{30}\right) = P(z > 2)$ *(1)*

$P(x < 440) = P\left(z < \frac{440 - 500}{30}\right) = P(z < -2)$ *(2)*

Now you can continue this problem by calculating probabilities (1) and (2) in exactly the same way as in Example 1(a) or you can use the symmetry of the curve to take a short cut. The ordinates at 2 and –2 are symmetrically placed with respect to the mean (you can see these two ordinates in Figure 12.6) so the areas beyond them will be equal, i.e.

$P(z < -2 \text{ or } z > 2) = 2 \times P(z > 2)$

Thus, $P(x < 440 \text{ or } x > 560) = 2 \times P(z > 2) = 2(0.5 - \Phi(2))$

$= 2(0.0228) = 0.0456$

i.e. the probability that one memory will last longer than 560 hours or less than 440 hours is 0.0456.

Therefore, in a random sample of 1,000 memories you would expect 0.0456 × 1,000 to last longer than 560 hours or less than 440 hours, i.e. 45.6.

Note that you can very often save time by using the symmetry of the normal curve to reduce the number of calculations.

G. USE OF THEORETICAL DISTRIBUTIONS

Types of Distribution

The theoretical frequency distribution of a variable is, in fact, the distribution of the whole population of the values that the variable can take. The method of calculating these distributions divides them into two types:

(a) Those the populations of which consist of a known limited number of values of the variable, so that we can construct their theoretical frequency distributions from known probabilities; the rectangular distribution is an example of this type. Then we can use any sample data to test whether the assumptions we have made are correct.

(b) Those the populations of which consist of an unknown or unlimited number of values of the variable so that we have no prior knowledge of the probabilities. Then we have to use the empirical probabilities calculated from sample data to deduce the population distribution. We have already discussed one very important distribution of this type – the normal distribution.

Use in Statistical Inference

There are three main ways in which theoretical distributions and simple data are used in statistical inference. They are:

(a) To find the shape of a population distribution.

(b) To estimate the population parameters from the sample statistics.

(c) To test whether sample data come from a given population or whether two samples come from the same population.

More about this in a later Study Unit.

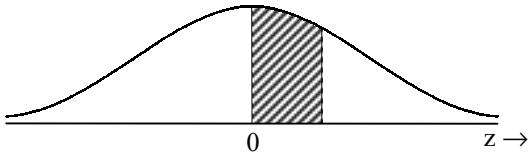
Use in This Course

Although you should appreciate the important part played by theoretical frequency distributions in statistical inference, for this course you will only need to know how to use these distributions to:

- Estimate the mean of a population.
- Test whether a sample comes from a population with a given mean.
- Test whether two samples come from the same population.

APPENDIX: STANDARD NORMAL TABLE – AREA UNDER THE NORMAL CURVE

This table gives the area under the normal curve between the mean and a point z standard deviations above the mean. The corresponding area for deviations below the mean can be found by symmetry.



$z = \frac{(x - \mu)}{\sigma}$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0159	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2133	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4430	.4441
1.6	.4452	.4463	.4474	.4485	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4762	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4865	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4072	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978.	.4979	.4980	.4980	.4981
2.9	.4981	.4982	.4983	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.49865	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.49903	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.49931	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.49952	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.49966	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.49977									

Study Unit 13

Probability Distributions

Contents *Page*

A. THE BINOMIAL EXPANSION

Some types of situation occur repeatedly in everyday life and there are various probability **distributions** that can be used to give us the probabilities associated with **all** the different possible outcomes under particular conditions. The first probability distribution we shall consider is the binomial distribution, but before we can continue with the probability aspect we need to remind ourselves of some algebra.

By multiplication, we can show that:

$$(a + b)^1 = a + b$$

$$(a + b)^2 = a^2 + 2ab + b^2$$

$$(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$$

$$(a + b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$$

All these products will be seen to fall under the general formula:

$$(a + b)^n = a^n + {}^nC_1a^{n-1}b + {}^nC_2a^{n-2}b^2 + {}^nC_3a^{n-3}b^3 + \ldots\ldots\ldots + b^n$$

$$= a^n + na^{n-1}b + \frac{n(n-1)}{1\times 2}a^{n-2}b^2 + \frac{n(n-1)(n-2)}{1\times 2\times 3}a^{n-3}b^3 + \ldots\ldots\ldots + b^n$$

We can check this as follows; if n = 4:

$$(a + b)^4 = a^4 + 4a^{4-1}b + \frac{4(4-1)}{1\times 2}a^{4-2}b^2 + \frac{4(4-1)(4-2)}{1\times 2\times 3}ab^3 + \frac{4(4-1)(4-2)(4-3)}{1\times 2\times 3\times 4}a^{4-4}b^4$$

$$= a^4 + 4a^3b + \frac{4\times 3}{1\times 2}a^2b^2 + \frac{4\times 3\times 2}{1\times 2\times 3}ab^3 + \frac{4\times 3\times 2\times 1}{1\times 2\times 3\times 4}a^0b^4$$

$$= a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$$

The best way to remember the general formula is in terms of combinations:

$$(a + b)^n = a^n + {}^nC_1a^{n-1}b + {}^nC_2a^{n-2}b^2 + {}^nC_3a^{n-3}b^3 + \ldots + {}^nC_ra^{n-r}b^r + \ldots + b^n$$

This is what is known as a **binomial expansion.** The binomial coefficients ${}^nC_1$, ${}^nC_2$, etc. are simply combinations:

i.e. ${}^nC_r = \dfrac{n!}{(n-r)!\ r!}$ and $n! = n(n-1)(n-2)\ldots\ldots 1$

0! is defined to equal 1.

Remember that ${}^nC_r$ is sometimes written $\left[\dfrac{n}{r}\right]$ or ${}^nC_r$.

Let us check this second version of the general formula again for n = 4:

From above: $^4C_1 = \dfrac{4!}{(4-1)!\ 1!} = \dfrac{4\times3\times2\times1}{3\times2\times1\times1} = 4$

$^4C_2 = \dfrac{4!}{(4-2)!\ 2!} = \dfrac{4\times3\times2\times1}{2\times1\times2\times1} = 6$

$^4C_3 = \dfrac{4!}{(4-3)!\ 3!} = \dfrac{4\times3\times2\times1}{1\times3\times2\times1} = 4$

$^4C_4 = \dfrac{4!}{(4-4)!\ 4!} = \dfrac{4\times3\times2\times1}{1\times4\times3\times2\times1} = 1$

We thus get:

$(a + b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$, as before.

If we extract the coefficients from each binomial expansion for successive values of n, we can arrange them in a special triangular form known as **Pascal's Triangle**:

Expansion	**Coefficients**
$(a + b)^1 = a + b$	1 1
$(a + b)^2 = a^2 + 2ab + b^2$	1 2 1
$(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$	1 3 3 1
$(a + b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$	1 4 6 4 1

The missing expansion is $(a + b)^0$ which equals 1, and this value can be inserted, giving the following triangular figure (Figure 13.1):

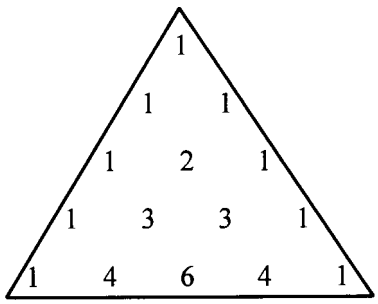


Figure 13.1: Pascal's Triangle

1 is always the outside figure on every line. The total number of figures increases by 1 on each line. The new inside values are found by adding together, in turn, consecutive pairs of figures in the previous row. Thus, from above, the next row would be 1, 1 + 4, 4 + 6, 6 + 4, 4 + 1, 1, i.e. 1, 5, 10, 10, 5, 1. This tells us without any more working that:

$(a + b)^5 = a^5 + 5a^4b + 10a^3b^2 + 10a^2b^3 + 5ab^4 + b^5$

Check for yourself that the next row is 1, 6, 15, 20, 15, 6, 1 and hence, write down the binomial expansion of $(a + b)^6$.

B. GENERAL FORMULA FOR THE BINOMIAL DISTRIBUTION

Introduction

The binomial probability distribution is applicable in situations where an experiment, consisting of a certain number of trials, satisfies **all of four conditions**. These are:

(a) The number of trials must be fixed and finite. This number is usually denoted by n.

(b) Every trial must result in one or other of only two mutually exclusive possible outcomes, which, for convenience, we usually label "success" or "failure". Examples are:

(i) When we roll a die we get a six or we do not get a six.

(ii) A product is either defective or not defective.

(iii) A tossed coin comes down heads or tails.

(iv) A child is either a boy or a girl.

We must, of course, define which event is the success before the term is used.

(c) The probability of a success or failure at each trial must remain constant throughout the experiment. The probability of a success is usually denoted by p, and that of a failure by q.

(d) The outcome of every trial must be independent of the outcome of every other trial. For example, if a coin is unbiased, then the probability of obtaining a head on the tenth time it is tossed remains ½, even if the previous nine tosses have all resulted in heads.

Provided these four conditions all hold, the binomial distribution enables us to work out the probability of any given number of successes or failures in a specified number of trials.

Die-Throwing Experiments

We shall work out from first principles the probabilities for the situation where we throw a fair die and observe whether we obtain a 6 or not. We shall label obtaining a 6 as success and not obtaining a 6 as failure.

For a fair die, P(Success) = $\frac{1}{6}$, P(Failure) = $\frac{5}{6}$

and we can draw a simple tree diagram (Figure 13.2):

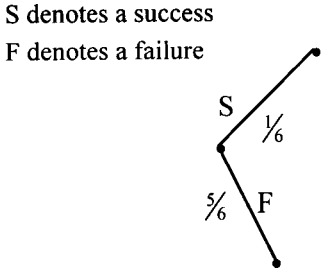


Figure 13.2: Die-Throwing Experiments – 1st Throw

Thus, for one throw of the die we can write:

$$P(0 \text{ successes}) = \frac{5}{6} \quad P(1 \text{ success}) = \frac{1}{6}$$

Let us now throw the die again. The probability of a success or failure at this second throw remains unchanged at $\frac{1}{6}$ or $\frac{5}{6}$ respectively. We can extend our tree diagram as follows (Figure 13.3):

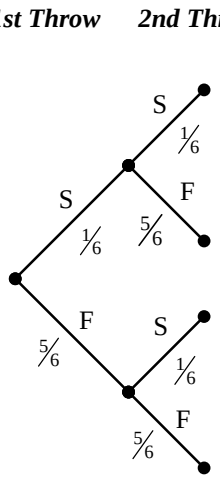


Figure 13.3: Die-Throwing Experiments – 2nd Throw

At the end of 2 throws, the different possible outcomes are as in Table 13.1.

Table 13.1

1st throw	2nd throw	Number of successes
6	6	2
6	Not a 6	1
Not a 6	6	1
Not a 6	Not a 6	0

We can put our results in a table.

Table 13.2

Event	Probability
0 successes	$\left[\frac{5}{6}\right]^2$
1 success	$\left[\frac{1}{6}\right]\left[\frac{5}{6}\right]+\left[\frac{5}{6}\right]\left[\frac{1}{6}\right]$ $=2\left[\frac{5}{6}\right]\left[\frac{1}{6}\right]$
2 successes	$\left[\frac{1}{6}\right]^2$

We then throw the die once more. The different possible outcomes after 3 throws are best found by looking at the tree, extended one stage more (see Figure 13.4 and Table 13.3):

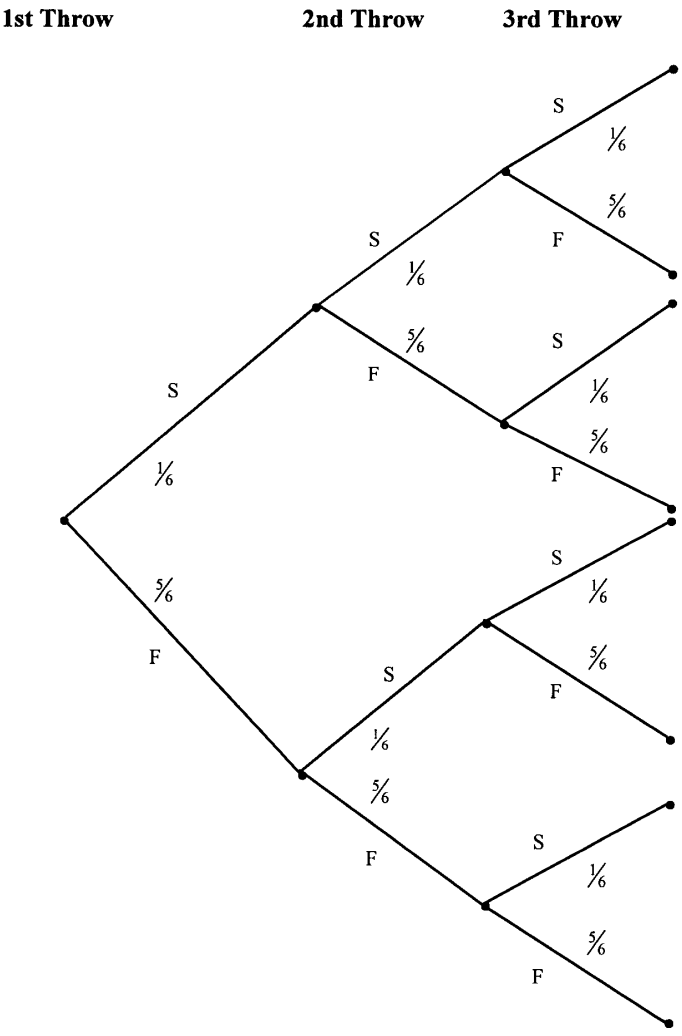


Figure 13.4: Die-Throwing Experiments – 3rd Throw

Table 13.3

Outcomes	Probability
SSS	$\left[\frac{1}{6}\right]^3$
SSF	$\left[\frac{1}{6}\right]^2\left[\frac{5}{6}\right]$
SFS	$\left[\frac{1}{6}\right]\left[\frac{5}{6}\right]\left[\frac{1}{6}\right]$
SFF	$\left[\frac{1}{6}\right]\ \left[\frac{5}{6}\right]^2$
FSS	$\left[\frac{5}{6}\right]\left[\frac{1}{6}\right]^2$
FSF	$\left[\frac{5}{6}\right]\left[\frac{1}{6}\right]\left[\frac{5}{6}\right]$
FFS	$\left[\frac{5}{6}\right]^2\left[\frac{1}{6}\right]$
FFF	$\left[\frac{5}{6}\right]^3$

and we can condense the results in a second table:

Table 13.4

Event	Probability
0 successes	$\left[\frac{5}{6}\right]^3$
1 success	$3\left[\frac{5}{6}\right]^2\left[\frac{1}{6}\right]$
2 successes	$3\left[\frac{5}{6}\right]\left[\frac{1}{6}\right]^2$
3 successes	$\left[\frac{1}{6}\right]^3$

Now try to extend the tree diagram one stage further corresponding to 4 throws of the die. List all the possible outcomes in a table with their probabilities and then construct a second table giving the probabilities of 0, 1, 2, 3 and 4 successes.

General Binomial Experiment

In a general experiment where each trial has just 2 mutually exclusive outcomes, we can denote

P(Success) = p and P(Failure) = q

We know that p + q = 1, as one or other of these outcomes is **certain to happen** (see Figure 13.5 and Table 13.5). Hence,

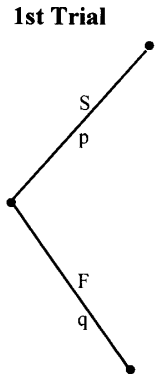


Figure 13.5: General Binomial Experiment – 1st Trial

Table 13.5

Event	Probability
0 successes	q
1 success	p

The trial is performed again (see Figure 13.6, Tables 13.6 and 13.7):

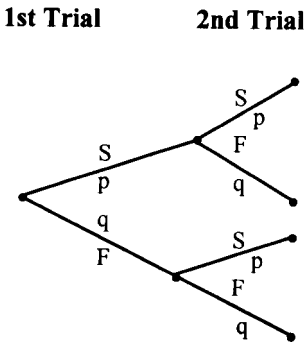


Figure 13.6: General Binomial Experiment – 2nd Trial

Table 13.6

Outcomes	Probability
SS	p^2
SF	pq
FS	qp
FF	q^2

Table 13.7

Event	Probability
0 successes	q^2
1 success	2qp
2 successes	p^2

and for a third time (see Figure 13.7, Tables 13.8 and 13.9):

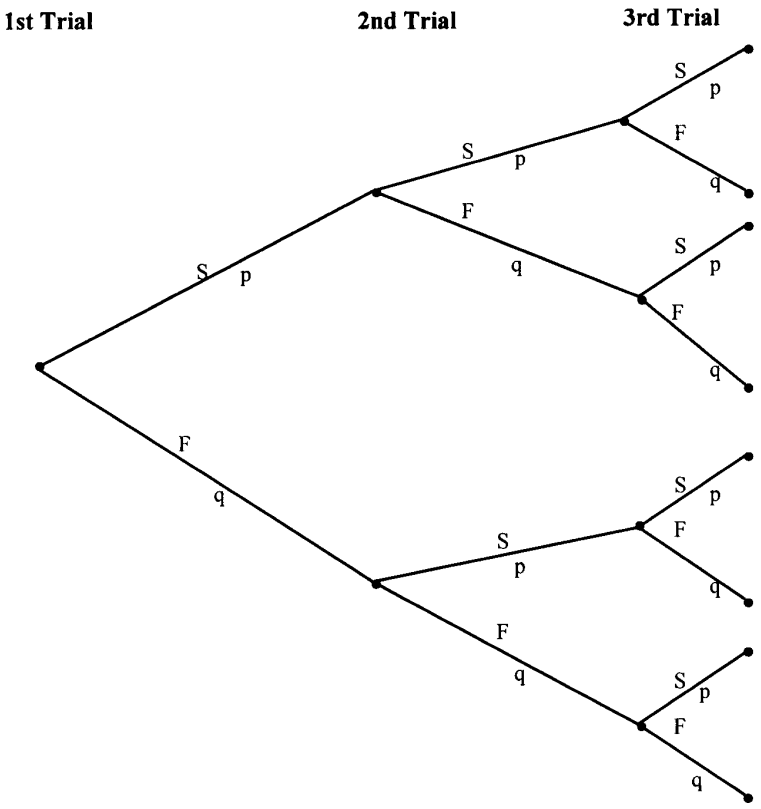


Figure 13.7: General Binomial Experiment – 3rd Trial

Table 13.8

Outcomes	Probability
SSS	p^3
SSF	p^2q
SFS	pqp
SFF	pq^2
FSS	qp^2
FSF	qpq
FFS	q^2p
FFF	q^3

Table 13.9

Event	Probability
0 successes	q^3
1 success	$3q^2p$
2 successes	$3qp^2$
3 successes	p^3

Now look back to the first section of this study unit and write down the binomial expansions of:

$(q + p)^2$ and $(q + p)^3$

You will see that the terms in the expansions are exactly the same as the probabilities we have worked out above.

See if you can write down, without using a tree diagram, the probabilities of 0, 1, 2, 3, 4 successes, when we perform this trial 4 times.

Although we can use tree diagrams to work out probabilities when we repeat our trial up to 4 or 5 times, once we have larger numbers of repetitions of the trial, the tree diagrams become unwieldy. It is then much better to use what we have discovered above. That is, if we perform our trial n times, the probabilities of 0, 1, 2, 3 .... up to n successes are given by successive terms in the binomial expansion $(q + p)$, where p is the probability of success and q the probability of failure at any one trial. Note that p and q must remain constant from trial to trial.

From our formula for the binomial expansion, the general term in the expansion of $(q + p)^n$ is ${}^nC_rq^{n-r}p^r$ and this gives us the probability of exactly r successes in n trials of the experiment. If we call x the number of successes, then we have:

$P(x = r) = {}^nC_rq^{n-r}p^r$ for r = 0, 1, 2.....n

P(r) is often used instead of P(x = r). As q always equals 1 – p, you will meet the formula on your examination paper as:

$P(r) = {}^nC_r(1 - p)^{n-r}p^r = {}^nC_rp^r(1 - p)^{n-r}$

This is the general formula for the binomial probability distribution. We shall see how to apply this formula in the next section, and you will find that it is not quite so fearsome as it may look at first.

C. APPLICATIONS OF THE BINOMIAL DISTRIBUTION

Example 1

The probability that a match will break on being struck is 0.04. What is the probability that, out of a box of 50:

(a) None will break;

(b) More than 2 will break?

A match will either break or not break when it is struck. Therefore:

P(Breaking) = 0.04 = P(Success) = p

P(Not breaking) = 1 – p = 1 – 0.04

= 0.96 = P(Failure)

We have a box of 50 matches, so n = 50.

(a) We require the probability that none will break, i.e. the probability of no successes, P(x = 0).

$P(x = 0) = P(0) = {}^{n}C_0 p^0 (1-p)^{n-0}$ using general formula

$$= \frac{n!}{n!\,0!}(1-p)^n = (1-p)^n$$

$$= (0.96)^{50} = 0.1299 \text{ to 4 dec. places.}$$

Therefore, probability none will break = 0.1299

(b) Probability that more than 2 will break = 1 – Probability that 2 or less will break

Probability that 2 or less will break = Probability that 0 or 1 or 2 will break

= P(0) + P(1) + P(2)

We thus need to work out P(x = 1) and P(x = 2):

$P(1) = {}^{n}C_1 p^1 (1-p)^{n-1}$

$$= \frac{n!}{(n-1)!}\,p(1-p)^{n-1} = \frac{50!}{49!}(0.04)(0.96)^{49}$$

$$= 50 \times (0.04)(0.96)^{49} = 0.2706 \text{ to 4 dp}$$

$P(2) = {}^{n}C_2 p^2 (1-p)^{n-2}$

$$= \frac{n!}{(n-2)!2!}p^2(1-p)^{n-2} = \frac{50!}{48!\,2!}(0.04)^2(0.96)^{48}$$

$$= \frac{50 \times 49}{2}(0.04)^2(0.96)^{48} = 0.2762$$

Therefore, probability that more than 2 will break = 1 – (0.1299 + 0.2706 + 0.2762)

= 1 – 0.6767 = 0.3233

= 0.323 to 3 dp

Example 2

It has been found that, on average, 5% of the eggs supplied to a supermarket are cracked. If you buy a box of 6 eggs, what is the probability that it contains 2 or more cracked eggs?

An egg is either cracked or not cracked:

P(Cracked) = 5% = 0.05 = P(Success) = p

P(Not cracked) = 1 – p = 1 – 0.05 = 0.95 = P(Failure)

We have a box of 6 eggs, so n = 6.

Probability of 2 or more cracked eggs in a box = 1 – Probability of less than 2 cracked eggs in a box

= 1 – Probability of 0 or 1 cracked eggs in a box

= 1 – [P(0) + P(1)].

$P(0) = {}^{n}C_0 p^0 (1-p)^{n-0} = (1-p)^n = (0.95)^6 = 0.7351$

$P(1) = {}^{n}C_1 p^1 (1-p)^{n-1} = \frac{n!}{(n-1)!\,1!} p(1-p)^{n-1}$

$= \frac{6!}{5!}(0.05)(0.95)^5 = 6(0.05)(0.95)^5 = 0.2321$

Therefore, probability of 2 or more cracked eggs in a box = 1 – (0.7351 + 0.2321)

= 1 – 0.9672 = 0.0328 = 0.033 to 3 dp

Example 3

A retail sales manager will accept delivery of a large consignment of goods if a random sample of 10 items contains no defectives. If 3% of the producer's total output is defective, what is the probability that delivery of a consignment will be accepted? How would the situation change if the random sample were of only 5 items?

An item is either defective or non-defective. Therefore:

P(Defective) = 3% = 0.03 = P(Success) = p

P(Non-defective) = 1 – p = 1 – 0.03 = 0.97 = P(Failure)

First, he takes a sample of 10, so n = 10.

We require the probability that this sample contains no defectives, i.e. P(0):

$P(0) = {}^{n}C_0 p^0 (1-p)^{n-0} = (1-p)^n$

$= (0.97)^{10}$

= 0.7374 to 4 dp

Therefore, probability that a delivery will be accepted is 0.7374.

Secondly, consider a sample of 5.

$P(0) = (1-p)^n = (0.97)^5 = 0.8587$ to 4 dp

Therefore, probability that delivery will be accepted is 0.8587, which is higher than when a larger sample was taken.

Notes

(1) One of the conditions for using the binomial distribution is that the chance of success, p, must be **constant** throughout the series of trials. This means that if we are, say, taking items from a batch and **not** replacing them before the next item is taken, then the binomial distribution does not apply because the batch is fractionally smaller (by 1 item each time). In practice, however, when the batch from which a sample is being taken is very large compared with the sample, the binomial distribution is a satisfactory approximation. As a rough guide, you can consider the batch to be very large if it is more than about 10 times the sample size.

(2) Tables are available giving values of ${}^{n}C_{1}$ for various values of n and r. This is particularly useful for large values of n but in examinations usually you are expected to be able to work them out for yourself.

D. MEAN AND STANDARD DEVIATION OF THE BINOMIAL DISTRIBUTION

Introduction

The binomial distribution is a probability distribution, and this is just another way of saying that it is a relative frequency distribution, as we can work out the **proportion** of times we get a given number of successes. Remember that with frequency distributions we know the **number** of times a given event has happened, i.e. the frequency of the event. For frequency distributions we can work out the arithmetic mean:

$$\bar{x} = \frac{\sum fx}{\sum f}$$

and the standard deviation:

$$\sqrt{\frac{f(x-\bar{x})^2}{f}} = \sqrt{\frac{\sum fx^2}{\sum f} - \left[\frac{\sum fx}{\sum f}\right]^2}$$

Similarly, we can work out the arithmetic mean of a probability distribution, denoted by the Greek letter μ (mu), and the standard deviation, denoted by σ (small sigma). The formulae in this case are:

$$\mu = \frac{\sum xP(x-x)}{\sum P(x-x)} \times \frac{\sum xP(x)}{\sum P(x)}$$

and $$\sigma = \sqrt{\frac{\sum (x-\mu)^2 P(x-x))}{P(x-x)}} = \sqrt{\frac{\sum (x-\mu)^2 P(x)}{\sum P(x)}}$$

i.e. where we had the frequency f previously, we have the probability (i.e. relative frequency), P(x). (It is usual to use x instead of r in this situation, i.e.

$P(x) == {}^{n}C_{x}p^{x}(1-p)^{n-x}$, for a binomial distribution.)

The sum of all the probabilities, P(x), must always be 1, as we are certain that one of the outcomes is bound to happen. Hence we can simplify the formulae to give:

$$\mu = \sum xP(x)$$

$$\sigma = \sqrt{\sum (x-\mu)^2 P(x)}$$

Reminder: variance = (standard deviation)^2 = σ^2

Die-Throwing Experiment

Let's consider again one of the particular binomial distributions we obtained earlier, i.e. that corresponding to 3 tosses of a die, when the successful outcome was obtaining a 6.

We obtained the results shown in Table 13.10:

Table 13.10

Event	Probability
0 successes	$\left[\frac{5}{6}\right]^3$
1 success	$3\left[\frac{5}{6}\right]^2\left[\frac{1}{6}\right]$
2 successes	$3\left[\frac{5}{6}\right]\left[\frac{1}{6}\right]^2$
3 successes	$\left[\frac{1}{6}\right]^3$

We will now evaluate these probabilities and construct another table similar to those we used when calculating means and standard deviations of frequency distributions (see Table 13.11):

Table 13.11

Number of successes x	P(x)	x.P(x)	$x - \mu$	$(x-\mu)^2$	$(x-\mu)^2 P(x)$
0	$\frac{125}{216}$		$-\frac{1}{2}$	$\frac{1}{4}$	$\frac{125}{864}$
1	$\frac{75}{216}$	$\frac{75}{216}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{75}{864}$
2	$\frac{15}{216}$	$\frac{30}{216}$	$\frac{3}{2}$	$\frac{9}{4}$	$\frac{135}{864}$
3	$\frac{1}{216}$	$\frac{3}{216}$	$\frac{5}{2}$	$\frac{25}{4}$	$\frac{25}{864}$
Totals	Check: 1	$\frac{108}{216}=\frac{1}{2}$			$\frac{360}{864}=\frac{15}{36}$

Mean, $\mu = \sum xP(x) = \frac{1}{2} = 0.5$.

Standard deviation, $\sigma = \sqrt{\frac{15}{36}} = 0.6454 = 0.645$ to 3 sig. figs.

When evaluating probabilities you will find that it is more accurate to work, if possible, in terms of exact fractions. If this is not possible, however, carry all your working to 4 significant figures and round the final result to 3.

In the case of a theoretical distribution like the binomial distribution there is usually a simple formula to show what the mean and standard deviation ought to be without needing to go through the lengthy calculation above.

For any binomial distribution:

$$\mu = np$$

$$\sigma = \sqrt{npq} = \sqrt{np(1-p)}$$

In our example above n = 3, p = $\frac{1}{6}$, 1 – p = $\frac{5}{6}$

Therefore:

$$\mu = 3 \times \frac{1}{6} = \frac{1}{2} = 0.5$$

$$\sigma = \sqrt{3 \times \frac{1}{6} \times \frac{5}{6}} = \sqrt{\frac{15}{36}} = 0.645 \text{ to 3 sig. figs}$$

Thus the result agrees with our earlier calculation. What we are saying is that if we repeat this experiment very many times, the mean number of successes is 0.5. As with a frequency distribution, the mean does not necessarily have to be one of the original values of x.

QUESTIONS FOR PRACTICE

Now try these questions and check your answers against those given at the end of the study unit.

1. A fair die with 6 sides is thrown 3 times. Show by means of a tree diagram that the probability of obtaining 0, 1, 2 or 3 sixes from the 3 throws is given by the binomial probability function:

   $${}^{3}C_{r}\left[\frac{1}{6}\right]^{r}\left[\frac{5}{6}\right]^{3-r}$$

   where r represents the number of successes.

2. A department produces a standard product. It is known that 60% of defective products can be satisfactorily reworked. What is the probability that in a batch of 5 such defective products, at least 4 can be satisfactorily reworked?

3. What is the probability of passing a batch of 5 units from a machine which averages 20% defectives, when only 2 of the 5 are tested?

4. Calculate the probability that, for 6 telephone lines:

   (a) At least 1 of the lines is engaged and

   (b) All 6 lines are engaged,

   when the probability of 1 line being engaged is ¼.

5. In a family with 10 children, if the probability of having a male child is the same as that of having a female child, what is the probability that:

   (a) 6 of the children will be boys?

   (b) None will be a girl?

   (c) At most, 2 will be boys?

6. This question tests your understanding of probability theory and the binomial distribution.

   The World Life Assurance Company Limited uses recent mortality data in the calculation of its premiums. The following table shows, per thousand of population, the number of persons expected to survive to a given age:

Age	0	10	20	30	40	50	60	70	80	90	100
Number surviving to given age	1,000	981	966	944	912	880	748	525	261	45	0

   Required:

   (a) Use the table to determine the probability that:

      (i) A randomly chosen person born now will die before reaching the age of 60.

      (ii) A randomly chosen person who is aged 30 now will die before reaching the age of 60.

      (iii) A randomly chosen person who is aged 50 now will die before reaching the age of 60.

      Comment on the order of magnitude of your three answers.

   (b) The company is planning to introduce a life insurance policy for persons aged 50. This policy requires a single payment paid by the insured at the age of 50 so that if the insured dies within the following ten years the dependant will receive a payment of £10,000; however, if this person survives, then the company will not make any payment.

      Ignoring interest on premiums and any administrative costs, calculate the single premium that the company should charge to break even in the long run.

   (c) If 12 people each take out a policy as described in (b) and the company charges each person a single premium of £2,000, find the probability that the company will make a loss.

   (d) The above table was based on the ages of all people who died in 1986. Comment on the appropriateness to the company when calculating the premiums it should charge.

   (e) The above table can be expanded to include survival numbers, per thousand of population, of other ages:

Age	50	52	54	56	58	60
Number surviving to given age	880	866	846	822	788	748

(i) Given that a person aged 50 now dies before the age of 60, use this new information to estimate the expected age of death.

(ii) Calculate a revised value for the single premium as described in part (b), taking into account the following additional information:

- The expected age of death before 60 as estimated at (i)
- A constant interest rate of 8% p.a. on premiums
- An administration cost of £100 per policy
- A cost of £200 to cover profit and commissions.

E. THE POISSON DISTRIBUTION

Introduction

The binomial distribution is useful in cases where we take a fixed sample size, and count the number of successes. Sometimes we do not have a definite sample size, and then the binomial distribution is of no use. For example, if we are studying the occurrence of accidents, we can count how many accidents occur in a month, but it is nonsense to talk about how many accidents did not occur! In such cases we use another theoretical distribution called the POISSON DISTRIBUTION, after a French mathematician.

The Exponential Function

Before we go on to study the Poisson distribution, there is some new mathematics to learn. In mathematics there are a few rather special numbers which are given special symbols. One of these you probably know very well already; that is π (pronounced "pie") which is used to calculate areas of circles. π cannot be given an **exact** arithmetical value, but it is very nearly 3.1416.

Another of these special symbols is e, and this is called the exponential number. Like π, it cannot be given an exact arithmetical value (although mathematicians can calculate it to as many decimal places as they wish). To 3 decimal places, the value of e is 2.718, which is accurate enough for almost all practical purposes.

Formula for Poisson Distribution

In the example of accidents mentioned above, we could count the number of accidents each month and work out the **mean number** of accidents per month. So, although we do not know the values of n or p, we **do** know the value of the mean np.

Mathematicians can prove that, if you know the value of this mean (let's call it m) then the theoretical probability that there will be r events (or "successes", if you prefer to keep the same word as before) is:

$$P(r) = \frac{e^{-m}m^r}{r!}$$

This is the general term of the Poisson distribution. Sometimes μ or λ (lambda) are used in place of m but the formula shown here is the one you are likely to be given in the examination.

Earlier we gave the conditions which must prevail in order that a binomial distribution of events can occur. The Poisson distribution is what is known as a limiting case of the binomial distribution. It is the result of letting n become very large, and p become very small (i.e. a very large number of trials

are conducted but the probability of a success in any one particular trial is exceedingly small), but in such a way that the mean number of successes, np, is of a moderate size and constant. What exactly constitutes "moderate size" is subjective, but certainly you need have no cause for concern working with any value less than 10.

In terms of our accident example, we could regard n, the number of trials, as being the number of one-minute intervals in a working month, during any one of which an accident could occur; p would then be the probability of an accident occurring during any particular one-minute interval, and np the mean number of intervals per month during which accidents would occur. This will be identical with the mean number of accidents per month, provided we make the assumption that there are no intervals in which more than one accident occurs (and this is a reasonable assumption).

The Poisson probability formula, as stated above, is obtained from the corresponding binomial formula by letting n tend towards infinity, p tend towards zero, and substituting m = np.

If the Poisson distribution is applicable in any given situation, this means that all the main conditions affecting the issue (in our example, the accident rate) do not alter. The month-to-month variations would then be due only to random chance causes, and they would be expected to fit a Poisson distribution. Other examples of the Poisson distribution will be given later.

If the Poisson distribution gives a good description of the practical situation, then the events are occurring randomly in time and there is no underlying reason why there should be more accidents in one particular month of the year. Any fluctuations in numbers of accidents are assignable in such cases to random variation.

The mean of a Poisson distribution is usually indicated on your exam paper by the letter m, and its value is all you need to know when calculating the probabilities – unlike the binomial, where you need two things, n and p. It can be proved that the **variance** of a Poisson distribution is also equal to m, which means that the standard deviation is $\sqrt{m}$. There is no upper limit to r but, in practice, there is a value beyond which the probability is so low that it can be ignored.

F. APPLICATION OF THE POISSON DISTRIBUTION

Applying the Poisson distribution to our accidents example, we might collect data over a year or so and find that the mean number of accidents in a workshop is 2.2 per month. We might now ask ourselves what the probability is of getting a month with no accidents, or only 1 accident, or 2 accidents, and so on. To get the answer, we work out the appropriate terms of the Poisson distribution with mean m = 2.2.

To find the value of $e^{-2.2}$ there are several options open to you:

(a) Some books of tables give you the values of e^{-x} directly.

You will find $e^{-2.2} = 0.1108$.

Other books of tables give e^{-x} for $0 < x < 1$ together with e^{-x} for x = 1, 2, 3 .... . Thus, to obtain $e^{-2.2}$, we have to use properties of indices, i.e.:

$$e^{-2.2} = e^{-2} \times e^{-0.2} = 0.13534 \times 0.8187$$

$$= 0.1108 \text{ to 4 sig. figs,}$$

(b) Some calculators have the facility to work out e^{-x} directly. Others have a y^x facility so you can put in y = e (i.e. 2.718) and x = –2.2 to evaluate $e^{-2.2}$.

(c) If neither of the above options is available, then you have to go back to log tables and use logs to work out e^{-m} by multiplying log e by –m and then antilogging. The only difficulty here is to remember to put the log in bar number form before antilogging, as only the whole number part of a log can be negative.

To find $e^{-2.2}$ using logs:

$\log e = 0.4343 \times -2.2 = -0.95546 = -1 + 0.04454 = \bar{1}.04454$

Antilog = 0.1108

Therefore $e^{-2.2} = 0.1108$

We can now proceed to work out the Poisson probabilities, $P(r) = \frac{e^{-m}m^r}{r!}$, for different values of r.

If you require the probability for just one or two values of r, then this expression may be evaluated fully in each case.

$$P(6) = \frac{\left(e^{-2.2}\right)(2.2)^6}{6!} = \frac{0.1108 \times 113.4}{720} = 0.0174$$

However, if you require a large number of terms of the probability distribution, it is not necessary to evaluate each term separately.

Using the Poisson probability formula, we have:

$$P(r-1) = \frac{e^{-m}m^{r-1}}{(r-1)!}$$

$$P(r) = \frac{e^{-m}m^r}{r!} = \frac{e^{-m}m^{r-1}}{(r-1)!} \times \frac{m}{f}$$

$$= P(r-1) \times \frac{m}{r}.$$

Therefore, the probability of r successes can be obtained by multiplying the probability of (r – 1) successes by the value $\frac{m}{r}$.

This technique is illustrated in the following table, using our accident example in which m = 2.2.

Table 13.12

r	P(r)		
0	$\frac{e^{-2.2} \times 2.2^0}{0!} = e^{-2.2}$	=	0.1108
1	$P(0) \times \frac{2.2}{1} = 0.1108 \times 2.2$	=	0.2438
2	$P(1) \times \frac{2.2}{2} = 0.2438 \times 1.1$	=	0.2682
3	$P(2) \times \frac{2.2}{3} = 0.2682 \times 0.7333$	=	0.1967
4	$P(3) \times \frac{2.2}{4} = 0.1967 \times 0.55$	=	0.1081
5	$P(4) \times \frac{2.2}{5} = 0.1081 \times 0.44$	=	0.0476
6	$P(5) \times \frac{2.2}{6} = 0.0476 \times 0.3667$	=	0.0174
7	$P(6) \times \frac{2.2}{7} = 0.0174 \times 0.3143$	=	0.0055
etc.			

Note that the value obtained for P(6) by this method is identical with that obtained by using the formula.

The values of P(r) are the probabilities that a particular month will have r accidents; or you may consider them as the relative frequencies with which months have 0, 1, 2, 3, etc. accidents. Note that, since our mean, 2.2, refers to per month, then the probabilities also refer to months. If we had used the mean per week, then the probabilities would have referred to weeks.

G. APPROXIMATION TO A BINOMIAL DISTRIBUTION

There is another very common use for the Poisson distribution, and that is as an approximation to the binomial distribution. If n is large, the calculations for the binomial distribution are often very tedious, even with a calculator. Fortunately, if we put m = np, the Poisson terms can, in certain circumstances, be used instead of the binomial terms. The conditions for using the Poisson as an approximation to the binomial are that:

- n should be large;
- p should be small, or q = 1 – p should be small.

Compare these conditions with those given earlier for the Poisson distribution.

There are no fixed rules to say what is meant by large and small, but it has been shown that the Poisson is a good approximation to the binomial if:

$n \geq 10$ and $p \leq 0.01$ *or* $n \geq 20$ and $p \leq 0.03$

or $n \geq 50$ and $p \leq 0.05$ *or* $n \geq 100$ and $p \leq 0.08$

Table 13.13 shows some probabilities for comparison.

Table 13.13

r	**Binomial** *n = 100, p = 0.01 np = 1*	**Binomial** *n = 50, p = 0.02 np = 1*	**Poisson** *m = 1*
0	0.3660	0.3642	0.3679
1	0.3697	0.3716	0.3679
2	0.1849	0.1858	0.1840
3	0.0610	0.0607	0.0613
4	0.0149	0.0145	0.0153
5	0.0029	0.0027	0.0031

It is also interesting to compare the histograms given by the binomial and Poisson distributions (Figures 13.8 and 13.9).

Whereas in Figure 13.8, where p is not small, the probabilities are noticeably different, in Figure 13.9, where p is smaller although n remains the same, the probabilities are in much closer agreement.

Note that the binomial and Poisson distributions are both discrete, i.e. there are gaps between consecutive values which a discrete variable may take. It is only possible, for example, to have 1, 2, 3, etc. accidents, not 1.1 or 2.7, etc. However, it is essential that in a histogram there should be no gaps between adjacent blocks. This problem is overcome by assuming that each value extends halfway to meet its neighbour in both directions.

Thus in the histograms which follow, each block representing the value 2 has been drawn to cover all values from 1.5 to 2.5, and similarly for other blocks, thus eliminating any gaps which would otherwise be present.

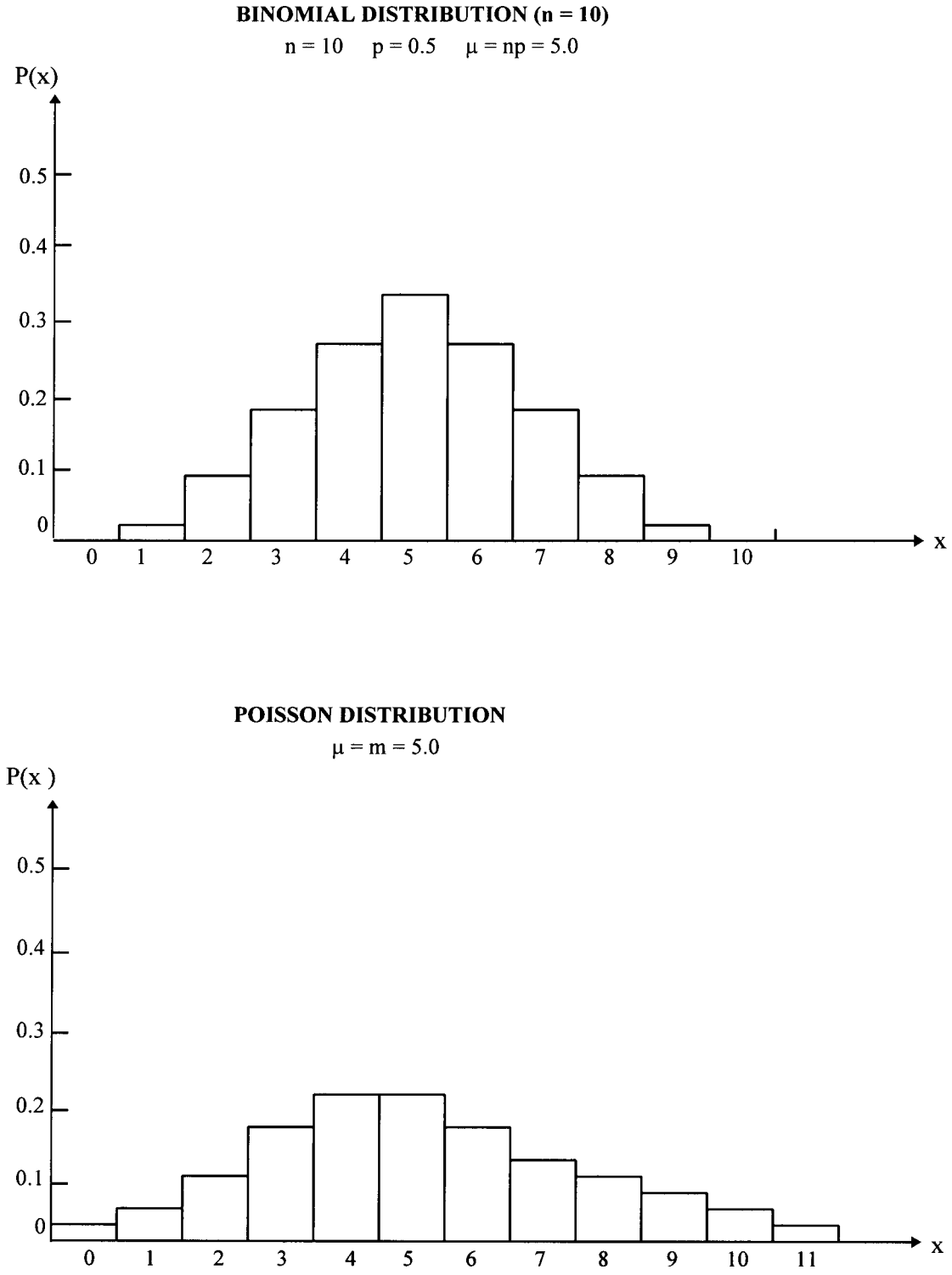
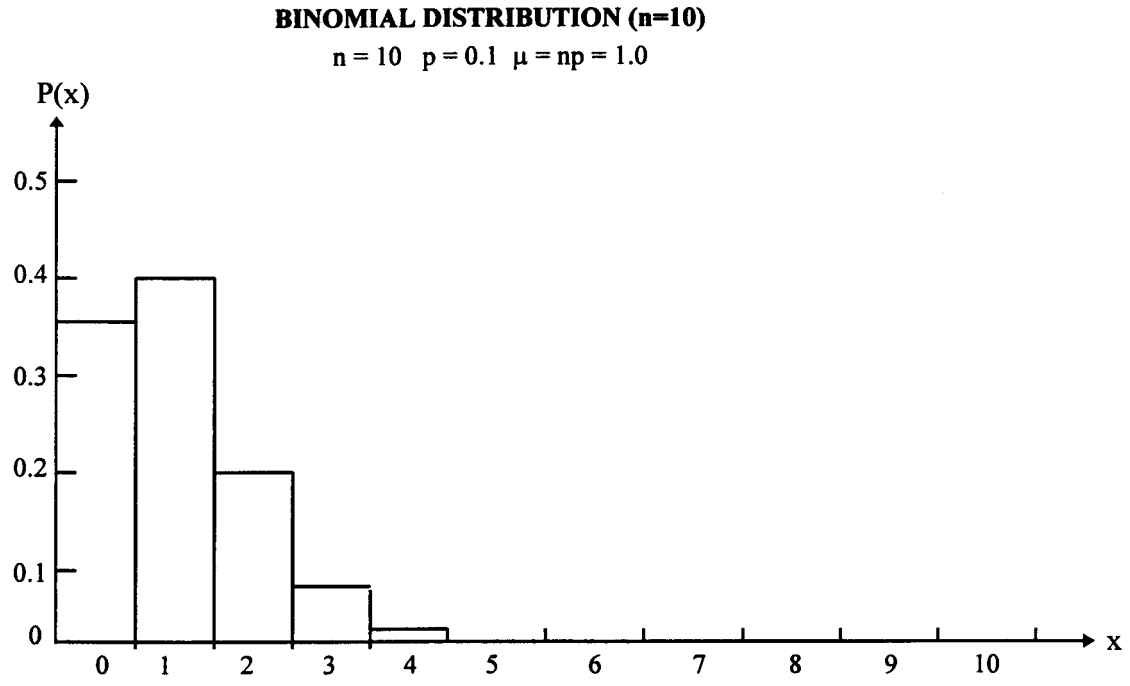


Figure 13.8: Comparative Histograms



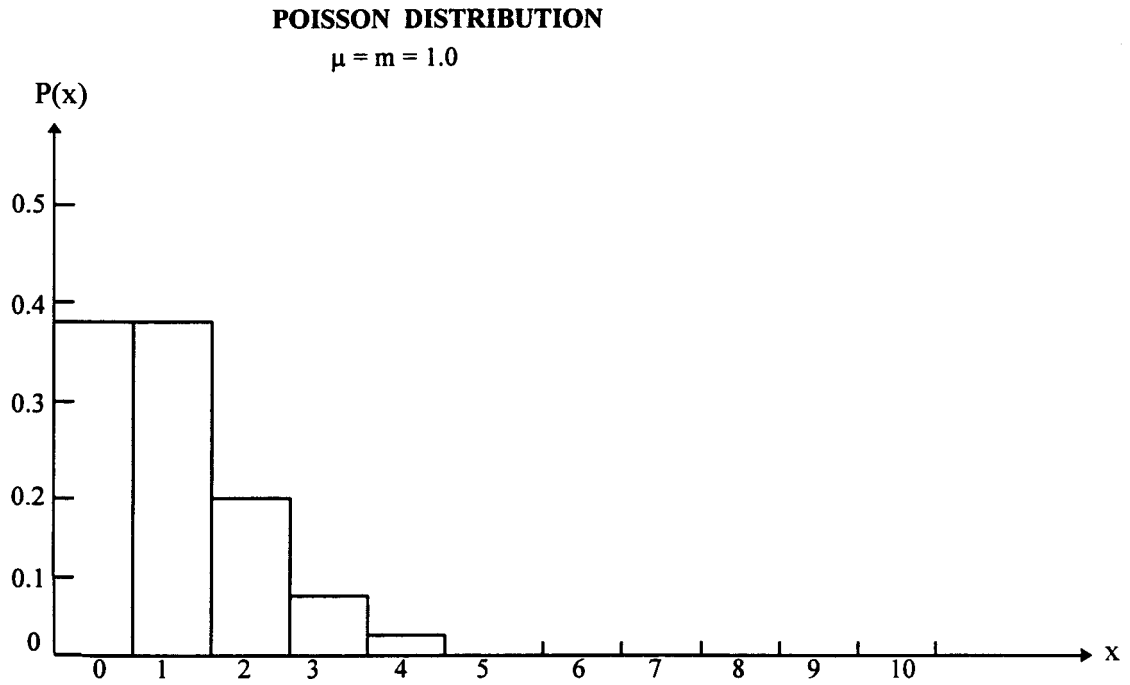


Figure 13.9: Comparative Histograms

H. APPLICATION OF BINOMIAL AND POISSON DISTRIBUTIONS – CONTROL CHARTS

One of the most important applications of the binomial and Poisson distributions is in **quality control**. You will probably have heard of quality control sampling schemes – where a sample of product is taken at regular intervals and checked for defectives. Suppose samples of 40 pieces are taken regularly from a production process. Past experience has shown that, in samples of 40, a defective piece can be expected in every other sample. Now let's see how the binomial and Poisson distributions can be used in order to develop a quality control chart.

The average number of defectives in a sample of 40 pieces is 0.5 (one defective unit in every other batch). The probability of finding 0, 1, 2, 3 ... defectives in a given sample could be accurately determined using the binomial expansion but, as the probability of finding a defective is relatively small and the sample of reasonable size, the Poisson distribution can be used as an approximation to the binomial. The individual probabilities for a Poisson distribution with mean 0.5 are as follows:

Number of defectives	Probability
0	0.6065
1	0.3033
2	0.0758
3	0.0126
4	0.0016
5	0.0002
	1.0000

(You may wish to practice use of the Poisson distribution by evaluating the probability formula $\frac{e^{-m}m^{r}}{r!}$ for r = 0, 1, 2 ... .)

You can see that the chance of a sample containing three or more defectives is about 1.4%, whilst the chance of a sample containing four or more defectives is only 0.2% (about twice in every thousand sample checks). Typically, a control chart would be set up identifying an **upper warning limit**, where the probability of a result arising by chance is 2.5%, and an **upper action limit**, where the probability of a result arising by chance is 0.1%. The control chart for our example would therefore look like Figure 13.10:

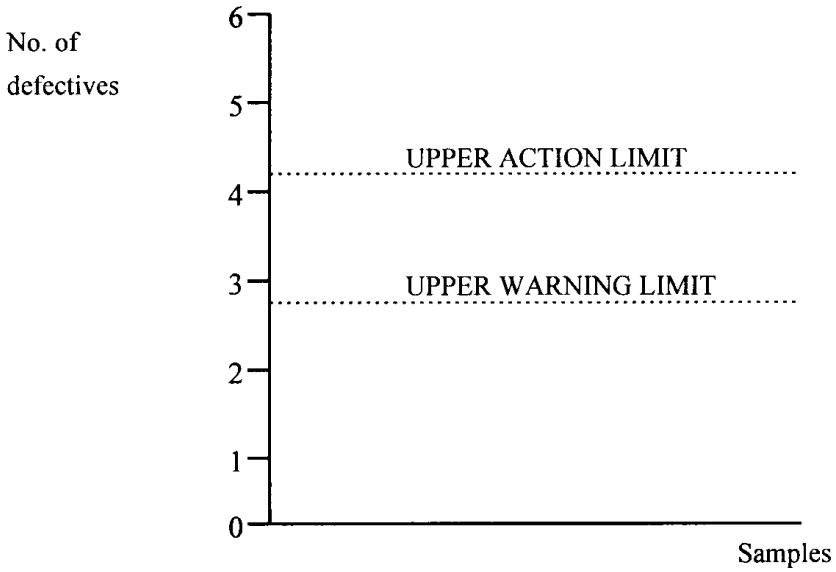


Figure 13.10

When samples are taken, the number of defectives is determined and points are entered on the control chart, as shown in Figure 13.11:

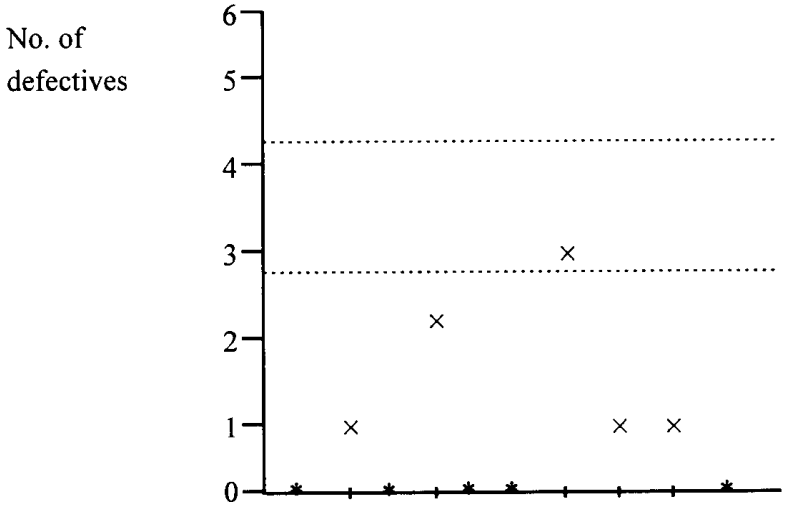


Figure 13.11

One point is above the upper warning limit but, unless more points fall above this line, no action would be taken – one point in every 40 would be expected to fall above the warning limit simply by chance.

Not only can upper action and warning limits be drawn on the chart, but also **lower** limits. At first this may seem strange, in that quality can hardly be too good. However, there are two reasons for identifying results which are better than expected. Firstly, it might be possible for someone to falsify results by making sure that samples chosen all contain good pieces. Secondly, if the process has actually been improved it is obviously important to determine exactly how it has happened, so that the change can be sustained.

ANSWERS TO QUESTIONS FOR PRACTICE

1. Let S represent a success, i.e. throwing a 6, and F represent a failure, i.e. not throwing a 6.

   Let $P(S) = p = \frac{1}{6}$

   $P(F) = q = (1-p) = \frac{5}{6}$

   The tree diagram is as shown in Figure 13.12:

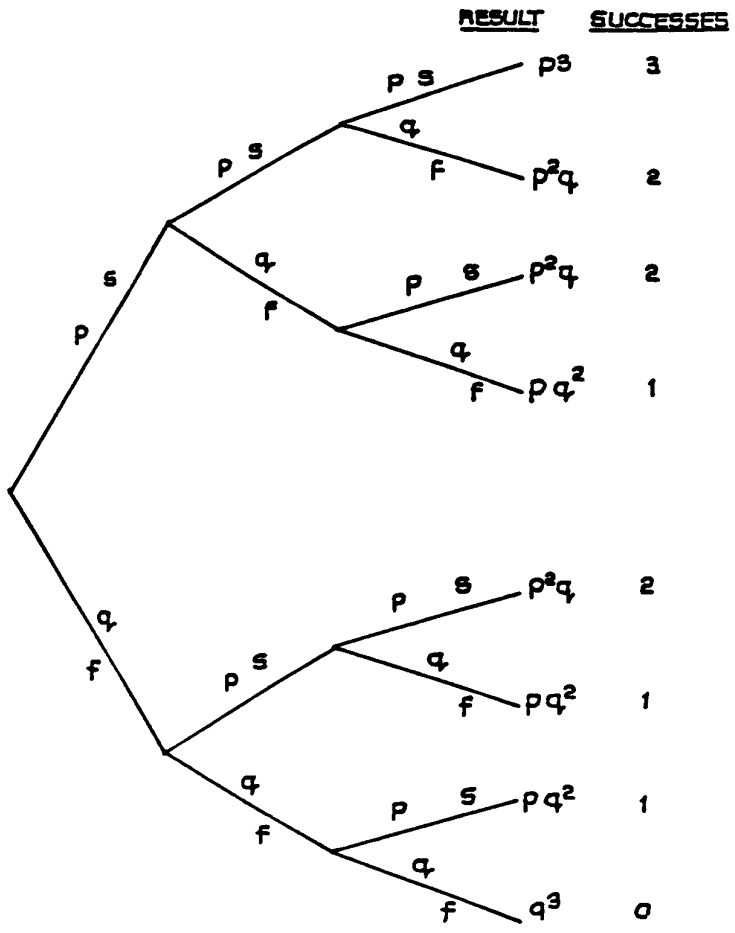


Figure 13.12

Summary of Results

$P(3 \text{ successes}) = 1 \times p^3 = \left[\frac{1}{6}\right]^3$

$P(2 \text{ successes}) = 3 \times p^2q = 3\left[\frac{1}{6}\right]^2\left[\frac{5}{6}\right]$

$P(1 \text{ success}) = 3 \times pq^2 = 3\left[\frac{1}{6}\right]\left[\frac{5}{6}\right]^2$

P(0 successes) $= 1 \times q^3 = \left[\frac{5}{6}\right]^3$

Combining these results gives:

$p3 + 3p^2q + 3pq^2 + q^3 = (p + q)^3$

or $\left[\frac{1}{6}\right]^3 + 3\left[\frac{1}{6}\right]^2\left[\frac{5}{6}\right] + 3\left[\frac{1}{6}\right]\left[\frac{5}{6}\right]^2 + \left[\frac{5}{6}\right]^3 = \left[\frac{1}{6} + \frac{5}{6}\right]^3$

which equals the binomial probability function

${}^3C_r \left[\frac{1}{6}\right]^r \left[\frac{5}{6}\right]^{3-r}$

2. Let p be the probability that an item can be satisfactorily reworked. Then:

$p = 0.6$, $q = (1 - p) = 0.4$, and $n = 5$

P(At least 4 can be reworked) = P(4 or 5 can be reworked)

$= P(4) + P(5)$

$= {}^5C_4(0.6)^4(0.4) + (0.6)^5$

$= 5(0.6)^4(0.4) + (0.6)^5$

$= 0.2592 + 0.07776 = 0.33696.$

3. Let p be the probability that an item is not defective. Then:

$p = 0.8$, $q = (1 - p) = 0.2$, and $n = 2$

P(Passing batch) = P(2 items tested are both good)

$= P(2) = (0.8)^2 = 0.64$

4. Let p be the probability that a line is engaged. Then:

$p = 0.25$, $q = (1 - p) = 0.75$, and $n = 6$

(a) P(At least 1 line is engaged) = 1 – P(0 lines are engaged)

$= 1 - P(0) = 1 - (0.75)^6$

$= 1 - 0.1780 = 0.8220$ to 4 sig. figs.

(b) P(All 6 lines are engaged) $= P(6) = (0.25)^6$

$= 0.0002441$ to 4 sig. figs.

5. Let p be the probability of having a boy. Then:

$p = ½$, $q = (1 - p) = ½$, and $n = 10$

(a) P(6 of the children will be boys) $= P(6) = {}^{10}C_6\left[\frac{1}{2}\right]^6\left[\frac{1}{2}\right]^4$

$= \frac{10!}{6!\ 4!}\left[\frac{1}{2}\right]^{10} = 210\left[\frac{1}{2}\right]^{10}$

$= 0.2051$ to 4 sig. figs

(b) P(None will be a girl) = P(10 will be boys) = P(10)

$$= \left[\frac{1}{2}\right]^{10} = 0.0009766 \text{ to 4 sig. figs}$$

(c) P(At most 2 will be boys) = P(0 or 1 or 2 will be boys)

$$= P(0) + P(1) + P(2)$$

$$= \left[\frac{1}{2}\right]^{10} + {}^{10}C_1 \left[\frac{1}{2}\right]\left[\frac{1}{2}\right]^{9} + {}^{10}C_2 \left[\frac{1}{2}\right]^{2} \left[\frac{1}{2}\right]^{8}$$

$$= \left[\frac{1}{2}\right]^{10} + 10 \left[\frac{1}{2}\right]^{10} + 45 \left[\frac{1}{2}\right]^{10}$$

$$= 56 \left[\frac{1}{2}\right]^{10} = 0.05469 \text{ to 4 sig. figs}$$

6. (a) (i) Probability – die before 60 $= \dfrac{1{,}000 - 748}{1{,}000} = 0.252$

(ii) Probability – die before 60, given already 30

$$= \frac{944 - 748}{944} = 0.208$$

(iii) Probability – die before 60, given already 50

$$= \frac{880 - 748}{880} = 0.150$$

(b) Expected value of pay-out = £10,000 × 0.15 = £1,500

Premium to break even = £1,500

(c) A loss will be made if pay-outs exceed premiums.

Premiums = 12 × £2,000 = £24,000

A loss is made if three or more die. Using binomial probability distribution:

$$P(r) = {}^{n}C_r p^r (1-p)^{n-r}$$

$$P(0) = (0.85)^{12} = 0.1422$$

$$P(1) = \frac{12!}{11!1!}(0.15)(0.85)^{11} = 0.3012$$

$$P(2) = \frac{12!}{10!\,2!}(0.15)(0.85)^{10} = 0.2924$$

Probability 3 or more die = $1 - P(0) - P(1) - P(2)$

$$= 1 - 0.1422 - 0.3012 - 0.2924$$

$$= 0.264$$

(d) The data relate to people born in the decades prior to 1986. There may be a trend in mortality rates such that people taking out policies now have different life expectancies from those who died in 1986. The statistics therefore need to be used with care.

(e) (i) Assume that those who die by the age of 52, die on their 51st birthday. Those who die by age 54 die on their 53rd birthday, and so on.

Age × Probability of Death
$51 \times \frac{14}{132} = 5.409$
$53 \times \frac{20}{132} = 8.030$
$55 \times \frac{24}{132} = 10.000$
$57 \times \frac{34}{132} = 14.682$
$59 \times \frac{40}{132} = 17.879$

Expected age of death (for someone who dies between ages 50 and 60)	= 56.000

(ii) Average discounted pay-out: $= \frac{10{,}000}{1.08^6} = £6{,}302.$

Premium for break-even = £100 + £200 + (0.15 × £6,302) = £1,245.30.

Study Unit 14

Decision Making

Contents *Page*

A. DECISION MAKING AND INFORMATION

The one activity which distinguishes management from the other functions in an organisation is decision making. We can study decision making from many aspects; in many circumstances decisions are taken for reasons which can only be studied by the use of psychology or sociology, but we can also consider what is called “rational decision making” and this makes use of quantitative measures to a large degree.

The purpose of quantitative methods of presenting and analysing data is to provide the decision maker with a basis on which to make his or her decisions. This of course tends to make decision making more difficult. The reason is that, in the absence of good information as to the possible consequences of a decision, decision making becomes a matter of choosing between those alternative courses of action which are before the decision maker. A person who makes decisions in a confident way may gain a good reputation as a decision maker; most people tend to prefer someone who is resolute and confident to someone who hesitates and tries to weigh up the alternatives. Having good leadership qualities in this manner does not, unfortunately, mean that the leader always makes correct decisions. If there is an extended interval between taking a decision and finding out what its ultimate consequences are, then the “leader” may escape censure for a bad decision. It may take someone studying the history of the events to make a clear judgement, by which time the “leader” will have moved on to other things!

When good quantitative information is available and presented in a meaningful way, it becomes a great deal easier to see what are the consequences of different courses of action. They can then be compared and a decision taken on the judgement as to which course of action is likely to have the best outcome. The more that is known about the possible outcomes, the more we have to try to bring the measures of success to a common basis, typically money. Attempts to do so when we are considering what is best for a community rather than just a commercial organisation have led to the “cost benefit analysis”.

It is convenient to classify decision situations under three headings: decisions under certainty; decisions under risk and decisions under uncertainty. We will look first at decisions under certainty.

B. DECISION MAKING UNDER CERTAINTY

Here the outcome to a choice of course of action (a strategy) can be evaluated and the same result will always occur for that choice of strategy. As the outcomes are determined, we call this a deterministic situation. When all possible strategies can be listed and their outcomes evaluated, the rational decision maker simply chooses the strategy which gives the best value to the outcome. The fact that there are many real-life situations in which people could take decisions in this manner but do not, either because they do not assess all strategies, or simply by perversity, shows the need to undertake behavioural studies.

There are three problems for the decision maker and his or her quantitative advisors:

- To find an appropriate measure by which to evaluate outcomes;
- To calculate the value of the measure for an outcome;
- To identify the best outcome from a large number of outcomes.

Outcome Measures

In most organisational situations there are two popular measures. "How much?" and "How long?". As most commercial decisions resolve to a matter of money or "the bottom line", the first of these is the dominant measure. In other organisations other measures may apply.

As the concept of organisational decision making originated in wartime, we can look at the choices facing a military commander when deciding on what to do next. One such problem concerned the supply of war materials from the USA to the UK and Europe across the Atlantic Ocean during World War II. The possible measures were:

- To maximise the total tonnage of goods delivered;
- To minimise the time taken to deliver goods;
- To minimise the loss of shipping.

They are conflicting objectives to some degree, and the "best" answer to the problem is not easy to see. In the event the answers adopted were related to the shipping available. The very fast ships, such as the liner Queen Mary, were used to minimise time, relying on their great speed to evade attack. Most other ships were operated in convoys, to minimise shipping losses, and to satisfy to some degree the total tonnage requirement.

Valuation of Measures

This is where the various mathematical techniques available come into their own. The characteristic of such methods is that there is a "right" answer to the question posed.

Optimal Outcome

Some situations involve a very large number of choices. Even to list them may be a daunting task, and to evaluate them all may be lengthy and costly. Fortunately some excellent methods and algorithms have been developed, which can ensure that we consider all alternatives and identify the "best" outcome. Critical path analysis is of this nature, providing a method of systemising our knowledge so that the effect of choices of order of doing things can be seen. Another very powerful method looks at the assignment of resources between a number of competing uses. This is "linear programming". We do not study linear programming (LP) in this course but you should know that it exists. It is used in areas such as:

- Investment decisions;
- Product mix situations;
- Product formulation;
- Production scheduling;
- Delivery scheduling.

C. DECISION MAKING UNDER RISK

In this situation the value of an outcome cannot be expressed as a single figure, but will take the form of a probability statement. Effectively we cannot say that a particular strategy is the "best", but only that it is "probably the best". Remembering what we said about the meaning of probability earlier, we are deriving measures which are correct "on average" or when taken over a large number of similar situations. We study these either by the use of statistical distributions or by "expectations",

which are average measures of outcome. Many decision situations fall into this category; the main characteristic is that we can measure our success on the basis of a long-term evaluation. As with a fairground operator, we look at the swings and roundabouts. It does not matter if one day swings are popular and roundabouts ignored, and another day the opposite applies, as long as the average revenue taken over both swings and roundabouts is good. When a decision can only be taken once, we may have to adopt a different approach, which is decision taking under uncertainty.

D. EXPECTATIONS

If we have a strategy which can result in a number of outcomes, and if we know the probability of each of the outcomes occurring, then we can calculate the expectation for the strategy:

$$\text{Expectation} = \Sigma(\text{Probability of outcome} \times \text{Value of outcome})$$

A good example of this method of decision making is given by the "newsboy problem". This is typical of several similar problems. A newsboy has a stand and can order so many of each newspaper or magazine each day. Suppose he knows the probability of any level of sales, and that he suffers a loss on unsold copies of the publication he is considering, say a specialist magazine.

We will work with the following data:

A magazine is issued monthly and remains on sale for one month only. Magazines are supplied in bundles of ten.

A magazine costs £1.50 to buy, sells for £2.50, unsold copies are taken back for £0.50. There is no penalty for failing to have a magazine available.

For simplicity we take sales rounded to tens.

Demand	Probability of this Level of Demand
10	0.1
20	0.15
30	0.25
40	0.20
50	0.20
60	0.10

First we calculate the value of outcomes for each choice of strategy and each possible outcome (in £):

Outcome Table

Strategy = Order		**10**	**20**	**30**	**40**	**50**	**60**
Demand:	10	10	0	–10	–20	–30	–40
	20	10	20	10	0	–10	–20
	30	10	20	30	20	10	0
	40	10	20	30	40	30	20
	50	10	20	30	40	50	40
	60	10	20	30	40	50	60

We now multiply the figures in each row by the probability that level of demand will result in, which gives a table of expectations. We then total the figures in each column to find the total expectation for each strategy:

Expectation Table

Strategy = Order		**10**	**20**	**30**	**40**	**50**	**60**
Demand	**Probability**						
10	0.1	1	0	–1	–2	–3	–4
20	0.15	1.5	3	1.5	0	–1.5	–3
30	0.25	2.5	5	7.5	5	2.5	0
40	0.2	2	4	6	8	6	4
50	0.2	2	4	6	8	10	8
60	0.1	1	2	3	4	5	6
Totals		10	18	23	23	19	11

We can see that the optimal choice is to buy either 30 or 40 copies of the magazine. A decision between them could now be made on the basis of confidence in increasing sales. If there seems to be any chance of sales being higher than described, then we would go for the higher order, 40 copies.

Expected Value of Perfect Information (EVPI)

If we knew the demand for each day in advance, we could order the exact amount required each day. The average profit would then be the sum of the profit for meeting the exact demand each month × probability of that demand, i.e.

$$(0.1 \times 10) + (0.15 \times 20) + (0.25 \times 30) + (0.2 \times 40) + (0.2 \times 50) + (0.1 \times 60)$$

$$= 1 + 3 + 7.5 + 8 + 10 + 6 = 35.5$$

The difference between this value and the average profit for the optimal order level is:

$$35.5 - 23 = 12.5$$

It would be worth paying up to this amount for the information, and this figure is called the **EVPI** or "**Expected Value of Perfect Information**".

Expectations and averages

An expectation is a weighted mean of the outcomes, using probabilities to provide the weights. We can identify the arithmetic mean as the expectation of the value of a variable in a distribution. It is quite evident when we look at the mean calculated from a grouped table. The midpoints of the classes are weighted by class frequencies, and the sum divided by the total frequency. This is equivalent to finding the sum of the midpoints multiplied by the probability that a measurement is in that class, using class frequency/total frequency as a measure of probability that a measurement is in a class. So we can refer to the mean as the expectation of the value of a variable.

E. COMPLEX DECISIONS: DECISION TREES

Some decisions are taken as part of a set, and if there is a mix of decisions and ranges of outcomes, it can become difficult to see what is happening. The tendering problem is of this type.

As an example, consider the case of a small builder who is invited to tender for a contract. He knows that there is a second tender coming up, more valuable than the first, but that unless he puts in a tender for the first he will not be invited to tender for the second. If he wins the first contract he must complete it and cannot bid for the second. He knows that his costs on the contracts will depend on the weather; also that his chance of winning the contract will depend on his bid.

He makes estimates as follows:

Contract A

Bid level high	0.2 chance of winning the contract
Bid level medium	0.6 chance of winning the contract
Bid level low	0.9 chance of winning the contract
Bid level high	Profit: £60,000
Bid level medium	Profit: £40,000
Bid level low	Profit: £25,000
Effect of weather:	
Very bad weather	Reduce profits by £15,000
Poor weather	Reduce profits by £5,000
Probability of very bad weather	= 0.2
Poor weather	= 0.4

Contract B

Bid level high	0.3 chance of winning the contract
Bid level medium	0.7 chance of winning the contract
Bid level low	0.9 chance of winning the contract
Bid level high	Profit: £100,000
Bid level medium	Profit: £80,000
Bid level low	Profit: £60,000
Effect of weather:	
Very bad weather	Reduce profits by £20,000
Poor weather	Reduce profits by £10,000
Probability of very bad weather	= 0.2
Poor weather	= 0.4

It costs £5 000 to prepare a bid, which has been allowed for in all the profit figures. What should he do?

We can draw a decision tree to review the choices and possible outcomes. A decision tree comprises branches coming out of nodes. The nodes denote points at which we either make a choice or we face several possible outcomes, and are known as "decision nodes" or "chance nodes" respectively. They are usually drawn as squares and circles, as shown by the partial tree in Figure 14.1.

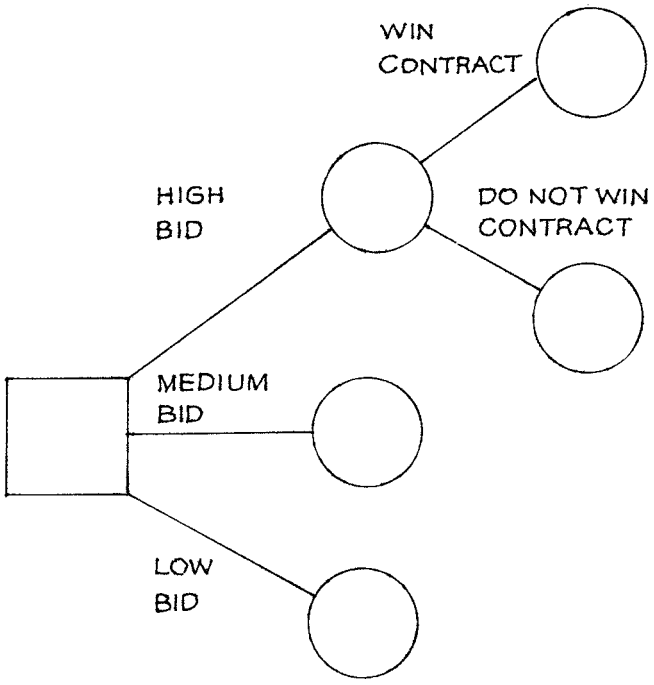


Figure 14.1: Section of a Tree only

Now try to draw the complete tree for yourself on the basis of the information given in the text. You can then check your version against Figure 14.2

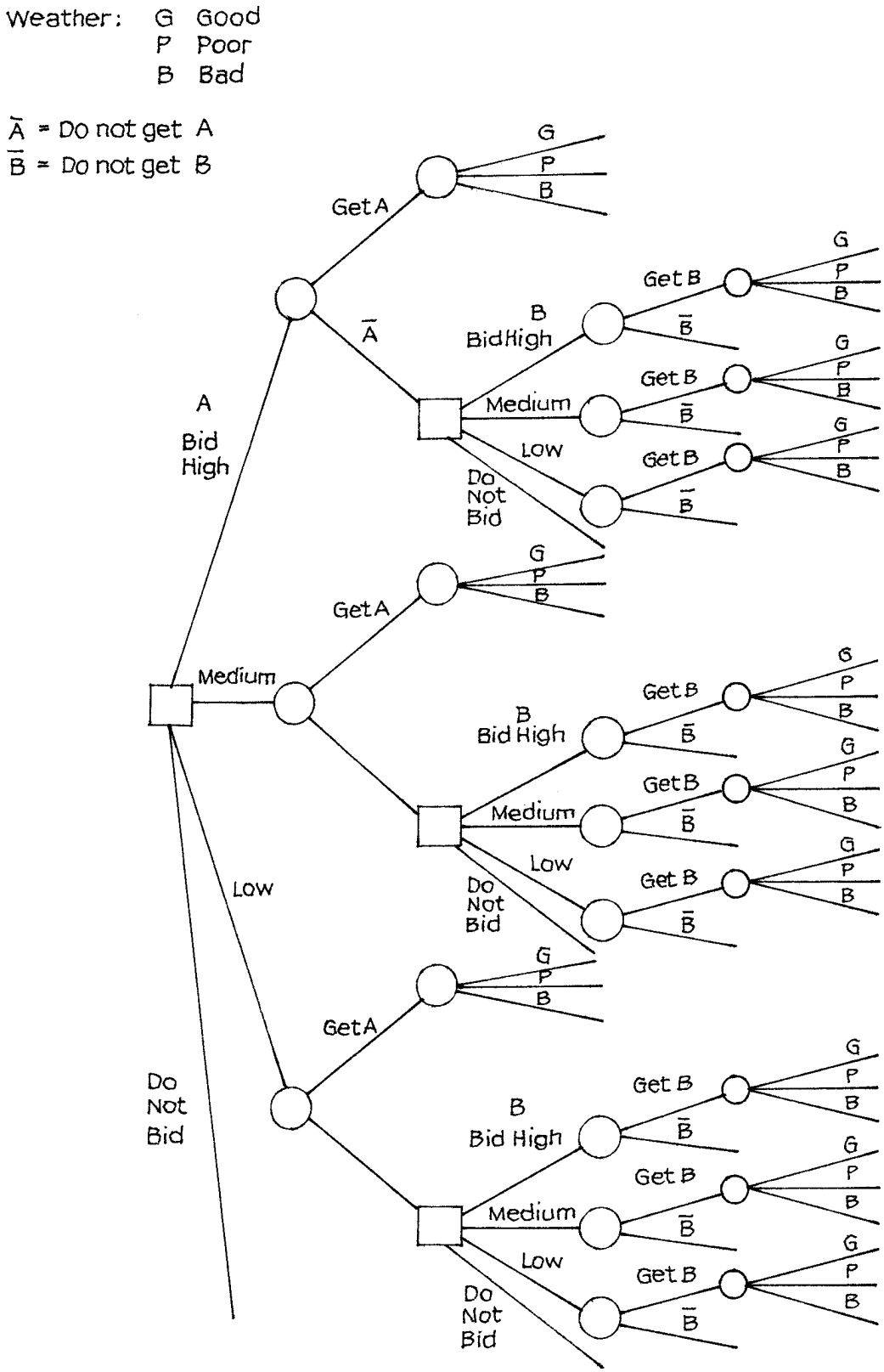


Figure 14.2: Completed Tree

For each “twig” on the extreme right of the tree in Figure 14.2 we can calculate the value of the outcome. For example, take the sequence:

- Bid high for first contract; do not get first contract
- Bid high for second contract; get second contract
- Experience poor weather.

The values are:

- –£5 000 Cost of first bid
- £100 000 Profit with high bid
- –£10 000 Reduction for poor weather.

Hence net profit is £85 000.

Now work out the values for all the other twigs in Figure 14.2, before looking at the tree in Figure 14.3.

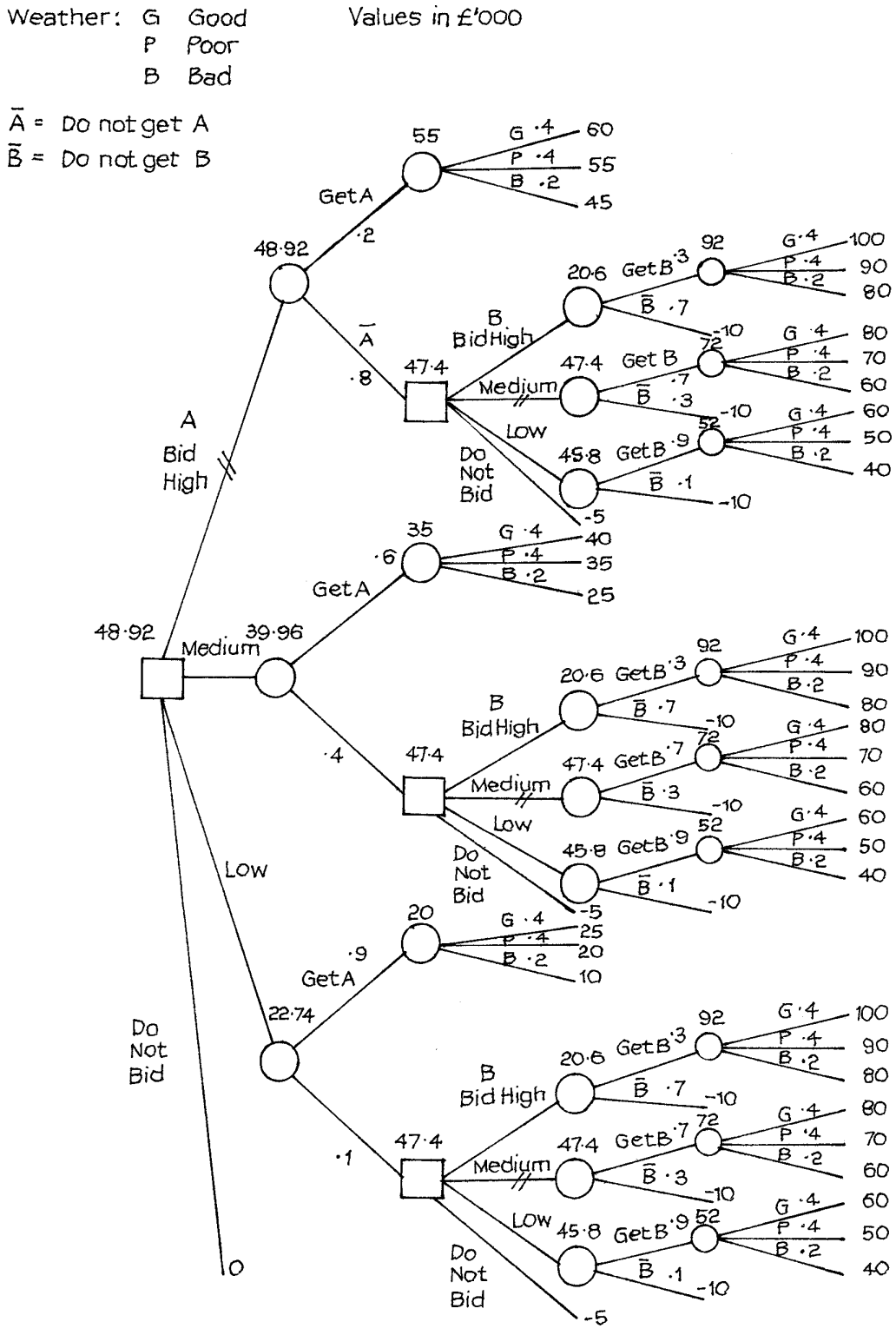


Figure 14.3

We can now work through the tree from right to left and, at each chance node, find the expectation for that node and, for each decision node, choose the strategy which gives the highest expectation.

Check the tree so that you can see where all the figures come from. Note that decisions are shown by putting a cross stroke on the branches chosen.

We find that the best decision is to make a high bid for the first contract; if this is not successful, make a medium bid for the second contract. The expectation of profit with this choice of strategies is

£48 920.

You may notice that the second part of the tree, concerning the second contract, repeated itself three times. We could have evaluated this as a separate decision tree and entered the result, which was medium bid each time, in a simpler tree for the first bid. Sometimes successive decisions interact; the profit on the second contract might have been linked to what happened in the first, in which case the tree must be analysed as a unit.

Interpreting Expectations

It is important to realise what an expectation, such as the answer to the last problem, actually means.

When the builder has completed a contract, his profit will be one of the figures at the ends of the twigs in the tree. The exact figure will depend on his decisions, which contract he gets and the weather. The figure of £48 920 we found is the average profit he would make if he were able to repeat the contracts many times, following the rule we derived. Of course in real-life this is not possible, so why bother? The answer is that if a company is in a situation where similar decisions have to be taken frequently, then it is reasonable to take an average view. We can afford to lose money on some contracts as long as on average, taken over many contracts, we make a profit. The analysis is then valid. In other words, for a large construction company which bids for many contracts the analysis is suitable. Let us return to our small builder. Suppose he is short of cash and worried that if he does not get a contract he will be unable even to pay for preparing the bid.

F. DECISION MAKING UNDER UNCERTAINTY

This scenario covers two situations. Firstly, when we cannot afford to take an average view, as with our builder; secondly, when nothing is known about the relative probabilities of the outcomes.

Using Guesstimates

It is often possible to make some sort of estimate of probabilities. Such expectations can be used , but we would probably carry out a sensitivity analysis to see how dependent the final answers are to changes in the probabilities utilised.

Principle of Insufficient Reason

A special case is where we can list the possible outcomes and their values for a strategy, but know nothing of their relative chances of occurring. If we know no reason why the chances of the outcomes occurring should be different, then we take them as all equal. We can then calculate expectations as before. We have already met a case. When we take a random sample in order to calculate a mean, we are deliberately trying to achieve a situation where all outcomes (the choices of item in the sample) are equally likely. A good test of random selection is to try to see if there is any possible reason why the chances of choosing a particular measurement should be different from the chance for other measurements. Note that if a human being is involved, chances are not equal, and therefore any selection process which depends on a personal choice cannot be random.

Criterion of Pessimism

As an alternative, consider the case when we have four strategies facing us, and we can identify three possible scenarios which will give a different outcome for each strategy. They are sometimes called the "States of Nature".

We will invent an example relating to the pricing of a product where we are concerned with the effect of competitors' actions. We can display the data in a table as follows, with the figures representing projected profits in £0,000s:

		State of Nature = Competitors' Action		
		Decrease by 10%	**Do Nothing**	**Increase by 10%**
Strategy:	Do nothing	−10	0	+20
	Increase by 10%	−50	+10	+5
Our Action:	Reduce by 10%	−5	+5	+40
	Reduce by 20%	0	+10	+30

The criterion of pessimism tells us to identify the worst outcome for each strategy, and then to choose the least worst. Thus we minimise our risk of loss, but without ensuring that we make the most profit.

The figures are:

Do nothing	−10
Increase by 10%	−50
Reduce by 10%	−5
Reduce by 20%	0

The least of the worst cases is 0 for "reduce by 20%".

This criterion is in fact widely used, although the users may not realise it. There are many people who have to take decisions who feel that they will be blamed if things go wrong, but will receive no praise if things go right. Anyone in this position will tend to play safe and allow chances for profit or other to gain go by if they involve any risk of loss for which the person could be blamed. A "risk averse" attitude is widespread. When applied to matters of safety it is probably correct. As an example, there is a small risk that some children who are vaccinated against whooping cough and other childhood diseases will react badly to the vaccine and suffer brain damage. The official attitude is that the benefits of mass vaccination to the population as a whole outweigh the risk. Many parents and some doctors argue that the tragic effect on a family when things go wrong must be regarded as too costly to balance the benefits for others. Here a risk averse attitude would seem to be very sensible to many people.

In contrast we might consider a manager making decisions on loan applications for a finance house. If the manager is too cautious in assessing the risk that a borrower will default on the loan, new business will be turned away. The profitable nature of most finance arrangements is such that a few bad debts can be balanced against the many which do not go wrong. A manager who makes only a few mistakes in granting loans is probably performing better and ensuring the finance house does not miss valuable business.

We have returned to expectations in this case, as the finance houses have developed sophisticated "scoring" systems which take into account a person's situation and give a rating which represents the chance that that person will default on a loan. The higher the interest rates charged on a loan, the greater risk of default the finance house is prepared to take. A simple rule used by some lenders is "If a person already has several credit cards and loans, grant the loan", based upon the premise that someone who is used to juggling with interest payments and loan repayments is less likely to get into trouble than someone who has not borrowed before, or who has had credit cards stopped or refused.

G. BAYESIAN ANALYSIS

There are two ways to consider probabilities. In the first, or classical way, we estimate probabilities either from theoretical grounds or from experience, and the probabilities remain constant through our analysis. The other, Bayesian way, is to review our estimates in the light of increasing knowledge.

Suppose we are throwing some dice. By classical probability we can calculate the chance of particular throws, say throwing two sixes, on the assumption that all faces of the dice are equally likely to end up on top. What if we find after a number of throws that there appear to be more sixes coming than predicted? If you have tried some of the exercises involving dice you will know that it can take some time before the values you get begin to settle down; the classical approach would say that we have to wait for many throws before we can test if our assumptions are true. However, the Bayesian analyst will modify the estimates of the probabilities continually, after every throw.

There is a lively debate in statistical circles as to precisely when it is best to adopt a classical or a Bayesian stance, as probability theory underpins the whole of statistics. Fortunately for many users who would find it difficult to follow, many statistical techniques are reliable regardless of the approach adopted. Bayesian methods tend to lead into complex mathematics, which require considerable computing power.

To make things difficult, some people use the term "Bayesian" to describe any methods which require the estimation of probabilities. Here, we shall give some idea of the real meaning of the term, but be prepared to meet the wider and vague use.

We will now look at a simple example of how Bayesian analysis can contribute to decision making by using "Bayes' Theorem" to modify probabilities.

Bayes' Theorem

Our application is drawn from test marketing, a common procedure when a company wants to see if a new or modified product, or a marketing campaign, is likely to be successful. We will assume that for the type of product the company has experience of previous test marketing, and has collected the data for cases when the product was placed on the market following a test marketing. It is not strange to find that some products are marketed despite a poor test marketing; marketeers do play hunches and take chances.

- Test market result rated "Very good": 15 cases.

  Subsequent sales assessed as: Excellent 4

  Good 10

  Poor 1

- Test market result rated "Good": 25 cases.

  Subsequent sales assessed as: Excellent 2
  Good 18
  Poor 5

- Test market result rated "Poor": 10 cases.

  Subsequent sales assessed as: Excellent 1
  Good 4
  Poor 5

We now derive some conditional probabilities, writing them in the form P($A1|B2$) which, as you will remember, would be read as the probability of observing outcome $A1$ given that event $B2$ has happened. We will let the outcomes to the test marketing be:

Very good test $B1$
Good test $B2$
Poor test $B3$

and to the actual marketing:

Excellent sales $A1$
Good sales $A2$
Poor sales $A3$

Regardless of testing, the probabilities of the selling results are estimated as:

$P(A1) = 0.2, \quad P(A2) = 0.6, \quad P(A3) = 0.2$

How should they be modified if we carry out a test marketing and find a "poor" result?

The answer is based on work done by the Rev. Thomas Bayes, the importance of whose work was not appreciated until many years after his death. Fortunately, a friend had read a paper by Bayes before the Royal Society, the implications of which were eventually recognised. He is now posthumously famous.

Let us collect our information together:

$P(A1) = 0.2, \quad P(A2) = 0.6, \quad P(A3) = 0.2$

From the tests:

$P(B1|A1) = 4/7 \quad P(B1|A2) = 10/32 \quad P(B1|A3) = 1/11$

$P(B2|A1) = 2/7 \quad P(B2|A2) = 18/32 \quad P(B2|A3) = 5/11$

$P(B3|A1) = 1/7 \quad P(B3|A2) = 4/32 \quad P(B3|A3) = 5/11$

The theorem states:

$$P(A1|B3) = \frac{P(A1) \times P(B3|A1)}{\sum\{[P(A1) \times P(B3|A1)] + [P(A2) \times P(B3|A2)] + [P(A3) \times P(B3|A3)]\}}$$

Putting figures in we get:

$$P(A1|B3) = \frac{(0.2 \times 1/7)}{(0.2 \times 1/7) + (0.6 \times 4/32) + (0.2 \times 5/11)}$$

$$= \frac{0.0286}{0.0286 + 0.075 + 0.0909} = 0.147$$

By following the same pattern we find:

$P(A2|B3) = 0.386$ and

$P(A3|B3) = 0.467$

Repeat the same calculation for the case when a very good test is observed,.

The probabilities we started with are called *a priori* probabilities; the ones we have just calculated are *a posteriori* probabilities.

Study Unit 15

Significance Testing

Contents ***Page***

(Continued over)

A. INTRODUCTION

We have discussed the practical details of collecting data and selecting samples for statistical surveys, and the theoretical concepts of elementary probability and theoretical probability distributions that you must know before you can carry out the process of statistical inference, i.e. before you can infer information about a whole population from the data collected in a sample survey. The normal distribution is, as far as you are concerned, the most important distribution.

In this study unit we shall be using information about population means and proportions from the means and proportions calculated from sample data. The inferences about these two measures are all that you require for this course, but similar methods are used in more advanced work for inferring the values of other population measures. We are concerned with two processes in this study unit:

(a) The estimation of population parameters from sample statistics and the confidence which can be attached to these estimates.

(b) Tests to decide whether a set of sample data belongs to a given population. This process is known as hypothesis testing.

Before explaining these processes, we need to state some necessary assumptions and definitions and to clarify some notation.

Assumptions

(a) The sample analysed is **large.**

(b) The sample is a **random sample**.

Definitions

(a) In this type of analysis, a sample which contains **more than 30 items** is counted as large. Different techniques must be used in the analysis of small samples, as we consider later in the unit. Because of the time factor in examinations, you will sometimes be given questions involving smaller samples and, in such cases, either you will be told that the sample is taken from a normal population or you should state, as an extra assumption, that the population is normal.

(b) A random sample is a sample that has been selected so that any other sample of the same size was **equally likely** to have been selected, or a sample that has been selected so that each individual item is equally likely to have been chosen. (These two definitions are equivalent.)

Note that we always assume that we are analysing a random sample, although in practice, because of the nature of the population or a restriction on the cost or time allowed for the survey, some other method of selection has been used. If, in an exam, you are asked to comment on the result of your analysis, you should state that a bias may have been introduced by the method of selection.

Notation

In all work on inference we will keep strictly to the notation listed below. If you are not careful in the use of notation, you may find that you have calculated two values for the same symbol. The general rule to remember is that Greek letters are used for populations, and Roman letters for samples. These symbols are used universally, so that they need not be defined in each problem.

	Size	Mean	Standard Deviation
Population	N	μ	σ
Sample	n	$\bar{x}$	s

You will notice that the rule is broken here with the population size, but the Greek equivalent of N is very easily confused with V. There is another common exception that we will use, i.e. because the Greek equivalent of p is π and we use this symbol for the ratio of the circumference of a circle to its diameter, we use p for the population proportion and $\hat{p}$ (read as p-hat) for the sample proportion.

In addition we need to define a special distribution and state an important theorem. These two items are so important that we will devote the whole of the next section to them.

B. THE SAMPLING DISTRIBUTION AND THE CENTRAL LIMIT THEOREM

The Sampling Distribution

Suppose you own a factory producing fluorescent light tubes. It would be very useful, both for advertising and for monitoring the efficiency of your plant, to know the average length of life of these tubes. Clearly it is not possible to test every tube, so you take a random sample of size n, measure x hours, the life length of each tube, and then calculate $\bar{x}$ and s. These values will be estimates of the population parameters μ and σ.

If you take a number of samples, each one will give you a different value for the statistics and so a different estimate. x is a variable so it will have a probability distribution and, since $\bar{x}$ is a statistic calculated from the sample, it is also a variable and so will have a probability distribution. The distribution of $\bar{x}$ is called the **sampling distribution of the mean** and it has the following properties:

(a) It can be shown that the mean of $\bar{x}$ is equal to the mean of x, i.e. $\mu_{\bar{X}} = \mu$

(b) It is obvious that the standard deviation of $\bar{x}$ will be smaller than the standard deviation of x because the extreme values of $\bar{x}$ must be smaller than the extreme values of x. The standard deviation, $\sigma_{\bar{X}}$, of $\bar{x}$ depends upon the size of the sample and is defined as:

$\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$. This is called the **standard error of the mean** (SE).

Of course, s will also have a sampling distribution, but that is not included in this course. The standard distribution is less variable than the mean, so its standard deviation will be smaller. In practice, since we rarely know the true value of σ, we can use the following formula for the standard error without significant loss of accuracy:

$$\frac{s}{\sqrt{n}}$$

Central Limit Theorem

The Central Limit Theorem forms the basis of the theory that we use for the statistical inference processes in this study unit. It states that:

(a) If the distribution of x is normal with mean μ and standard deviation σ and samples of size n are taken from this distribution, then the distribution of $\bar{x}$ is also normal with mean μ and standard deviation (SE)

$$\frac{\sigma}{\sqrt{n}}$$ whatever the size of n.

(b) If the distribution of x is not normal, as the size of the sample increases, the distribution of x approaches the normal distribution with mean μ and standard deviation

$$\frac{\sigma}{\sqrt{n}}$$

i.e. for values of $n > 30$ we can assume that $\bar{x}$ is normal.

Application of Central Limit Theorem

The Central Limit Theorem means that if we have a large sample from a population, the distribution of which we do not know, we can use the standard normal distribution to calculate the probability of occurrence of values of $\bar{x}$ since the standardised value of x, which is

$z = \dfrac{\bar{x} - \mu}{\left(\dfrac{\sigma}{\sqrt{n}}\right)}$, will have the standard normal distribution.

In particular, we can find the length of the interval within which a given percentage of the values of $\bar{x}$ will lie. We are most interested in those intervals for which $\bar{x}$ is the centre, and these intervals are equivalent to the standardised intervals with 0 as the centre.

Example 1

The statement that about 95% of the values of $\bar{x}$ lie within two standard deviations of the mean implies that exactly 95% of the values will lie within about two standard deviations of the mean. Let z_1 be the exact value; then we can calculate z_1 from the equivalent probability statement, i.e.:

$$P(-z_1 < z < z_1) = 0.95$$

$$2P(z < z_1) = 0.95 \text{ (using the symmetry of the curve)}$$

$$P(z < z_1) = \frac{1}{2} \times 0.95 = 0.475$$

Then, from the standard normal table, $z_1 = 1.96$.

Example 2

Suppose we need the value of z_1 so that exactly 99% of the values will lie in the interval. Then as in Example 1:

$$P(-z_1 < z < z_1) = 0.99$$

$$P(z < z_1) = \frac{1}{2} \times 0.99 = 0.495$$

From the table, $z_1 = 2.57 + \frac{1}{2} \times 0.01 = 2.576$ to 3 dp (though 2.58 is usually accurate enough)

In the same way you can calculate the z value for any percentage or probability that you choose.

C. CONFIDENCE INTERVALS

Means

Usually, because of the time or cost of sample surveys, we have to base decisions on the analysis of one set of sample data. Having calculated the mean and standard deviation of this sample, we say that the estimated value of the population mean, μ, is the sample mean, $\bar{x}$. This information will be much more valuable if we can also estimate the difference between $\bar{x}$ and the true value of μ. We do this by finding the interval, with centre the calculated value of $\bar{x}$, which will contain a chosen percentage of the other possible values of $\bar{x}$. This is called a confidence interval for μ, and its boundaries are called the **confidence limits**.

Let z_1 be the value of the standard normal corresponding to the chosen percentage; then the confidence interval is defined by:

$$-z_1 < z < z_1$$

$$-z_1 < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_1 \text{ since } z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$-z_1 \frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < z_1 \frac{\sigma}{\sqrt{n}} \text{ multiplying the inequality by the positive } \frac{\sigma}{\sqrt{n}}$$

$$-\bar{x} - z_1 \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{x} + z_1 \frac{\sigma}{\sqrt{n}} \text{ subtracting } \bar{x}$$

$$\bar{x} + z_1 \frac{\sigma}{\sqrt{n}} > \mu > \bar{x} - z_1 \frac{\sigma}{\sqrt{n}} \text{ dividing by } -1 \text{ and so reversing the inequality sign}$$

i.e. $$\bar{x} - z_1 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_1 \frac{\sigma}{\sqrt{n}}$$

Thus, we say that the confidence interval for $\mu = \bar{x} \pm z_1 \frac{\sigma}{\sqrt{n}}$.

The working given above need not be included in the solution of problems, but you must learn the final formula so that you can quote it.

Example

A sample of 100 fluorescent tubes gives a mean length of life of 20.5 hours with a standard deviation of 1.6 hours. Find (a) a 95% confidence interval; (b) a 99% confidence interval for the average length of life of those tubes. Interpret the result in each case.

(a) $\bar{x} = 20.5$, $n = 100$, $s = 1.6$ $(= \sigma)$, $z_1 = 1.96$

A 95% confidence interval for $\mu = \bar{x} + z_1 \frac{\sigma}{\sqrt{n}}$

$$= 20.5 \pm 1.96 \times \frac{1.6}{10}$$

$$= 20.5 \pm 0.3136$$

$$= 20.19 \text{ to } 20.81$$

(Note that as $\bar{x}$ is given to only 1 dp the limits should be given to 2 dp.)

This means that for 95% of the possible samples that could be taken, the estimate you would give for μ would lie in the interval 20.19 to 20.81 hours, i.e. you are 95% confident that μ lies between these values.

(b) $\bar{x} = 20.5$, $n = 100$, $s = 1.6$, $z_1 = 2.58$

A 99% confidence interval for $\mu = \bar{x} \pm z_1 \frac{\sigma}{\sqrt{n}}$

$$= 20.5 \pm 2.58 \times \frac{1.6}{10}$$

$$= 20.5 \pm 0.4128$$

$$= 20.09 \text{ to } 20.91$$

This means that you are 99% confident that the true value of μ lies between 20.09 and 20.91 hours.

Proportions

Suppose we need to find the proportion of individuals in a population who possess a certain attribute. For instance, for planning purposes we may require to know:

- The proportion of defective items coming off a production line in a shift.
- The proportion of pensioners in a country.
- The proportion of voters who are in favour of reintroducing the death penalty.
- The proportion of householders in a major city who wish to possess cable television.

Provided the proportion is not expected to be very small, we can use the same technique to find this information as we used for measurements of continuous variables.

The results of a sample survey are used to estimate the population proportion. For the population:

Let $p = \frac{\text{Number of individuals possessing the attribute}}{N}$

$q = \frac{\text{Number of individuals not possessing the attribute}}{N}$

and $p + q = 1$

For the sample:

$$\hat{p} = \frac{\text{Number of individuals possessing the attribute}}{n}$$

$$\hat{q} = \frac{\text{Number of individuals not possessing the attribute}}{n}$$

and $\hat{p} + \hat{q} = 1$

Then $\hat{p}$ is the estimate of p and, by the Central Limit Theorem, $\hat{p}$ is normally distributed, so:

- A confidence interval for $p = \hat{p} \pm z_1(\text{SE of } \hat{p})$ where z_1 is the value of the standard normal of the chosen percentage and
- The standard error of $\hat{p}$ can be shown to be $\sqrt{\frac{pq}{n}}$ which is estimated by $\sqrt{\frac{\hat{p}\hat{q}}{n}}$.

Example

In a sample of 200 voters, 80 were in favour of reintroducing the death penalty. Find a 95% confidence interval for the proportion of all voters who are in favour of this measure.

$$\hat{p} = \frac{80}{200} = 0.4, \quad \hat{q} = 1 - \hat{p} = 0.6, \quad \text{SE of } \hat{p} = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{0.4 \times 0.6}{200}}$$

Thus, a 95% confidence interval for $p = \hat{p} \pm 1.96 \sqrt{\frac{\hat{p}\hat{q}}{n}}$

$$= 0.4 \pm 1.96 \times 0.0346$$

$$= 0.4 \pm 0.0679$$

$$= 0.332 \text{ to } 0.468$$

Therefore, the proportion of voters in favour lies between 33% and 47%.

D. HYPOTHESIS TESTS

In order to discover whether or not a certain set of data comes from a particular population, we choose one of the descriptive measures of that population and then test to see if it is likely that the sample belongs to that population. To carry out the test we need to formulate two hypotheses:

(a) The null hypothesis, H_o: the sample belongs to the population.

(b) The alternative hypothesis, H_1: the sample does not belong to the population.

The principles behind hypothesis testing are the same whichever measure we choose for the test.

Because the sample measures are variables, we run a risk of making the wrong decision in two ways. We may reject H_o when it is true, making what is called a type I error, or we may accept H_o when it is false, making what is called a type II error.

Table 15.1 gives a summary of the possible decisions.

Table 15.1: Types of Error

	Accept H_o	*Reject H_o*
H_o true	Correct decision	Make type I error
H_o false	Make type II error	Correct decision

There is no direct algebraic connection between the values of the risks of making the two errors, but as one risk is decreased the other is increased. We use the type I error in hypothesis tests and the type II error to find the power of the test, which is outside the scope of this course.

Having formulated the hypotheses, you must decide on the size of the risk of making a type I error that you are prepared to accept, e.g. you may decide that a 5% risk is acceptable (making the wrong decision once in 20 times) or you may require only a 1% risk (wrong once in a 100 times). Then divide all the possible values of the sample measure into two sets, putting (say) 5% of them in the critical or rejection region and 95% in the acceptance region. The percentage chosen is called the level of significance. Next calculate the value of the measure for the sample you are using, see which region it belongs to and state your decision.

Hypothesis Tests About One Mean

The important thing to remember about hypothesis tests is that there should be no personal bias in the decision, i.e. the test must be set up so that the same decision is reached whoever carries out the test. With this in mind, a procedure has been designed that should be followed for every test. The steps are listed below for tests about means, but these steps should be included in tests about any measure.

The necessary steps are as follows:

(1) State the null hypothesis, H_o.

(2) Decide on the alternative hypothesis, H_1.

(3) Choose the level of significance, thus fixing the critical region.

(4) Calculate the mean and standard deviation of the sample, if these are not given in the problem, and the standard error of the mean.

(5) Calculate the standardised z statistic.

(6) Accept or reject the null hypothesis.

Now we will look at these steps in detail:

(1) The null hypothesis is H_o: μ = the value of the population mean (μ_o say) given in the problem.

(2) The alternative hypothesis depends on the wording of the problem. The wording can suggest one of three possible meanings:

(a) The sample comes from a population the mean of which is not equal to μ_o, i.e. it may be smaller or larger. Then you take H_1: $\mu \neq \mu_o$.

For this alternative you divide the critical region into two equal parts and put one in each tail of the distribution, as shown in Figure 15.1. This is called a **two-tailed test**.

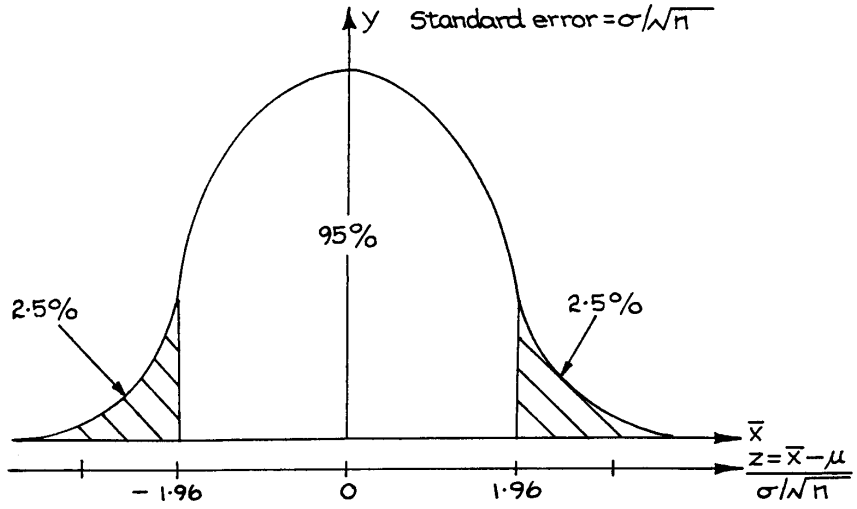


Figure 15.1: Two-Tailed Test

Figure 15.1 shows the critical region (hatched areas) for a two-tailed test when the level of significance is 5%. The z-scale at the bottom shows the critical values of the z statistic. (Notice that the acceptance region is the region used for a 95% confidence interval for μ.)

(b) The sample comes from a population the mean of which is larger than μ_o. Then you take H_1: $\mu > \mu_o$ and put the whole of the critical region in the right-hand tail of the distribution, as shown in Figure 15.2. This is called an **upper-tail test**.

(c) The sample comes from a population with μ smaller than μ_o, so take H_1: $\mu < \mu_o$ and put the whole of the critical region in the left-hand tail of the distribution. This is called a **lower-tail test** and it would be shown in a figure similar to Figure 15.2 with the hatched areas in the left-hand tail and the value –1.645 shown on the z-scale.

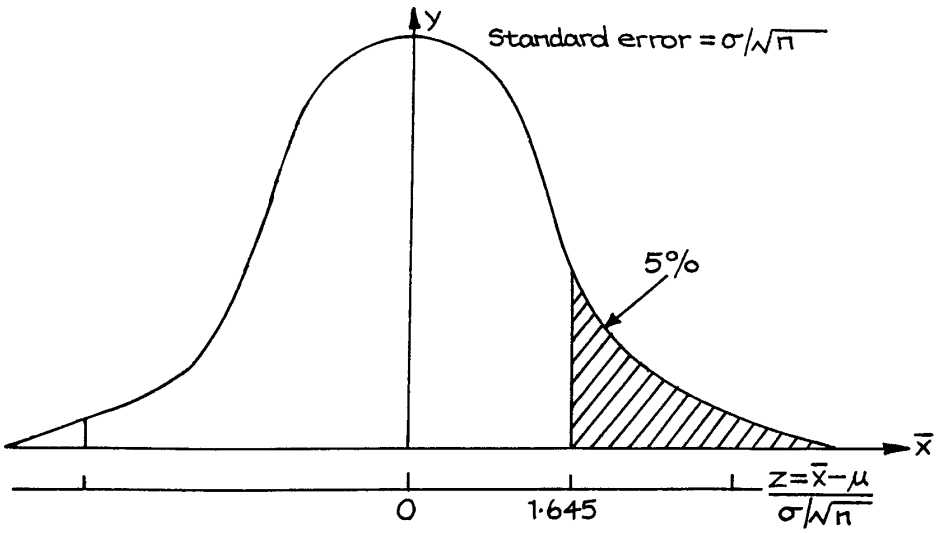


Figure 15.2: Upper-Tail Test

(3) Decide what risk of making the wrong decision you are prepared to accept. State this risk as the value of the level of significance and also state the corresponding critical values of z and

define the critical region. Table 15.2 shows the critical z values for the three most commonly used levels of significance. These values have been calculated from the standard normal tables.

Table 15.2: Critical z Values

Level of Significance	U-Tail Test	L-Tail Test	Two-Tailed Test
5%	1.645	–1.645	–1.96 and 1.96
1%	2.326	–2.326	–2.576 and 2.576
0.1%	3.09	–3.09	–3.29 and 3.29

You will use these values (particularly the first two) so often that it is worth memorising them.

(4) In exam questions you will usually be given $\bar{x}$ and s, but in practice you may have to calculate them from the sample data (see earlier study units).

If you are given σ in the problem, calculate $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

If you are not given σ, use SE = $\frac{s}{\sqrt{n}}$.

(5) Calculate $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

and state whether the value lies in the critical region or the acceptance region.

(6) **Reject H_o if z lies in the critical region, accept H_o if z lies in the acceptance region** and then state the meaning of your decision in terms of the actual problem you are solving. H_o is rejected if z is equal to the critical value used.

If z lies in the **acceptance region** for the level used, the result of the test is said to be **not significant**.

If z lies in the **critical region at the 5% level**, the result of the test is said to be **significant**.

If z lies in the **critical region at the 1% level**, the result of the test is said to be **highly significant**.

If z lies in the **critical region at the 0.1% level**, the result of the test is said to be **very highly significant**.

In the following examples, the steps in the testing procedure are numbered so that you can follow the method easily.

Example 1

A sample of 100 fluorescent light tubes has a mean life of 20.5 hours and a standard deviation of 1.6 hours. Test:

(a) At the 1% level whether the sample comes from a population with mean 23.2 hours.

(b) At the 5% level whether it comes from a population with mean 20.8 hours.

(c) At the 5% level whether it comes from a population with mean less than 20.8 hours.

Answer

(a) (1) H_o: $\mu = 23.2$

(2) H_1: $\mu \neq 23.2$.

(3) Level of significance = 1%.

Critical values of z are –2.576 and 2.576.

Therefore, critical region is either $z < -2.576$ or $z > 2.576$

(4) $\bar{x} = 20.5$, s = 1.6, n = 100, so SE = $\frac{1.6}{10} = 0.16$

(5) $z = \frac{20.5 - 23.2}{0.16} = \frac{-2.7}{0.16} = -16.875 < -2.576$

So z lies in the critical region.

(6) This result is highly significant so reject H_o, i.e. the evidence of this sample suggests that the population mean is not 23.2 hours.

(b) (1) H_o: $\mu = 20.8$

(2) H_1: $\mu \neq 20.8$

(3) Level of significance = 5%

Critical values of z are –1.96 and 1.96

Therefore, critical region is either $z < -1.96$ or $z > 1.96$

(4) As in (a), SE = 0.16

(5) $z = \frac{20.5 - 20.8}{0.16} = \frac{-0.3}{0.16} = -1.875 > -1.96$

So z does not lie in the critical region.

(6) This result is not significant so accept H_o, i.e. the evidence of this sample suggests that the population mean is 20.8 hours.

(c) (1) H_o: $\mu = 20.8$

(2) H_1: $\mu < 20.8$

(3) Level of significance = 5%

Critical value of z is –1.645

Therefore, critical region is $z < -1.645$

(4) As in (a), SE = 0.16

(5) As in (b), $z = -1.875 < -1.645$

So z lies in the critical region.

(6) This result is significant so reject H_o, i.e. the evidence of this sample suggests that the population mean is less than 20.8.

Example 2

A sample of 150 students had an average IQ of 112 with a standard deviation of 9.

(a) At what level of significance would you accept that this sample is taken from a student population with average IQ of 110?

(b) At the 5% level would you accept that the student population had an average IQ greater than 113?

Answer

(a) (1) H_o: $\mu = 110$

(2) H_1: $\mu \neq 110$

(3) (To answer this question you have to test with all three levels of significance, beginning with the largest critical region.)

(i) Level of significance = 5%

Critical values of z are –1.96 and 1.96

Therefore, critical region is $z < -1.96$ or $z > 1.96$

(ii) Level of significance = 1%

Critical values of z are –2.576 and 2.576

Therefore, critical region is $z < -2.576$ or $z > 2.576$

(iii) Level of significance = 0.1%

Critical values of z are –3.29 and 3.29

Therefore, critical region is $z < -3.29$ or $z > 3.29$

(4) $\bar{x} = 112$, $s = 9$, $n = 150$, $SE = \dfrac{9}{\sqrt{150}} = 0.73$

(5) $z = \dfrac{112 - 110}{0.73} = \dfrac{2}{0.73} = 2.74$

(i) $z > 1.96$, so z lies in the critical region.

(ii) $z > 2.576$, so z lies in the critical region.

(iii) $z < 3.29$, so z does not lie in the critical region.

(6) H_o would be rejected at the 5% and 1% levels but accepted at the 0.1% level, i.e. at the 0.1% level the sample provides evidence that the student population has an average IQ of 110.

(b) (1) H_o: $\mu = 113$

(2) H_1: $\mu > 113$

(3) Level of significance = 5%

Critical value of z is 1.645

Therefore, critical region is $z > 1.645$

(4) As in (a), SE = 0.73

(5) $z = \dfrac{112 - 113}{0.73} = \dfrac{-1}{0.73} = -1.37 < 1.645$

So z does not lie in the critical region.

(6) This result is not significant so accept H_o, i.e. the sample evidence suggests that the student population does not have an average IQ greater than 113.

Hypothesis Tests About Two Means

There are occasions when we are not particularly interested in the population from which a sample is taken, but we need to know whether two samples come from the same population. We use the same test procedure to test the difference between the two means. Using the suffixes 1 and 2 to distinguish between the two samples, the hypotheses become:

H_o: $\mu_1 = \mu_2$, i.e. $\mu_1 - \mu_2 = 0$

H_1: (i) $\mu_1 \neq \mu_2$, i.e. $\mu_1 - \mu_2 \neq 0$ Two-tailed

(ii) $\mu_1 > \mu_2$, i.e. $\mu_1 - \mu_2 > 0$ Upper tail

(iii) $\mu_1 < \mu_2$, i.e. $\mu_1 - \mu_2 < 0$ Lower tail

$$\text{SE of } (\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \text{if } \sigma_1 \text{ and } \sigma_2 \text{ are not known}$$

$$= \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{if both populations are known to have the same standard deviations}$$

$$\text{Then } z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\text{SE of } (\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2}{\text{SE } (\bar{x}_1 - \bar{x}_2)} \quad \text{if } H_o \text{ is true}$$

Example

Two factories are producing visual display units for computers. Use the following sample data to test whether the two production lines are producing units with the same mean life length:

	$\bar{x}$	s	n
Sample 1	20.5	3.4	125
Sample 2	19	2.1	180

Answer

(1) H_o: $\mu_1 - \mu_2 = 0$

(2) H_1: $\mu_1 - \mu_2 \neq 0$

(3) Let level of significance = 5%

Critical values of z are −1.96 and 1.96

Therefore, critical region is $z < -1.96$ or $z > 1.96$

(4) $SE = \sqrt{\frac{(3.4)^2}{125} + \frac{(2.1)^2}{180}} = \sqrt{\frac{11.56}{125} + \frac{4.41}{180}} = 0.34$

(5) $z = \frac{20.5 - 19}{0.34} = \frac{1.5}{0.34} = 4.41 > 1.96$

So z lies in the critical region.

(6) This result is significant so reject H_o, i.e. the test result suggests that the two production lines are producing units with different mean life lengths.

Hypothesis Tests About One Proportion

Example 1

A manufacturer of computers claims that his computers are operational for at least 80% of the time. During the course of a year one computer was operational for 270 days. Test, at the 1% level, whether the manufacturer's claim was justified.

Answer

(1) H_o: $p = 0.8$

(2) H_1: $p > 0.8$

(3) Level of significance = 1%

Critical value of z is 2.326

Therefore, critical region is $z > 2.326$

(4) $\hat{p} = \frac{270}{365}$, $n = 365$, $SE = \sqrt{\frac{0.8 \times 0.2}{365}} = 0.021 = \sigma_{\bar{x}}$ since H_o gives p

(5) $z = \frac{\hat{p} - p}{SE} = \frac{0.74 - 0.8}{0.021} = -2.86 < 2.326$

So z does not lie in the critical region.

(6) This result is not significant so accept H_o, i.e. the evidence does not support the manufacturer's claim.

Example 2

The proportion of drivers who plead guilty to driving offences is usually 60%. Out of 750 prosecutions 400 pleaded guilty. Is this proportion significantly different from usual?

Answer

(1) H_o: $p = 0.6$

(2) H_1: $\hat{p} \neq 0.6$

(3) Let the level of significance = 5%

Critical values of z are –1.96 and 1.96

Therefore, critical region is $z < -1.96$ or $z > 1.96$

(4) $\hat{p} = \frac{400}{750} = 0.53$, $n = 750$, $\sigma_{\bar{x}} = \sqrt{\frac{0.6 \times 0.4}{750}} = 0.018$

(5) $z = \frac{0.53 - 0.6}{0.018} = \frac{-0.07}{0.018} = -3.89 < -1.96$

So z lies in the critical region.

(6) This result is significant to reject H_o, i.e. this sample of prosecutions suggests that the proportion of drivers pleading guilty is changing.

Hypothesis Tests About Two Proportions

Example

A market research organisation carried out a sample survey on the ownership of washing machines and concluded that 64% of all households owned a washing machine, out of 200 households sampled. Six months later they repeated the survey, using the same questionnaire, and concluded that 69% owned a washing machine, out of 150 households sampled. Is the difference due to a significant increase in ownership or is it a random sampling error?

The sample data is $\hat{p}_1 = 0.64$, $n_1 = 200$, $p_2 = 0.69$, $n_2 = 150$

Answer

(1) H_o: $p_1 = p_2$, i.e. $p_1 - p_2 = 0$

(2) H_1: $p_1 < p_2$, i.e. $p_1 - p_2 < 0$

(3) Let the level of significance = 5%

Critical value of z is –1.645

Therefore, critical region is $z < -1.645$

(4) As p_1 and p_2 are not known, we have to estimate the SE of their difference. The best estimate of this SE is found by combining the two samples. Let $\hat{p}_o$ = proportion of owners in combined sample.

Number of owners in 1st sample = $n_1 \hat{p}_1$

Number of owners in 2nd sample = $n_2 \hat{p}_2$

Size of combined sample = $n_1 + n_2$

Therefore, $\hat{p}_o = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$ and SE = $\sqrt{\frac{\hat{p}_o\hat{q}_o}{n_1} + \frac{\hat{p}_o\hat{q}_o}{n_2}}$

Substituting the sample values gives:

$$\hat{p}_o = \frac{200 \times 0.64 + 150 \times 0.69}{200 + 150} = \frac{231.5}{350} = 0.66 \text{ and } \hat{q}_o = 0.34$$

$$\text{So SE} = \sqrt{\frac{0.66 \times 0.34}{200} + \frac{0.66 \times 0.34}{150}} = 0.051$$

(5) $$z = \frac{\hat{p}_1 - \hat{p}_2}{SE} = \frac{0.64 - 0.69}{0.051} = \frac{-0.05}{0.051} = -0.98 > -1.645$$

So z does not lie in the critical region

(6) This result is not significant so accept H_o, i.e. the test result suggests that the apparent small rise in ownership is due to sampling error.

E. NEGATIVE AND POSITIVE PROOF

Now that you are familiar with the basic routine of statistical significance testing we can look a little deeper at the basis of this technique and extend the areas in which we can use it.

When we set up a null hypothesis to test, we are actually assuming the opposite to what we want to find. In general, when we carry out a test we hope to be able to show that there is a change in some quantity or a difference between two quantities. The null hypothesis H_o takes the form "no change, no increase, no difference." To meet what we want, we hope to get the result "reject H_o"; if we can reject the hypothesis of no change or no difference we are left with the conclusion that there has been a significant change or there is a significant difference. This is a powerful result. To see why let us take an example.

Suppose I am standing by the side of a road, along which cars are passing. I make the hypothesis that there are no cars painted yellow in the area. I watch for half an hour and I see no yellow cars. This supports my hypothesis, but does not prove it. I watch for another half an hour, and again see no yellow cars. I may start to think that my hypothesis is true, but I have got no nearer to proving it. In fact, I can watch for as long as I want to and I will never be able to say with certainty that there are no yellow cars in the area. Positive proof, which agrees with our hypothesis, is weak. Now suppose I am watching and I see a yellow car. Immediately the hypothesis has been disproved. The negative proof is powerful, as one piece of information which contradicts the hypothesis is sufficient to disprove it.

The only difference in statistical testing is that we do not have a yes/no answer; our hypothesis takes the form of a distribution and so we can never be completely certain that we are right in rejecting a hypothesis. The type I error is the error of rejecting as false a hypothesis which is true.

When we are testing in this way there is always the possibility that a better test, which normally means one based on a bigger sample, will lead to a rejection. It may be preferable to quote the two alternative results of a significance test in the form:

"Reject H_o" , ***or*** "Do not reject H_o"

The second alternative then includes the possibilities that either the hypothesis is true, or we did not have a powerful enough test to disprove it. We can never tell which is correct.

F. DIFFERENCES

When we are using significance tests there are three types of differences or changes which can concern us:

- Observed or expected differences or changes;
- Significant differences or changes;
- Important differences or changes.

In general we carry out a test because we either have seen or suspect a change or difference, or because we expect one. Of course if we collect data and find no change or difference there is no need to go further with the test. The supporters of the idea of global warming base their case on observations which show an increase, on average, in the temperature around the world. As there is a small increase in the average temperature over recent years, there is a basis for carrying out a test.

In the second case we do or observe something which we think will cause a difference or change. Examples from industry or commerce would be:

- Expected reduction in breakdowns per week for machinery after an increase in expenditure on maintenance;
- Expected increase in sales after an increase in expenditure on advertising;
- Expected reduction in absenteeism after the introduction of an incentive scheme.

Notice that in each case we would carry out a one-tailed test, as we know before we collect the data what change we are looking for. Again, we only proceed with a test if there turns out to be a change of the sort we are looking for.

A significant difference or change is one which satisfies the requirements to reject the null hypothesis. The level of significance should be stated, i.e. “Significant at 1%”. If this is omitted you can assume that the test was carried out at the 5% level, which is the default level used by all statisticians.

Lastly we must consider if the change or difference detected matters. The word “significant” used in everyday speech carries the idea of importance with it, but this is not so in statistics. A difference may be significant statistically, but unimportant economically. To make such a judgement you must have a knowledge of the context in which the test was carried out. For example, a sales promotion campaign may produce a statistically significant increase in sales, but if the increase does not produce more in extra revenue than we spent on the campaign, it is of no interest or value.

G. SIGNIFICANCE LEVELS

It is essential that you can realise and explain the implications of choosing a particular significance level. The most useful way to get a “feel” for this measure is to think of it as the chance of being wrong when you say there is a change or difference. If we know the cost of being wrong, we can work out the expectation of being wrong:

$$\text{Expectation} = \text{P(Wrong)} \times \text{Cost of being wrong}$$

If the cost of being wrong is small, we can afford to take a higher risk of being wrong. A good example would be in the early stages of an investigation of a pharmaceutically-active substance which might be developed into a new drug. If we decide on the basis of a test that something may be causing a useful effect and we keep it for further testing, the cost will be low at the start of testing if

we later reject it. We would err on the side of keeping the substance in our tests. In this circumstance, a 10% chance of keeping something which later turns out to be useless would be quite acceptable.

Consider now the end of the process when a new drug has been produced. It must be tested extensively for undesirable side effects, because we know that if it is sold and side effects develop, the company selling it may face legal actions which might result in the payment of extremely large damages. The risk of incurring such costs has to be kept very low, and so very small values of the significance level will be used. Just to confuse us, they are sometimes referred to as "very high significance", meaning we have very little doubt that our conclusions are correct.

If we cannot estimate costs in this way, the default value is the 5% significance level. The best way to explain it is to say "In the absence of any information which would make a different significance level a better choice, we use the 5% level."

H. SMALL SAMPLE TESTS

The t-distribution

In the tests we did earlier, the central limit theorem allowed us to state that the means or differences we were testing were normally distributed, so we could calculate the standard deviation of a sample mean or difference from the population standard deviation and the sample size. This is fine if we know the population standard deviation. Often we do not, all the data we have is represented by a small sample. For example, suppose a supplier claims that his product has a content of a main ingredient of 40%. We measure the content in a sample of 5 bags of the product (chosen at random) and we obtain the following values:

38%, 32%, 44%, 37%, 34% with a mean of 37%.

Does this sample invalidate the supplier's claim?

Before we carry out the test, consider our information. We can use the given figures to estimate the standard deviation of the population of all bags of the product. How reliable is a sample of 5? Not very. So it would seem to make sense to demand a greater difference between the observed mean value of 37% and the claimed value of 40%, than if we could carry out the test with a value of the standard deviation based on a large amount of data. How much allowance for this doubt in the value of the standard deviation must we put into the test? The answer was given by a statistician called W S Gosset. He published his work under the pseudonym of "A Student of Statistics", and used the letter *t* for the test parameter. Ever since, the distribution he derived has been referred to as "Students *t*-distribution".

Here is how it works: in the tests we have done already we calculated a z value, and tested it against critical values derived from the normal distribution, i.e. + or – 1.96 for a two-tailed test at the 5% level. The distribution of *t* which replaces *z* looks like the normal distribution, but has a sharper central peak and higher tails. To get the same tail areas which determine the critical values we have to go further from the mean. In other words we need a larger difference for it to be taken as significant than when we had better information. The difference between the normal distribution and the *t*-distribution with the same mean and standard deviation is shown in Figure 15.3. For a sample of 5 we require values of *t* beyond + or – 2.78.

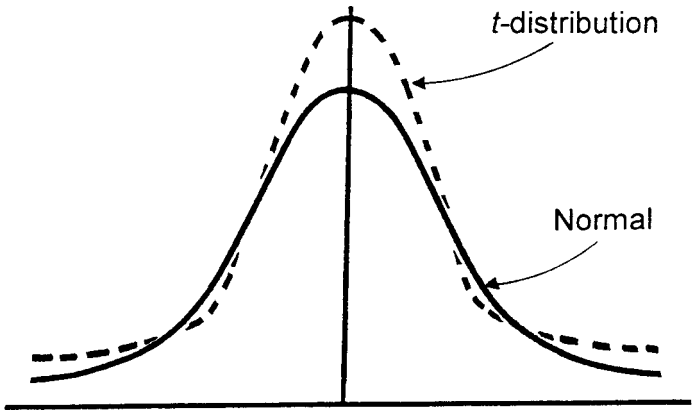


Figure 15.3

Standard Errors for Small Samples

If we were to take a large number of samples, each of 5 items from the population of all bags of the products, and for each calculate the standard deviation using the formula we have used so far, then found the mean of these values, we would get a value for the standard deviation which is lower than the true value. This would be so no matter how many samples we averaged the value over. We say that the estimate of the population standard deviation is biased.

We can correct for the bias by multiplying the sample standard deviation *s* by:

$$\sqrt{\frac{n}{(n-1)}}.$$

Alternatively, in calculating the standard deviation we use the divisor $(n - 1)$ instead of n in the calculation, which will give us the best estimate of the population standard deviation, σ. This is written $\hat{\sigma}$.

Degrees of Freedom

The divisor $(n - 1)$ is called the "degrees of freedom" (do not worry about why). The degrees of freedom also determine how close the *t*-distribution is to the normal distribution. In fact when the degrees of freedom is greater than about 30, we usually ignore the difference and carry out tests as we did earlier.

To find the critical values for *t* we have to know three things:

- The significance level.
- One or two-tailed test?
- Degrees of freedom.

Tables are available to give us the values. If you look at a *t* table you will see the familiar values of 1.96 and 1.65 at the bottom of the columns when the degrees of freedom become very large.

Let's Do the Test

$H_o : \mu = 40\%$

Sample values of x are 38%, 32%, 44%, 37%, 34%.

Sample mean = $\bar{x} = \frac{\Sigma x}{5} = \frac{185}{5} = 37\%$

$$\hat{\sigma} = \sqrt{\frac{\Sigma(x-\bar{x})^2}{(n-1)}} = \sqrt{\frac{(1^2+5^2+7^2+0^2+3^2)}{4}}$$

$$= 4.5826$$

SE of sample mean = $\frac{\hat{\sigma}}{\sqrt{n}} = \frac{4.5826}{\sqrt{5}} = 2.05$

From tables, critical value of *t* at 5% level, two-tailed test with degrees of freedom = (n – 1) = 4, *t* + or – 2.78,

$t = \frac{37-40}{2.05} = 1.46$ (N.S.)

The value of *t* lies well within the range which is consistent with the null hypothesis, so we cannot reject the null hypothesis (shown by the letters N.S. for "Not Significant" above). In terms of our original question, we do not have enough evidence to say that the average content differs from 40%.

The t test and Proportions

When testing proportions or differences in proportions you will rarely, if ever, meet a case where the sample sizes are small enough to require the use of the *t*-distribution.

Difference of Means Test

Two groups of students are asked to take a test. The results are:

- Group A scores 45, 87, 64, 92, 38.
- Group B scores 40, 35, 61, 50, 47, 32.

Is there a significant difference in mean score between the two groups?

Group A: mean = $\bar{x} = \frac{326}{5} = 65.2$

$$\hat{\sigma}_A = \sqrt{\frac{2342.8}{4}} = 24.2012$$

Group B: mean = $\bar{x} = \frac{265}{6} = 44.17$

$$\hat{\sigma}_B = \sqrt{\frac{574.8333}{5}} = 10.7226$$

H_o : no difference in mean score.

SE of difference in scores $= \sqrt{\frac{\hat{\sigma}_A^{\,2}}{n_A} + \frac{\hat{\sigma}_B^{\,2}}{n_B}}$

$$= \sqrt{\frac{24.2012^2}{5} + \frac{10.7226^2}{6}}$$

$$= 11.67$$

Degrees of freedom = $(n_A - 1) + (n_B - 1) = (5 - 1) + (6 - 1) = 9$.

Critical values of t, 5% level, two-tailed test = + or – 2.26

$$t = \frac{65.2 - 44.17}{11.67} = 1.80 \text{ (N.S.)}$$

There is no significant difference in scores for the two groups.

You should now realise that to have a significant result when using small samples, large changes or differences are essential. The statistical test will often say "not significant" when there does appear to be a change or difference. In this case it is likely that with more data a significant result will be found.

I. SUMMARY

This study unit is very important – examination questions are frequently set on confidence intervals and hypothesis tests. Although you may use any levels of significance you like, it is best to keep to the three given here as experience has shown that they are of the most practical use.

You must realise that statistical tests do not prove or disprove hypotheses in the mathematical sense. They only provide evidence in support of one or other of the two hypotheses chosen.

You should be quite sure that you know the following thoroughly:

- How to use the standard normal table.
- The formulae for confidence intervals.
- The procedure for carrying out hypothesis tests.

Whenever we have good information about the variability in a given situation, i.e. σ has been estimated from a large amount of data, we can use a z-test, no matter how small the samples we are testing. It is not the size of sample which decides whether or not we use z or t, but whether we used the sample to provide an estimate of σ.

Study Unit 16

Non-parametric Tests and Chi-squared

Contents ***Page***

A. NON-PARAMETRIC TESTS

In Unit 15, we learned how to test hypotheses using data from one or two samples. We used one-sample tests to determine if a mean was significantly different from a hypothesised value and two-sample tests to determine if the difference between two means was significant. These tests are known as parametric tests, because they involve testing the parameters of a population, such as the mean and the proportions. They use the parametric statistics of samples that come from the population being tested. To formulate these tests, we make assumptions about the population, for example, that the population is normally distributed.

There are certain kinds of data that cannot be tested in this way, such as data which was not collected in a random sample and therefore does not have a normal distribution; ordinal data; ranked data; and data from more than two populations. In business, we often encounter data of this type, such as:

- the results of a survey of which brand of washing powder consumers prefer;
- an analysis of the arrival of customers at supermarket checkouts;
- a survey of employees' attitudes towards performance appraisal in different departments;
- a study of whether male staff have been more successful in passing professional examinations than female staff.

For these types of data, it is necessary to use tests which do not make restrictive assumptions about the shape of population distributions. These are known as non-parametric tests. Non-parametric tests have certain advantages over parametric tests:

- it is not necessary to assume that the population is distributed in the shape of a normal curve, or any other specific shape;
- generally, they are quicker to do and easier to understand. Sometimes, formal ranking or ordering is not necessary.

But non-parametric methods are often not as precise as parametric tests, because they use less information.

In this Unit, we are going to consider one of the most commonly used non-parametric tests, called the chi-squared test (pronounced "ki-squared"). Critical values in the chi-squared test are denoted by the symbol χ^2. The chi-squared test is principally used to determine:

- if two population attributes are independent of each other; or
- if a given set of data is consistent with a particular distribution, known as the "goodness of fit" test.

We will consider each of these versions of the chi-squared test in turn.

B. CHI-SQUARED AS A TEST OF INDEPENDENCE

Managers often need to know whether observable differences between several sample proportions are significant or only due to chance. For example, if the evaluation of data shows that a new method of training staff by using open learning materials results in higher outputs than the old method of on-the-job training with an experienced employee, the personnel manager may decide that the new method of training should be introduced throughout the organisation.

There is a series of steps which need to be undertaken to determine if two attributes are dependent or independent of one another:

(a) formulating the null and alternative hypotheses;

(b) constructing a contingency table;

(c) calculating the chi-squared statistic of the sample;

(d) determining the appropriate number of degrees of freedom;

(e) ascertaining whether the sample statistic falls inside the acceptance region.

We will now consider each of these in more detail, taking as an example the results of a survey of staff preferences in different locations towards two car leasing schemes.

Formulating the hypotheses

In Unit 15, we examined how to formulate null and alternative hypotheses in order to discover whether a certain set of data came from a particular population. In the same way, we also formulate null and alternative hypotheses to determine whether the two population attributes are independent of each other. In our example, the two hypotheses would be:

- the null hypothesis, H_0, that employees' preferences for the two car leasing schemes are independent of their work location;
- the alternative hypothesis, H_1, that employees' preferences for the two car leasing schemes are not independent of their work location.

Having formulated the hypotheses, we then need to decide upon an appropriate level of significance. As we have already learned in Unit 15, the higher the significance level we use for testing a hypothesis, the greater the probability of a Type I error, that is, of rejecting a null hypothesis when it is true. In our example, let us assume that the organisation wants to test the null hypothesis at a 10% level of significance.

Constructing a contingency table

The results of the survey of employees' preferences for the two car leasing schemes in different locations are presented in Table 16.1 below. This is known as a **contingency table**. It is made up of rows and columns, each showing a basis of classification – the rows classify the information according to preference and the columns classify the information according to location. Because this particular table has 2 rows and 3 columns, it is called a "2×3 contingency table" (note that the "Total" row and "Total" column are not counted).

Table 16.1

Preference	Location A No. of staff	Location B No. of staff	Location C No. of staff	Total No. of staff
Scheme 1	76	66	61	203
Scheme 2	42	35	49	126
Total	118	101	110	329

Calculating the chi-squared statistic of the sample

To evaluate the results of the survey, we start by assuming that there is no connection between the two attributes of preference and location. In other words, we assume that the null hypothesis is correct. If this is so, then we would expect the results of the staff survey to be in proportion at each location. To carry out the test, we therefore need first to calculate what the **expected** frequencies would be, assuming that there is no connection, and then to compare these with the **observed** frequencies.

We calculate the expected frequencies by first combining data from all three locations in order to estimate the overall proportion of employees who prefer each scheme. We start first by combining the data of all those who prefer Scheme 1, as follows:-

$$\frac{76+66+61}{118+101+110} = \frac{203}{329} = 0.6170$$

If 0.6170 is the estimate of the proportion of employees who prefer Scheme 1, then 0.3830 (1 – 0.6170) will be the estimate of the proportion of employees who prefer Scheme 2. Using these two values, we can estimate the number of employees in each location whom we would expect to prefer each of the two car leasing schemes. These calculations are shown in Table 16.2 below.

Table 16.2

	Location A	Location B	Location C
Total number sampled	118	101	110
Estimated proportion who prefer 1	× 0.6170	× 0.6170	× 0.6170
Number expected to prefer 1	72.81	62.32	67.87
Total number sampled	118	101	110
Estimated proportion who prefer 2	× 0.3830	× 0.3830	× 0.3830
Number expected to prefer 2	45.19	38.68	42.13

The expected (estimated) and observed (actual) frequencies at each location are brought together in Table 16.3 below.

Table 16.3

	Location A	**Location B**	**Location C**
Frequency preferring 1			
Observed	76	66	61
Expected	72.81	62.32	67.87
Frequency preferring 2			
Observed	42	35	49
Expected	45.19	38.68	42.13

We then calculate the chi-squared statistic of the sample, which is expressed as:

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

where: f_0 = observed frequency, and

f_e = expected frequency.

To obtain χ^2, we first subtract f_e from f_0 for each combination of preference and location shown in Table 16.3. For example, the calculation for those preferring Scheme 1 at Location A is:

$f_0 - f_e = 76 - 72.81$

$= 3.19$

When all six calculations have been carried out, we then square each of the resulting values. For example, $3.19^2 = 10.18$. Remember that when negative values are squared, the result is always a positive figure, so that, for example, $-6.87^2 = 47.20$.

Next, we divide each squared value by f_e. For example, 10.18 is divided by 72.81, which gives 0.1398.

Finally, the six resulting values are summed.

The results of these calculations are shown in Table 16.4 below.

Table 16.4

f_0	f_e	$f_0 - f_e$	$(f_0 - f_e)^2$	$\frac{(f_0 - f_e)^2}{f_e}$
76	72.81	3.19	10.18	0.1398
66	62.32	3.68	13.54	0.2173
61	67.87	–6.87	47.20	0.6954
42	45.19	–3.19	10.18	0.2253
35	38.68	–3.68	13.54	0.3501
49	42.13	6.87	47.20	1.1203
				2.7482

The chi-squared statistic of the sample is therefore 2.7482.

Determining the appropriate number of degrees of freedom

To use the chi-squared test to ascertain whether the sample statistic falls inside or outside the acceptance region, we first have to calculate the number of **degrees of freedom** (known as **df**), in the contingency table, using the formula:

(number of rows r – 1) (number of columns c – 1)

As we saw earlier, the contingency table we are using (Table 16.1) has 2 rows and 3 columns, so the appropriate number of degrees of freedom is as follows:

df = (r–1) (c–1)

= (2–1) (3–1)

= 1×2

= 2

Ascertaining whether the sample statistic falls inside the acceptance region

We then use the statistical table for the Area in the Right Tail of a Chi-squared Distribution, set out in the Appendix to this unit, to find the value of the chi-squared statistic. In our example, we set a significance level of 10% and therefore we look in the column headed .010. Then we read down to the 2 degrees of freedom row and find that the value of the chi-squared statistic is 4.61.

This means that the region to the right of the chi-squared value of 4.61 contains 10% of the area under the distribution curve, as we can see from Figure 16.1 below. Therefore the acceptance region for the null hypothesis goes from the left tail of the curve along to the chi-squared value of 4.61. The sample chi-squared value which we calculated is 2.7482 and because this value is less than 4.61, it falls within the acceptance region. We can therefore accept the null hypothesis that employees' preference about the two car leasing schemes is independent of their work location.

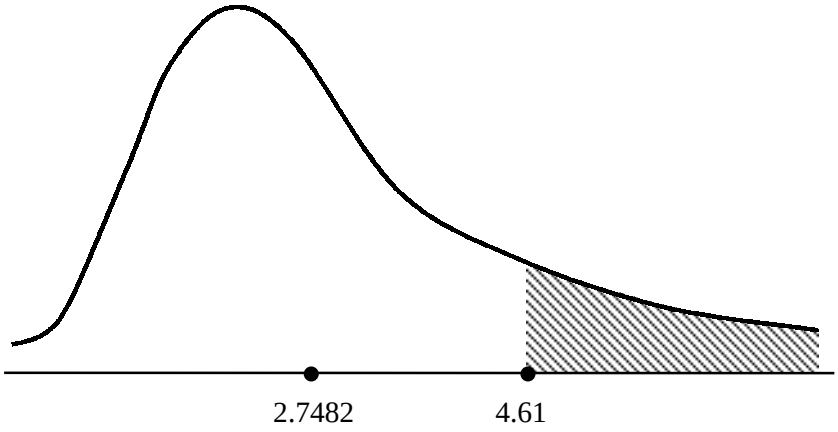


Figure 16.1

Using the chi-squared hypothesis test

When using the chi-squared hypothesis test, the size of your sample must be large enough to ensure that the theoretically correct distribution (the distribution assumed by the chi-squared table in the Appendix) and your sample distribution are similar. If the expected frequencies are too small (less than about 5), the sample value of chi-squared will be overestimated and this will therefore result in too many rejections of the null hypothesis.

When interpreting the chi-squared statistic of a sample, a large value, such as 20, indicates a big difference between observed and expected frequencies, whereas a value of zero indicates that the observed and expected frequencies exactly match. If you find that the sample value of chi-squared is at or near zero, but you believe that a difference between observed and expected values should exist, it is sensible to re-examine the data and the way it was gathered and collated, to make sure that any genuine differences have not been missed.

C. CHI-SQUARED AS A TEST OF GOODNESS OF FIT

The chi-squared test can also be used to determine if there is a similarity – that is, a **good fit** – between the distribution of observed data and a theoretical probability distribution, such as the normal distribution (which we considered in Unit 12) or the binomial or Poisson distributions (which we considered in Unit 13). We first perform the chi-squared test to establish whether there is a significant difference between our observed distribution and the theoretical distribution we have chosen and this information then enables us to decide whether the observed data is a sample from our hypothesised theoretical distribution. The chi-squared goodness of fit test is therefore a useful tool for managers, who often need to make decisions on the basis of statistical information. For example, a maintenance manager at a factory may use information about the frequency of breakdowns to decide how many engineers to deploy on each shift.

There is a series of steps which need to be undertaken to determine goodness of fit:

(a) formulating the null and alternative hypotheses;

(b) calculating the expected frequencies;

(c) calculating the chi-squared statistic of the sample;

(d) determining the appropriate number of degrees of freedom;

(e) ascertaining whether the sample statistic falls inside the acceptance region.

We will now consider each of these in more detail, taking as an example data collected on the amount of money consumers spend on chocolate each week.

Formulating the null and alternative hypotheses

The sales manager of a confectionery manufacturer has commissioned a survey of the amount of money consumers spend on chocolate each week. He/she believes that the variable – the amount of money spent – may be approximated by the normal distribution, with an average of £5.00 per week and a standard deviation of £1.50.

The hypotheses would therefore be:

- the null hypothesis, H_0, that a normal distribution is a good description of the observed data;
- the alternative hypothesis, H_1, that a normal distribution is not a good description of the observed data.

Let us assume that the sales manager wants to test the null hypothesis at a 10% level of significance.

Calculating the expected frequencies

The results of the consumer survey are shown in Table 16.5 below.

Table 16.5

Weekly expenditure £	**Number of consumers**
< £2.60	12
£2.60 – £3.79	60
£3.80 – £4.99	82
£5.00 – £6.19	104
£6.20 – £7.39	24
≥ £7.40	18
Total	300

To determine what the expected frequencies would be under a normal distribution, we use the same techniques as we have already used in Unit 12 to ascertain areas under the normal distribution curve.

First, we start with the formula:

$$z = \frac{x - \mu}{\sigma}$$

where: x = value of the random variable

μ = mean of the distribution of the random variable

σ = standard deviation of the distribution

z = number of standard deviations from x to the mean.

Then we look up the value given for z using the Standard Normal Table set out as the Appendix to unit 12.

Finally, we multiply this value by the size of the sample to obtain the expected frequency.

For example, to calculate the expected frequency of consumers spending less than £2.60 per week under a normal distribution, we first obtain z as follows:

$$z = \frac{x - \mu}{\sigma}$$

$$= \frac{2.59 - 5}{1.5}$$

$$= \frac{-2.41}{1.5}$$

$$= -1.61$$

Then we look up the z value of 1.61 in the Appendix to unit 12 (we can disregard the minus sign, because a normal distribution is symmetrical), which gives 0.4463. This is the area under a normal curve between the mean and £2.60. Since the left half of a normal curve (between the mean and the left hand tail) represents an area of 0.5, we can obtain the area below £2.60 by subtracting 0.4463 from 0.5, which gives 0.0537.

Next, we multiply 0.0537 by 300 (the size of the sample population), to obtain the expected frequency. This gives a value of 16.11. Under a normal distribution, the expected frequency of consumers spending less than £2.60 per week on chocolate is therefore 16.11.

The expected frequencies calculated for each class of expenditure are shown in Table 16.6 below.

Table 16.6

Weekly expenditure £	**Observed frequency**	**Normal probability**	**Population**	**Expected frequency**
< £2.60	12	0.0537	× 300	16.11
£2.60 – £3.79	60	0.1571	× 300	47.13
£3.80 – £4.99	82	0.2881	× 300	86.43
£5.00 – £6.19	104	0.2852	× 300	85.56
£6.20 – £7.39	24	0.1560	× 300	46.8
≥ £7.40	18	0.0548	× 300	16.4
Total	300			298.43

Calculating the chi-squared statistic of the sample

We can now calculate the chi-squared statistic of the sample, using the formula:

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

where: f_0 = observed frequency, and

f_e = expected frequency.

The calculations are shown in Table 16.7 below.

Table 16.7

f_0	f_e	$f_0 - f_e$	$(f_0 - f_e)^2$	$\frac{(f_0 - f_e)^2}{f_e}$
12	16.11	–4.11	16.89	1.0484
60	47.13	12.87	165.64	3.5145
82	86.43	–4.43	19.63	0.2271
104	85.56	18.44	340.03	3.9742
24	46.8	–22.8	519.84	11.1077
18	16.4	1.6	2.56	0.1561
				20.03

The chi-squared statistic of the sample is therefore 20.03.

Determining the appropriate number of degrees of freedom

The number of degrees of freedom is determined by the number of classes (expressed as *k*), for which we have compared the observed and expected frequencies. However, the actual number of observed frequencies that we can freely specify is $k - 1$, because the last one is always determined by the size of the sample. One additional degree of freedom must also be subtracted from *k* for each population parameter that has been estimated from the sample data.

In our example, there are six classes, so $k = 6$, but because the total number of observed frequencies must add up to 300, the actual number of frequencies that we can freely specify is only 5. We therefore subtract one degree of freedom from *k*, so the number that we can freely specify is $k - 1$, which is 5.

Furthermore, we had to use the sample mean to estimate the population mean, so we must subtract another degree of freedom, leaving $k - 2$, which is 4.

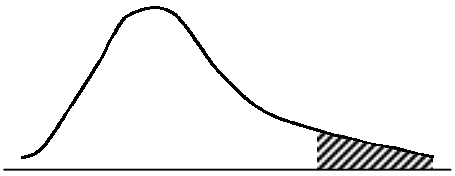
Finally, we also had to use the sample standard deviation to estimate the population standard deviation, so we must subtract another degree of freedom, leaving $k - 3$, which is 3.

Ascertaining whether the sample statistic falls inside the acceptance region

We then use the statistical table for the Area in the Right Tail of a Chi-squared Distribution, set out in the Appendix, to find the value of the chi-squared statistic. In our example, we set a significance level of 10% and therefore we look in the column headed 0.10. Then we read down to the 3 degrees of freedom row and find that the value of the chi-squared statistic is 6.25.

This means that the region to the right of the chi-squared value of 6.25 contains 10% of the area under the distribution curve. Therefore the acceptance region for the null hypothesis goes from the left tail of the curve along to the chi-squared value of 6.25. The sample chi-squared value which we calculated is 20.03 and because this value is greater than 6.25, it falls outside the acceptance region. We therefore cannot accept the null hypothesis that a normal distribution is a good description of the observed frequencies.

APPENDIX: AREA IN THE RIGHT TAIL OF A CHI-SQUARED (χ^2) DISTRIBUTION



Degrees of freedom	0.25	0.10	0.05	0.025	0.01	0.005	0.001
1	1.32	2.71	3.84	5.02	6.63	7.88	10.8
2	2.77	4.61	5.99	7.38	9.21	10.6	13.8
3	4.11	6.25	7.81	9.35	11.3	12.8	16.3
4	5.39	7.78	9.49	11.1	13.3	14.9	18.5
5	6.63	9.24	11.1	12.8	15.1	16.7	20.5
6	7.84	10.6	12.6	14.4	16.8	18.5	22.5
7	9.04	12.0	14.1	16.0	18.5	20.3	24.3
8	10.2	13.4	15.5	17.5	20.1	22.0	26.1
9	11.4	14.7	16.9	19.0	21.7	23.6	27.9
10	12.5	16.0	18.3	20.5	23.2	25.2	29.6
11	13.7	17.3	19.7	21.9	24.7	26.8	31.3
12	14.8	18.5	21.0	23.3	26.2	28.3	32.9
13	16.0	19.8	22.4	24.7	27.7	29.8	34.5
14	17.1	21.1	23.7	26.1	29.1	31.3	36.1
15	18.2	22.3	25.0	27.5	30.6	32.8	37.7
16	19.4	23.5	26.3	28.8	32.0	34.3	39.3
17	20.5	24.8	27.6	30.2	33.4	35.7	40.8
18	21.6	26.0	28.9	31.5	34.8	37.2	42.3
19	22.7	27.2	30.1	32.9	36.2	38.6	43.8
20	23.8	28.4	31.4	34.2	37.6	40.0	45.3
21	24.9	29.6	32.7	35.5	38.9	41.4	46.8
22	26.0	30.8	33.9	36.8	40.3	42.8	48.3
23	27.1	32.0	35.2	38.1	41.6	44.2	49.7
24	28.2	33.2	36.4	39.4	43.0	45.6	51.2
25	29.3	34.4	37.7	40.6	44.3	46.9	52.6
26	30.4	35.6	38.9	41.9	45.6	48.3	54.1
27	31.5	36.7	40.1	43.2	47.0	49.6	55.5
28	32.6	37.9	41.3	44.5	48.3	51.0	56.9
29	33.7	39.1	42.6	45.7	49.6	52.3	58.3
30	34.8	40.3	43.8	47.0	50.9	53.7	59.7
40	45.6	51.8	55.8	59.3	63.7	66.8	73.4
50	56.3	63.2	67.5	71.4	76.2	79.5	86.7
60	67.0	74.4	79.1	83.3	88.4	92.0	99.6
70	77.6	85.5	90.5	95.0	100	104	112
80	88.1	96.6	102	107	112	116	125
90	98.6	108	113	118	124	128	137
100	109	118	124	130	136	140	149

Study Unit 17

Applying Mathematical Relationships to Economic Problems

Contents *Page*

A. FUNCTIONS, EQUATIONS AND GRAPHS

Revision

The purpose of mathematics in this course is to provide a theoretical basis to some of the concepts we study in economics. We start this unit by revising some basic arithmetic and algebra, which we will then go on to use in our analysis of demand and supply.

You will already be familiar with the four basic arithmetical operations:

Addition +

Subtraction –

Multiplication ×

Division ÷

To undertake a series of calculations, we use the **order of operations** rule, which sets out the order in which we need to perform the calculations, as follows:

1. Remove brackets
2. Divide
3. Multiply
4. Add
5. Subtract

For example, if we have to calculate:

$9 \times (5 + 7 - 4) \div 2$

we perform the following steps.

First, we calculate the part in brackets:

$5 + 7 - 4 = 8$

Then we perform the division:

$8 \div 2 = 4$

Finally, we perform the multiplication:

$9 \times 4 = 36$

For any three numbers, the **distributive law** states:

$a\,(b + c) = ab + ac$

This can be applied whatever the number of terms inside the brackets. For example, if there are two pairs of brackets, we multiply each term in the first pair of brackets by each term in the second pair of brackets, as follows:

$(a + b)\,(c + d)$

$= (a + b)\,(c + d)$

$= ac + ad + bc + bd$

For example:

$$(2 + 5)\,(7 + 3) = (2 \times 7) + (2 \times 3) + (5 \times 7) + (5 \times 3) = 70$$

There are also some important rules about how the signs of positive and negative numbers change, when we are multiplying or adding them together, which are shown in the diagrams below. For example, multiplying a positive number by a negative number always results in a negative number.

Adding

	+	–
+	+	#
–	#	–

Multiplying

	+	–
+	+	–
–	–	+

 # sign of largest number

A mathematical expression containing letters is known as an **algebraic expression**, for example:

$$4\,(x + 5)$$

An algebraic expression takes different values for different values of x – a procedure known as **substitution**. For example:

When x = 1

$$4\,(x + 5) = 4\,(1 + 5)$$
$$= 4 \times 6$$
$$= 24$$

The different parts of an algebraic expression that are separated by + or – signs are known as **terms**. If they contain the same combinations of letters, they are known as **like terms**, such as 2x and –8x, and they can be added and subtracted together. Terms which are **unlike** cannot be added or subtracted together; examples of these would be 2x, $4x^3$ and 7xy.

Functions

As you will have already seen in your studies of Economics, we often want to analyse how one variable affects another, such as how the price of strawberries affects the amount that consumers will buy. We also noted that these relationships are known as functional relationships, because one variable is dependent upon the other, that is, it is a function of the other.

A function is expressed as:

$$y = f\,(x)$$

which means that y is a function of x. Y is the dependent variable and x is the independent variable.

A **linear** function is a relationship which when plotted on a graph produces a straight line.

An example of linear function is:

$$y = 4 + 2x$$

As you can see, this is also an equation.

Functional relationships can be expressed mathematically as equations or graphs.

Equations

A mathematical expression which comprises two parts separated by an equals sign (=) is known as an equation. As we have just seen, a function can be expressed as an equation. If it is a linear function, the equation is known as a **linear** equation.

$y = 4 + 2x$ is therefore a linear equation.

Expressing functions as equations enables us to apply to them the mathematical techniques which are used to manipulate equations. For example, the equation:

$$8x = 3x + 2y + 5$$

is also a function y = f(x).

We can re-arrange the equation so that x becomes the subject of the equation, as follows:

$$8x - 3x = 2y + 5$$

$$5x = 2y + 5$$

$$x = \frac{2y + 5}{5}$$

$$x = \frac{2}{5}y + 1$$

The solution of an equation is the value of the unknown variable which satisfies the equation.

To solve a linear equation, we carry out the following operations:

1. Simplify the expression by multiplying out the brackets.
2. Move all the unknown variables to one side and all the numerical terms to the other side.
3. Perform the arithmetical operations, leaving an expression of the following form:

   $ax = k$

   where: a = constant;

   k = constant;

   $a \neq 0$.
4. Divide both sides by the value in front of x to obtain the solution, expressed as:

   $$x = \frac{k}{a}$$

Here is an example of a linear equation:

$$2(x - 1) = \frac{4x}{3}$$

To solve the equation, we first multiply out the brackets:

$$2x - 2 = \frac{4x}{3}$$

Then we move the unknown variables to one side and the numerical terms to the other side:

$$4x - 2x = 2 \times 3$$

Next we perform the arithmetical operations:

$$2x = 6$$

Finally, we divide both sides by the value in front of x to obtain the solution:

$$x = \frac{6}{2}$$

$$x = 3$$

If there are two unknown variables that we have to find, we need two equations, which are known as **simultaneous equations**. We can then eliminate one of the unknown variables, producing just one equation with one unknown, which we can solve. Then we substitute the known variable into one of the original equations and solve the equation for the other variable.

A set of simultaneous equations is shown below:

$$4x + 3y = 11$$

$$2x + y = 5$$

Let us proceed to solve these.

First, we eliminate x by making the numbers in front of x the same. In this example, we can multiply the first equation by 2 and the second equation by 4:

$$8x + 6y = 22$$

$$8x + 4y = 20$$

Then we subtract one equation from the other to eliminate x:

$$\begin{array}{rrcr} 8x & + 6y & = & 22 \\ -\,8x & + 4y & = & 20 \\ \hline & 2y & = & 2 \\ & y & = & 1 \end{array}$$

Finally, we substitute the value of y into one of the equations:

$$4x + 3y = 11$$

$$4x + 3(1) = 11$$

$$4x + 3 = 11$$

$$4x = 8$$

$$x = 2$$

Graphs

We use graphs in economic analysis when we want to depict a functional relationship.

To graph a function, we carry out the following operations:-

1. Select a series of values for the independent variable x.
2. Substitute the values of x into the function to find the corresponding values of the dependent variable y. Each pair of values represents a point on a graph, known as **co-ordinates** and expressed as (x, y).

3. Plot the co-ordinates on a graph and join them up. If we are depicting a linear function, we will find that we can join the co-ordinates with a straight line.

Let us consider how to graph the following example of a linear function:

$y = 3 + 2x$

First, we select two values for the independent variable (x). These can be any values which are convenient to graph, such as:

x = 1

x = 3

Then we substitute these values of x into the equation to find corresponding values of the dependent variable y. This operation gives us the following pairs of values:

when x = 1, y = 5;

when x = 3, y = 9.

Each pair of values are co-ordinates:

(1, 5)

(3, 9)

Next, we graph these co-ordinates and join them up, as shown below.

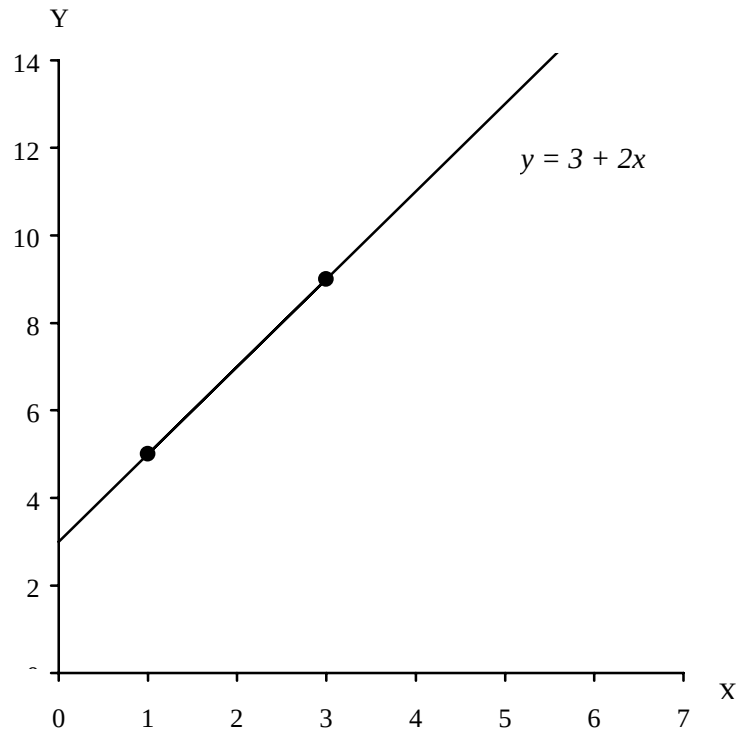


Figure 17.1: Graph of y = 3 + 2x

The graph of a linear function is always a straight line. The general form of a linear function is:

$$y = a + bx$$

where: a = the point where the line crosses the vertical axis;

b = the slope of the line.

To draw a straight line, we only need two points, (x_1, y_1) and (x_2, y_2).

For any two points, we can obtain the gradient of a straight line by using the following formula:

$$b = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\text{difference in y coordinates}}{\text{difference in x coordinates}}$$

B. USING LINEAR EQUATIONS TO REPRESENT DEMAND AND SUPPLY FUNCTIONS

The demand function

You are already familiar with the concept of the demand curve, which shows how much of a product will be demanded by consumers at a range of possible prices. The demand curve is obtained by taking data on the amount demanded at different prices, known as the demand schedule, and plotting the resulting values on a graph. An example of a demand schedule is shown in Table 17.1 below.

Table 17.1

Price	Quantity
1	20
2	15
3	10
4	5
5	0

The demand schedule is shown in graphical form, that is, in the form of a demand curve, in Figure 17.2 below. (Confusingly, a depiction of a demand or supply schedule is always known as a "curve", even when it is a straight line!)

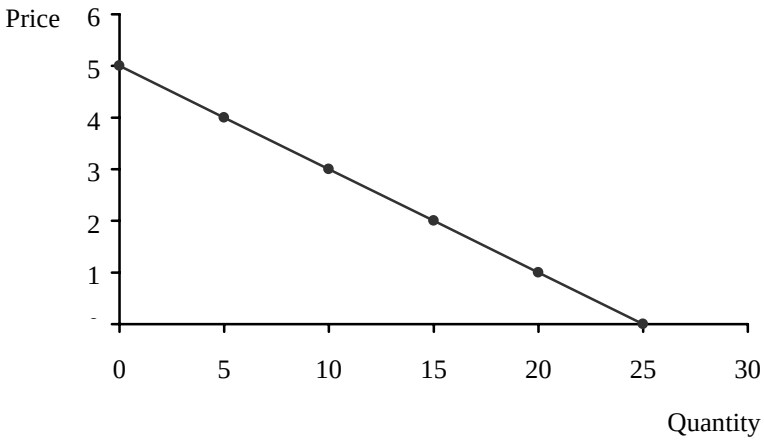


Figure 17.2

Because the demand curve is a straight line, we can express it in the form of a linear equation. The line has a negative slope and therefore it is expressed as:

$$q = a - bp$$

where: q = quantity demanded;

p = price;

and a and b are constants.

For example, the demand function shown above can be expressed as:

$$q = 25 - 5p$$

Sometimes, the equation for the demand function may be given, and from this, we can determine the quantity that will be demanded at any particular price. For example, if the demand function is:

$$q = 50 - 2p$$

we can calculate the values of q at a range of different values of p. These are shown in Table 17.2 below.

Table 17.2

Price	Quantity
5	40
10	30
15	20
20	10
25	0

The equation can also be written to express price as a function of quantity, so we can use it to determine the price at which particular quantities will be demanded. In the example above, the equation would be re-written as:

$$p = 25 - 0.5q$$

The determination of equilibrium price and quantity

Supply functions can also be expressed as equations, in the same way as demand functions. As we saw in Module …, in a competitive market, we assume that price is one of the most important influences on the quantity supplied: as the price of a product increases, the quantity supplied will increase. We can therefore state that supply is a function of price and we can express the supply function mathematically as:

$$q = a + bp$$

where: q = quantity supplied;

p = price;

and a and b are constants.

Note the + sign in the equation, because the supply curve has a positive slope.

If we consider the equations for the demand and supply functions together, we can determine the equilibrium price and quantity. The two equations contain like terms, so we can treat them as simultaneous equations and proceed to solve them for the unknown variables, p and q. For example, let us consider the following demand and supply functions:

Demand function:

$$q^d = 50 - 2p$$

Supply function:

$$q^s = 2 + 4p$$

The equilibrium condition is that quantity demanded is equal to quantity supplied, which can be expressed as:

$$q^d = q^s$$

Considering these as simultaneous equations, we can proceed to solve them, as follows:

$$50 - 2p = 2 + 4p$$

$$4p + 2p = 50 - 2$$

$$6p = 48$$

$$p = 8$$

The equilibrium price is therefore 8.

The equilibrium quantity can then be determined by substituting p =8 into one of the equations. Let us take the equation of the demand function as an example:

$$q^d = 50 - 2p$$

$$q^d = 50 - (2 \times 8) = 50 - 16 = 34$$

The equilibrium quantity is therefore 34.

We can check this by performing the same operations on the equation of the supply function:

$$q^s = 2 + 4p$$

$$q^s = 2 + (4 \times 8) = 2 + 32 = 34$$

The operations which we have performed can be expressed algebraically as follows:

$q^d = a + bp$, where b is less than 0

$q^s = c + dp$, where d is greater than 0

$q^d = q^s$

To obtain the equilibrium price:

$$a + bp = c + dp$$

which can be re-arranged to give:

$$bp - dp = c - a$$

$$p = \frac{c-a}{b-d}$$

The equilibrium quantity can now be obtained from either the demand or supply equation. Using the demand equation, we have:

$$q^d = a + b\left(\frac{c-a}{b-d}\right)$$

$$= \frac{a(b-d)+b(c-a)}{b-d}$$

$$= \frac{ab-ad+bc-ba}{b-d}$$

$$= \frac{bc-ad}{b-d}$$

Shifts in the demand and supply functions

Changes in the demand for a product are brought about by a number of different influences. It is assumed that the price of the product is the principal influence, but there are also others, such as the prices of other products (complementary goods and substitutes), income, tastes and expectations, which were explored in your study of economics. We also noted that because a normal graph can only depict the relationship between two variables – in this case, quantity and price – a change in the quantity demanded which is caused by a change in one of the other influences has be shown graphically by a shift in the whole demand curve.

In the same way that the demand function can be expressed mathematically through an equation or a graph, shifts in the demand function can also be expressed through equations or graphs. Let us consider again the following demand and supply functions:

Demand function

$$q^d = 50 - 2p$$

Supply function

$$q^s = 2 + 4p$$

The demand schedule for a range of possible prices can be calculated from the equation, as shown in Table 17.3 below.

Table 17.3

Price	Quantity
1	48
2	46
3	44
4	42
5	40
6	38
7	36
8	34
9	32
10	30

Let us suppose that there is then a change in consumer tastes away from the product. As a result, the quantity demanded at any particular price will be less than before. The new equation for demand is:

$q^d = 44 - 2p$

The new demand schedule is shown in Table 17.4 below.

Table 17.4

Price	Quantity
1	42
2	40
3	38
4	36
5	34
6	32
7	30
8	28
9	26
10	24

The old and the new demand schedules are shown in graphical form in Figure 17.3 below. As you can see, the demand curve has shifted to the left, showing that at each particular price, demand for the product will be less. We can read the new equilibrium price and quantity off the graph – the new equilibrium price is 7 and the new equilibrium quantity is 30.

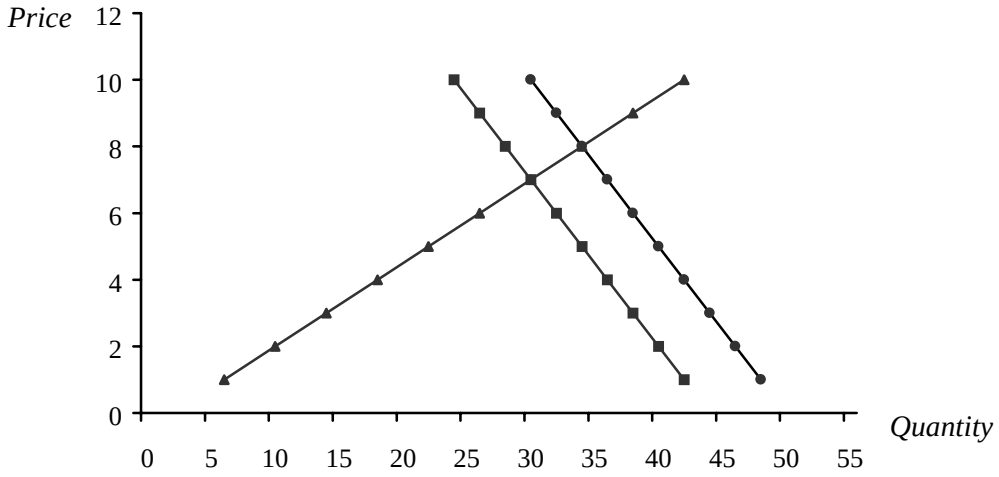


Figure 17.3

We can also use equations to find the new equilibrium price and quantity, in exactly the same way as we did to find the old equilibrium price and quantity.

The new demand function is:

$q^d = 44 - 2p$

The supply function is:

$q^s = 2 + 4p$

The equilibrium condition is :

$q^d = q^s$

Considering these as simultaneous equations, we can proceed to solve them, as follows.

$44 - 2p = 2 + 4p$

$4p + 2p = 44 - 2$

$6p = 42$

$p = 7$

The new equilibrium price is therefore 7.

The equilibrium quantity can then be determined by substituting p = 7 into one of the equations. Let us take the equation of the demand function as an example:

$q^d = 44 - 2p$

$q^d = 44 - (2 \times 7) = 44 - 14 = 30$

The new equilibrium quantity is therefore 30.

If we want to analyse the effects of shifts in the demand curve to the right, or the effects of shifts in the supply curve, we can use exactly the same method.

C. PROBLEMS IN ESTIMATING THE DEMAND AND SUPPLY FUNCTIONS

In economics, we assume that the demand and supply functions for any product can be expressed mathematically by means of equations, and from this information, we can calculate the market equilibrium price and quantity. But in the real world, it is seldom possible to estimate accurately demand and supply functions. Why is this?

- If we gather data about the sales (or production) of a particular product, showing how much is bought (or sold) at different prices, that data will have been collected over a period of time. We therefore cannot be sure that all the other influencing factors, apart from price, will have remained unchanged. If any of these factors had changed, then the data would not represent a demand (or supply) curve. In order to address this problem, we have to use sophisticated econometric techniques, such as multiple regression, to enable us to estimate the influences of different variables.
- We have seen that demand and supply of a product depend on its price, but also that their interaction determines price. If we obtain data about different combinations of price and quantity of a product, we therefore cannot be sure exactly what is happening and whether we are observing demand or supply. The difficulty of identifying the demand curve separately from the supply curve is known as the **identification problem**. Furthermore, if the demand curve shifts, we often do not know whether the supply curve has also shifted, and if so, by how much.

Even if these difficulties could be resolved, any data about price and quantity can only provide an estimate of demand (or supply) at the time that it was obtained. It does not necessarily follow that the same relationship between price and quantity would hold in the future.

D. DISEQUILIBRIUM ANALYSIS

As we have seen, the concepts of demand and supply are central in the study of economics and from our analysis of the demand and supply for individual products, we can develop an equilibrium model of the economy as a whole. Much economic analysis is concerned with comparing one equilibrium state with another, known as **comparative static equilibrium analysis**.

In equilibrium analysis, it is assumed that if the factors which influence demand or supply change, supply and demand may temporarily go into disequilibrium, but equilibrium will soon be restored. Let us consider, for example, a situation where firms are willing to supply more of a product than the equilibrium quantity. This is illustrated in Figure 17.4 below.

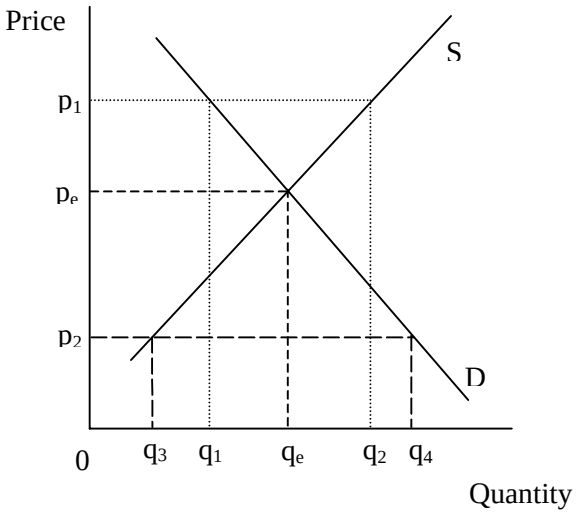


Figure 17.4

At price is $0p_1$, firms are willing to produce an amount $0q_2$, but demand for the product is only $0q_1$. There is therefore excess supply, equal to q_1q_2. Firms may respond by reducing prices to $0p_2$ to clear their stocks, and hence excess supply will exert a downward pressure on prices. At price $0p_2$, the quantity demanded is $0q_4$, which is greater than the amount supplied, so there is now excess demand. We assume that this will cause competition among consumers, firms will be able to sell more than they are producing, and so the price will rise. Excess demand will therefore exert an upward pressure on price. Equilibrium, at price $0p_e$ and quantity q_e, will quickly be found.

But in the real world, disequilibria can often persist in markets for some time. Here are some common examples of disequilibria:

- where there are significant time lags before supply and/or demand responds to price changes, often as a result of imperfect knowledge, for example, consumers may not become aware immediately that prices have changed;
- where the government intervenes in the market, such as rationing or fixing maximum prices;
- where there are significant differences between the quantities producers plan to supply and those which are actually produced, as in agriculture, for example, where unanticipated weather conditions affect production.

One of the main reasons for the persistence of disequilibria is imperfect knowledge among producers and consumers. **Search theory** suggests that it is therefore necessary for them to search for relevant price and quantity information. For example, if different firms are seeking information to help them to set prices appropriately, consumers will find that there is a range of prices for the same product. Once consumers have searched for information about the different prices available, those firms selling at the highest prices will experience difficulty selling their products and those selling at the lowest prices will realise that they could charge more; prices will therefore tend to move towards a middle point. Search theory indicates that an equilibrium price can only be determined if perfect information is available. In the real world, perfect information is not usually available and therefore such a price is not likely to be achieved.

Study Unit 18

Breakeven Analysis

Contents ***Page***

A. AN INTRODUCTION TO COSTS

When we use a mathematical model to solve a financial problem there is usually a function (or functions) which include **costs**. It is important that you understand the nature of costs, and how this affects your model formulation and interpretation. The subject of cost accounting is a large one and here we are only concerned with the broad principles involved in identifying particular types of cost.

Structure of Costs

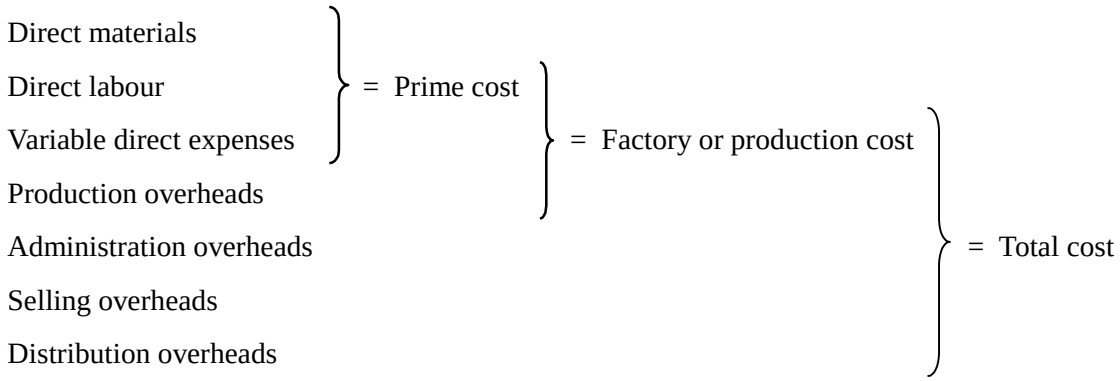
The word "cost" may be used to describe expenditure, or the act of ascertaining the amount of expenditure. The two definitions of cost are:

(a) The amount of expenditure (actual or notional) incurred on, or attributable to, a specified thing or activity.

(b) To ascertain the cost of a specified thing or activity.

Note: Costs can rarely stand on their own. They should always be qualified as to their nature or limitations (e.g. historical, variable) and related to a particular thing (e.g. a given quantity or unit of goods made or services performed).

Total cost is built up of the following constituent elements:



The terminology was originally determined by factory or manufacturing environments, but is equally applicable to a service or computing business. Production overheads can apply in a service industry, as the production of a service incurs overhead expenditure.

Direct and Indirect Costs

Direct costs are any expense which can be **wholly associated** with a particular product or service. This may be broken down into a number of components.

- **Direct Materials**

  These are materials **entering into and becoming constituent elements of a product or saleable service** – the metal used to make a car is a direct material, but the oil used to lubricate the production machinery is an indirect material (part of the manufacturing overheads). Cost of carriage inwards is usually added to material cost.

- **Direct Labour Cost**

  This is the **cost of remuneration for employees' efforts and skills applied directly to a product or saleable service**. The wages of workers on the production line are, therefore,

direct wages, but the wages of the foreman or supervisor are indirect wages. The cost of any idle time of the productive workers is not a direct wages cost.

- **Direct Expenses**

  These are **costs, other than materials or labour, which can be identified in a specific product or saleable service**. Examples are buying special tools for one particular production order, the cost of special designs, royalties payable, the cost of contract computer programs.

By contrast, indirect costs are those items of material, wages or expense which, because of their general nature, cannot be charged direct to a particular job or process. They are often described as **overheads** and are usually classified by reference to the activity from which they derive (administration, distribution, premises, etc.).. Such costs have to be spread – or **apportioned** – in some way over the **various** jobs or processes to which they relate.

Fixed, Variable and Semi-Variable Costs

An alternative classification is to consider the nature of costs in relation to the volume of activity involved on producing the goods or services.

- **Fixed Costs**

  A fixed cost is one which accrues in relation to the passage of time and which, within certain output and turnover limits, tends to be **unaffected by fluctuations in the level of activity**. Examples are rent, Council tax, salary of the production manager. Any expense classified as fixed is fixed only for a certain time, and only within certain levels of production. For instance, the local authority can increase Council tax once a year or once every few years, so clearly Council tax is not fixed for ever. However, within the year they are fixed regardless of the level of production at the factory. If, however, production increased so greatly that it was necessary to acquire a new factory, clearly there would be another lot of Council tax to pay.

- **Variable Costs**

  These are costs which tend to **follow** (in the short term) **the level of activity**. Direct costs are by their nature variable. Consider a selling expense such as travellers' commission. If the organisation makes no sales, no payment or expense will arise, but as sales rise the cost of commission will increase according to the **level of sales** achieved. Further examples of variable overheads are: lubrication of machinery, repairs and maintenance of machinery, consumable stores used in the factory.

- **Semi-Variable (or Semi-Fixed) Cost**

  Between these two extremes, one of which reacts in complete sympathy with activity, while the other is not affected by activity at all, there is another type of overhead which is partly fixed and partly variable. It is known as a semi-fixed or semi-variable cost. An example is the charge for electricity, which consists of a **standing charge** per quarter (the fixed element) and a charge **per unit** of usage (the variable element). Any semi-variable cost can be separated into fixed and variable components.

When the total variable cost of a number of products is deducted from the total sales revenue, the amount that is left over is called the **contribution**. Since fixed costs have not yet been taken into account, this contribution has to cover fixed costs and then any amount remaining is profit. (That is why it is called the contribution – it contributes to fixed costs and then profit.)

We can also talk about the **contribution per unit** of a product: this is simply the selling price minus the variable cost.

We can write this symbolically as:

S – V = F + P

(Sales revenue – Variable cost = Fixed cost + Profit).

(This is the basic equation of **marginal costing**. You will probably be aware of this approach to cost accounting from you studies in Finance and Accounting and we shall not go into the details here.)

B. BREAKEVEN ANALYSIS

For any business, there is a certain level of sales at which there is neither a profit nor a loss. Total income and total costs are equal. This point is known as the **breakeven point**. It is easy to calculate, and can also be found by drawing a graph called a **breakeven chart**.

Calculation of Breakeven Point

Example

The organising committee of a Christmas party have set the selling price at £21 per ticket. They have agreed with a firm of caterers that a buffet would be supplied at a cost of £13.50 per person. The other main items of expense to be considered are the costs of the premises and discotheque, which will amount to £200 and £250 respectively. The variable cost in this example is the cost of catering, and the fixed costs are the expenditure for the premises and discotheque.

Answer

The first step in the calculation is to establish the amount of contribution per ticket:

	£
Price of ticket (sales value)	21.00
less Catering cost (marginal cost)	13.50
Contribution per ticket	7.50

Now that this has been established, we can evaluate the fixed costs involved. The total fixed costs are:

	£
Premises hire	200
Discotheque	250
Total fixed expenses	450

The organisers know that for each ticket they sell, they will obtain a contribution of £7.50 towards the fixed costs of £450. Clearly it is necessary only to divide £450 by £7.50 to establish the number of contributions which are needed to break even on the function. The breakeven point is therefore 60 - i.e. if 60 tickets are sold there will be neither a profit nor a loss on the function. Any tickets sold in excess of 60 will provide a profit of £7.50 each.

Formulae

The general formula for finding the breakeven point (BEP) is:

$$\text{BEP} = \frac{\text{Fixed costs}}{\text{Contribution per unit}}$$

If the breakeven point (BEP) is required in terms of sales **revenue**, rather than sales **volume**, the formula simply has to be multiplied by selling price per unit, i.e:

$$\text{BEP (sales revenue)} = \frac{\text{Fixed costs}}{\text{Contribution per unit}} \times \text{Selling price per unit}$$

In our example about the party, the breakeven point in revenue would be 60 × £21 = £1,260. The committee would know that they had broken even when they had £1,260 in the kitty.

Suppose the committee were organising the party in order to raise money for charity, and they had decided in advance that the function would be cancelled unless at least £300 profit would be made. They would obviously want to know how many tickets they would have to sell to achieve this target.

Now, the £7.50 contribution from each ticket has to cover not only the fixed costs of £450, but also the desired profit of £300, making a total of £750. Clearly they will have to sell 100 tickets (£750 ÷ £7.50).

To state this in general terms:

$$\text{Volume of sales needed to achieve a given profit} = \frac{\text{Fixed costs + desired profit}}{\text{Contribution per unit}}$$

Suppose the committee actually sold 110 tickets. Then they have sold 50 more than the number needed to break even. We say they have a **margin of safety** of 50 units, or of £1,050 (50 × £21) – i.e.:

Margin of safety = Sales achieved – Sales needed to break even

It may be expressed in terms of sales volume or sales revenue.

Margin of safety is very often expressed in percentage terms:

$$\frac{\text{Sales achieved} - \text{Sales needed to break even}}{\text{Sales achieved}} \times 100\%$$

i.e. the party committee have a percentage margin of safety of $\frac{50}{110} \times 100\% = 45\%$.

The significance of the margin of safety is that it indicates the amount by which sales could fall before a firm would cease to make a profit. If a firm expects to sell 2,000 units, and calculates that this would give it a margin of safety of 10%, then it will still make a profit if its sales are at least 1,800 units (2,000 – 10% of 2,000), but if its forecasts are more than 10% out, then it will make a loss.

The profit for a given level of output is given by the formula:

(Output × Contribution per unit) – Fixed costs

It should not be necessary for you to memorise this formula, since when you have understood the basic principles of marginal costing you should be able to work out the profit from first principles.

Example

Using the data from the first example, what would the profit be if sales were:

(a) 200 tickets?

(b) £2,100 worth of tickets?

Answer

(a) We already know that the contribution per ticket is £7.50.

Therefore, if they sell 200 tickets, total contribution is 200 × £7.50 = £1,500.

Out of this, the fixed costs of £450 must be covered; anything remaining is profit.

Therefore profit = £1,050. (Check: 200 tickets is 140 more than the number needed to break even. The first 60 tickets sold cover the fixed costs; the remaining 140 show a profit of £7.50 per unit. Therefore profit = 140 × £7.50 = £1,050, as before.)

(b) £2,100 worth of tickets is 100 tickets since they are £21 each.

	£
Total contribution on 100 tickets =	750
less Fixed costs	450
Profit	300

C. BREAKEVEN CHARTS

A number of types of breakeven chart are in use. We will look at the two most common types:

- Cost/volume charts
- Profit/volume charts

Information Required

(a) Sales Revenue

When we are drawing a breakeven chart for a single product, it is a simple matter to calculate the total sales revenue which would be received at various outputs.

As an example, take the following figures:

Table 18.1

Output *(Units)*	**Sales Revenue** *(£)*
0	0
2,500	10,000
5,000	20,000
7,500	30,000
10,000	40,000

(b) Fixed Costs

We must establish which elements of cost are fixed in nature. The fixed element of any semi-variable costs must also be taken into account.

We will assume that the fixed expenses total £8,000.

(c) Variable Costs

The variable elements of cost must be assessed at varying levels of output.

Table 18.2

Output *(Units)*	**Variable costs** *(£)*
0	0
2,500	5,000
5,000	10,000
7,500	15,000
10,000	20,000

Cost/Volume Chart

The graph is drawn with level of output (or sales value) represented along the horizontal axis and costs/revenues up the vertical axis. The following are the stages in the construction of the graph:

(a) Plot the **sales line** from the above figures.

(b) Plot the **fixed expenses line**. This line will be parallel to the horizontal axis.

(c) Plot the **total expenses line**. This is done by adding the fixed expense of £8,000 to each of the variable costs above.

(d) The **breakeven point** is represented by the meeting of the sales revenue line and the total cost line. If a vertical line is drawn from this point to meet the horizontal axis, the breakeven point in terms of units of output will be found.

The graph is illustrated in Figure 18.1, a typical cost/volume breakeven chart.

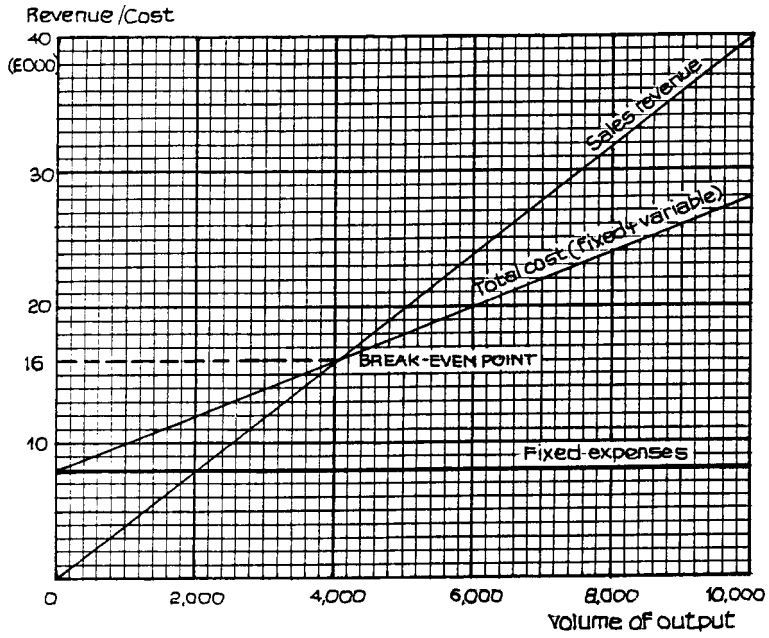


Figure 18.1

Note that although we have information available for four levels of output besides zero, one level is sufficient to draw the chart, provided we can assume that sales and costs will lie on straight lines. We can plot the single revenue point and join it to the origin (the point where there is no output and therefore no revenue). We can plot the single cost point and join it to the point where output is zero and total cost = fixed cost.

In this case, the breakeven point is at 4,000 units, or a revenue of £16,000 (sales are at £4 per unit).

This can be checked by calculation:

Sales revenue	=	£4 per unit
Variable costs	=	£2 per unit
Thus, contribution	=	£2 per unit
Fixed costs	=	£8,000
Breakeven point	=	Fixed costs ÷ Contribution per unit
	=	4,000 units

The relationship between output and profit or loss is shown in Figure 18.2, a typical cost/volume chart.

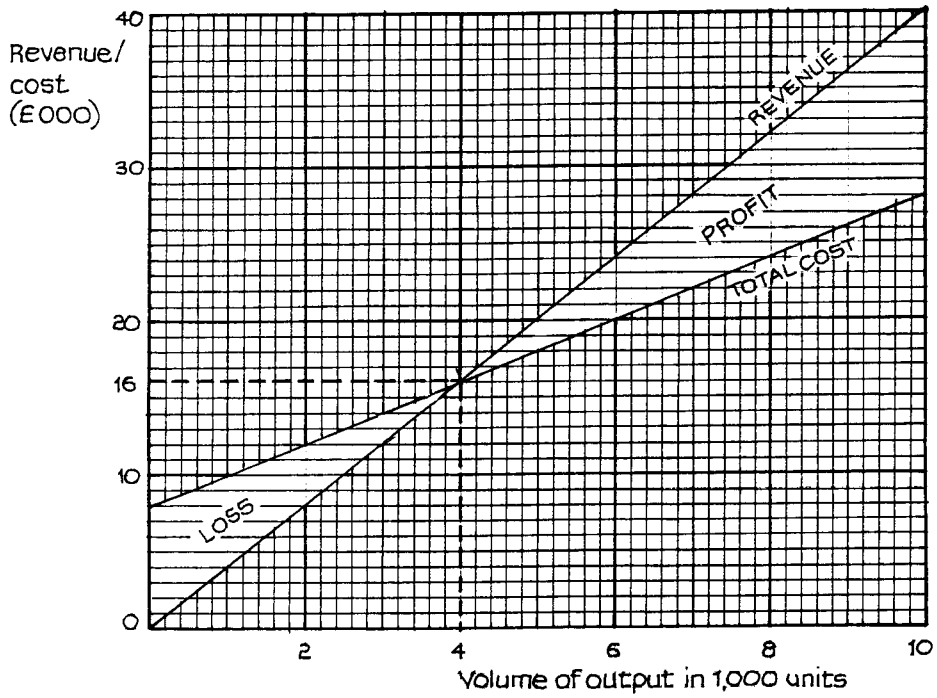


Figure 18.2

Profit/Volume Chart

With this chart the **profit line** is drawn, instead of the revenue and cost lines. It does not convey quite so much information, but does emphasise the areas of loss or profit compared with volume.

The contribution line is linear, so we need only two plotting points again.

When the volume of output is zero, a loss is made which is equal to fixed costs. This may be one of our plotting points. The other plotting point is calculated at the high end of the output range:

i.e. When output = 10,000 units

Revenue = £40,000

Total costs = £(8,000 + 20,000) = 28,000

Profit = £(40,000 – 28,000)

= 12,000 (see Figure 18.3)

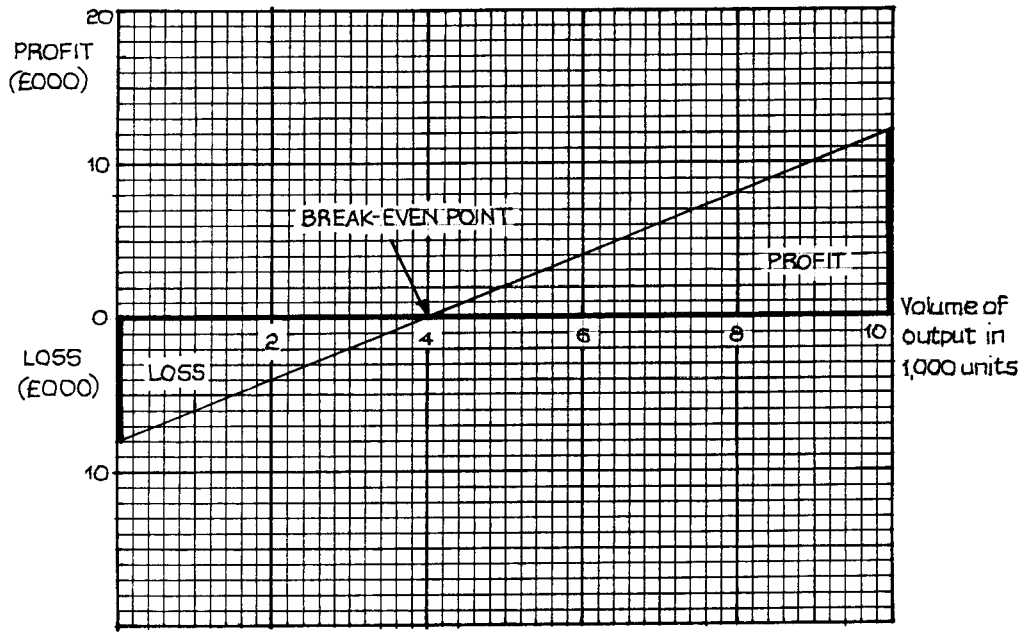


Figure 18.3

When drawing a breakeven chart to answer an exam question, it is normal to draw a cost/volume chart unless otherwise requested in the question. The cost/volume chart is the more common type, and does give more detail.

Margin of Safety

If management set a level of budgeted sales, they are usually very interested in the difference between the budgeted sales and the breakeven point. At any level between these two points, some level of profit will be made. This range is called the **margin of safety** (see Figure 18.4), where the level of activity is budgeted (planned) at 8,000 units.

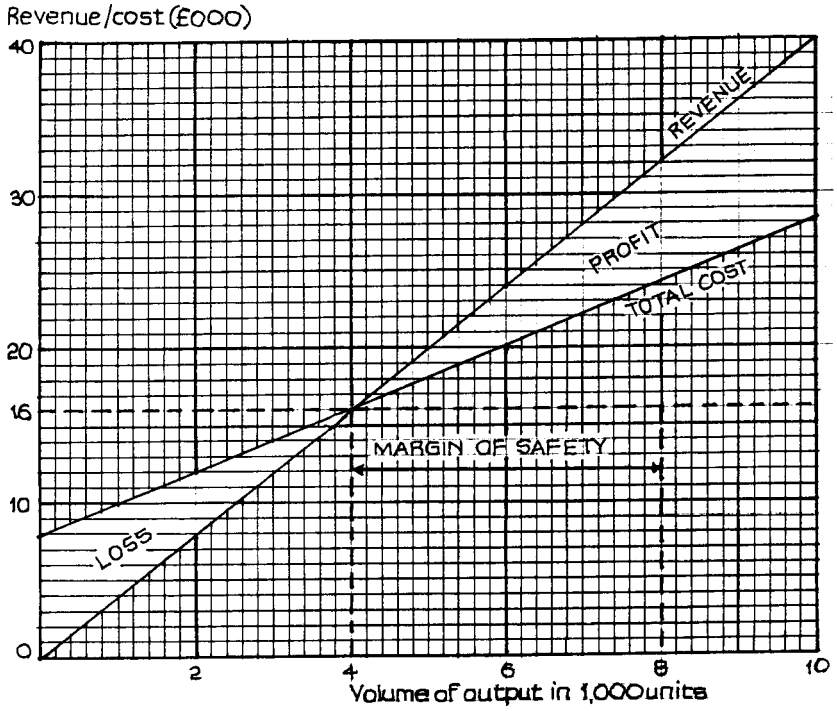


Figure 18.4

Assumptions and Limitations of Breakeven Charts

- It is difficult to draw up and interpret a breakeven chart for more than one product.
- Breakeven charts are accurate only within fairly narrow levels of output. This is because if there was a substantial change in the level of output, the proportion of fixed costs could change.
- Even with only one product, the income line may not be straight. A straight line implies that the manufacturer can sell any volume he likes at the same price. This may well be untrue: if he wishes to sell more units he may have to reduce the price. Whether this increases or decreases his total income depends on the elasticity of demand for the product. The sales line may therefore curve upwards or downwards, but in practice is unlikely to be straight.
- Similarly, we have assumed that variable costs have a straight line relationship with level of output - i.e. variable costs vary directly with output. This might not be true. For instance, the effect of diminishing returns might cause variable costs to increase beyond a certain level of output.
- Breakeven charts hold good only for a limited time.

Nevertheless, within these limitations a breakeven chart can be a very useful tool. Managers who are not well-versed in accountancy will probably find it easier to understand a breakeven chart than a calculation showing the breakeven point.

D. THE ALGEBRAIC REPRESENTATION OF BREAKEVEN ANALYSIS

Using linear equations to represent cost and revenue functions

We have already seen how equations can be used to represent demand and supply functions and hence to determine equilibrium price and quantity. Similarly, equations can be used to represent cost and revenue functions and to calculate profit and output.

Let us consider a simple example. Table 18.1 shows the sales revenue which is yielded at different levels of output – it is a **revenue schedule**. The schedule is depicted graphically in Figure 18.1, where we can see that it takes the form of a straight line. We already know that a relationship which, when plotted on a graph, produces a straight line is a linear function and hence can be described by means of a linear equation. It therefore follows that the revenue schedule we are considering is a linear function and can be described by a linear equation.

We know that the general form of a linear function is:

$$y = a + bx$$

where: a = the point where the line crosses the vertical axis;

b = the slope of the line.

We also know that for any two points, we can obtain the gradient of a straight line by using the following formula:

$$b = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\text{difference in y coordinates}}{\text{difference in x coordinates}}$$

From Figure 18.1, we can see that the line crosses the vertical axis at 0.

To find the gradient, we perform the following calculation:

$$\frac{20{,}000 - 10{,}000}{5{,}000 - 2{,}500} = \frac{10{,}000}{2{,}500} = 4$$

We can therefore state the equation for revenue (R) as follows:

$$R = 4x$$

where: x = output.

This is known as the **revenue function**.

We can also perform a similar calculation to find the equation of the total cost line – the **cost function** – depicted in Figure 18.1. Remember that we need first to sum fixed costs (set at £8,000) and variable costs (shown in Table 18.2) to obtain values for total costs; then we can carry out the calculation as before.

The breakeven point

We have already seen that the breakeven point corresponds to the volume of output at which total revenue equals total cost. At this point, profit is zero; beyond this point, any increase in output will yield a profit.

In algebraic terms, profit can be expressed as:

$$\mu = Pq - (F + Vq)$$

where: μ = profit;

P = unit selling price;

q = sales volume in units;

F = total fixed costs;

V = unit variable cost.

The breakeven point at which total revenue equals total cost and profit equals zero can be expressed as:

$$Pq_b - (F + Vq_b) = 0$$

where: q_b = breakeven volume.

We can re-arrange the equation to express breakeven volume as:

$$q_b = \frac{F}{P - V}$$

where P – V is the contribution per unit. Therefore the breakeven point equals total fixed costs (F) divided by the contribution per unit (P – V). To convert q_b into breakeven sales (Y), we multiply both sides of the q_b formula by P, as follows:

$$Y = Pq_b = \frac{PF}{P - V}$$

This can also be expressed as:

$$Y = \frac{F}{1 - V/P}$$

where: 1 – V/P = contribution ratio.

This formula gives us breakeven sales.

Let us consider an example of a company that produces a product which sells for 50 pence per unit. Total fixed costs amount to £10,000 and the variable cost per unit is 30 pence.

The unit contribution (or the excess of unit sales price over unit variable cost) is:

$$P - V = 0.50 - 0.30 = 0.20$$

The breakeven point is:

$$q_b = \frac{10{,}000}{0.20} = 50{,}000 \text{ units}$$

The contribution ratio is:

$$1 - V/P = 1 - \frac{0.30}{0.50} = 40\%$$

Breakeven sales is:

$$Y = \frac{10{,}000}{0.40} = £25{,}000$$

which can also be expressed as:

$$Y = Pq_b = 0.50 \times 50{,}000 \text{ units} = £25{,}000$$

Changes in the cost and revenue functions

We can use the breakeven formulae above to analyse the effect of changes in the cost and revenue functions – that is, in the parameters and variables, such as the unit selling price, variable costs and fixed costs. Let us consider each of these in turn.

A reduction in the unit selling price will decrease the contribution and hence increase the breakeven volume. If we assume that the unit price is reduced from 50 pence to 40 pence, while all the other variables remain unchanged, we can find the new breakeven point as follows:

$$q_b = \frac{10{,}000}{0.40 - 0.30} = 100{,}000 \text{ units}$$

and $Y = 100{,}000 \times 0.40 = £40{,}000$

or $$Y = \frac{10{,}000}{1 - (0.30/0.40)} = £40{,}000$$

An increase in the unit variable cost will decrease the unit contribution and increase the breakeven volume. If we assume that the price of raw materials increases by 10 pence per unit, while the other variables remain unchanged, we can find the new breakeven point as follows:

$$q_b = \frac{10{,}000}{0.50 - 0.40} = 100{,}000 \text{ units}$$

and $Y = 100{,}000 \times 0.50 = £50{,}000$

or $$Y = \frac{10{,}000}{1 - (0.40/0.50)} = £50{,}000$$

Similarly, a decrease in unit variable cost will decrease the breakeven volume.

An increase in total fixed costs will increase breakeven volume, while a decrease in total fixed costs will decrease breakeven volume. If we assume that fixed costs increase by £2,000, while the other variables remain unchanged, we can find the new breakeven point as follows:

$$q_b = \frac{10{,}000 + 2{,}000}{0.50 - 0.30} = 60{,}000 \text{ units}$$

and $Y = 60{,}000 \times 0.50 = £30{,}000$

or $$Y = \frac{12{,}000}{1 - (0.30/0.50)} = £30{,}000$$

Calculating profit at different output levels

We have already seen that profit at breakeven point equals zero. Therefore, the profit for any volume of output greater than breakeven equals the profit generated by the additional output beyond the breakeven volume. We can express profit for any given sales volume (q_1) as:

$$(q_1 - q_b) \times (P - V)$$

In our example, the breakeven volume is 50,000 units. Let us assume that we now want to find the profit generated by sales of 70,000 units. Using the formula above:

$$(70{,}000 - 50{,}000) \times (0.50 - 0.30) = £4{,}000$$

The profit generated by sales of 70,000 units is therefore £4,000.