Edition

3

TEXT BOOK ON

Information and Data Analysis

Written and compiled by

Martin Ziarati
Basak Akdemir

2015

Based on the work and contributions from

Professor Dr Reza Ziarati

Information and Data Analysis

A TEXT BOOK ON

Information And Data Analysis

Publisher: Centre for Factories of the Future,
University of Warwick Science Park,
Barclays Venture Centre,
Sir Williams Lyons Rd,
Coventry CV4 7EZ

Year of Publication of Professor Ziarati's notes: 2003, 1st and 2010, 2nd Edition

Table of Contents

Preface

Why having skills and a good knowledge and understanding of methods to analyse data and information are important?

Engineering and social science students often presume that they can analyse data and information and make correct decisions without the need to consider proven techniques in problem definitions and solutions. Many have studied one or two basic statistical subjects and may not be aware that the subject of data and information analysis is not primarily based on statistical topics but a whole range of subjects and recently developed techniques and tools.

All graduates, but particularly engineering, science and management, handle data and information on a daily basis and in recent years have been using computers and associated software for analysis of a given situation or problem.

The topics included in this textbook have been carefully selected and introduced using a novel approach which introduces the students to a given topic through a real example after presenting the minimum amount of information relating to various concepts, definitions and techniques as well as tools. The students subconsciously become aware of the need for various concepts and the reason for the definitions.

Students often assume that simple concepts and basic definitions are also easy to understand. To this end, basic concepts, terms and statements are introduced in such a way that they test the students for understanding and awaken their curiosity.

Each chapter commences with the principle objectives and these are carefully inter-related with each other both within the chapters and with the objectives of earlier chapters.

The textbook places equal emphasis on practical aspects. The use of this textbook for the delivery of topics presented should take advantage of the popular hardware and software systems, techniques and tools. The tutorials involve the use of well-known management and engineering application software packages.

I would like to thank the authors for their informal approach to learning creating a student-centred means to learning. I found the book interesting and commend the authors for their efforts and contributions.

Professor Dr Reza Ziarati

June 2005

Chapter 1

Introduction to information System

Principle Objectives

- *To introduce the concepts of information and system*
- *To explain why information systems have gained importance*
- *To ascertain the need for information system in an organisation*
- *To describe how information systems are modelled.*

Introduction

Before referring directly to the principle objectives it may be appropriate to we have to look inside an organization, say a company or a firm for instance, and see what they are and what they do and what factors influence their decisions? On a detailed level a company is made up of different departments each responsible for a specific function.

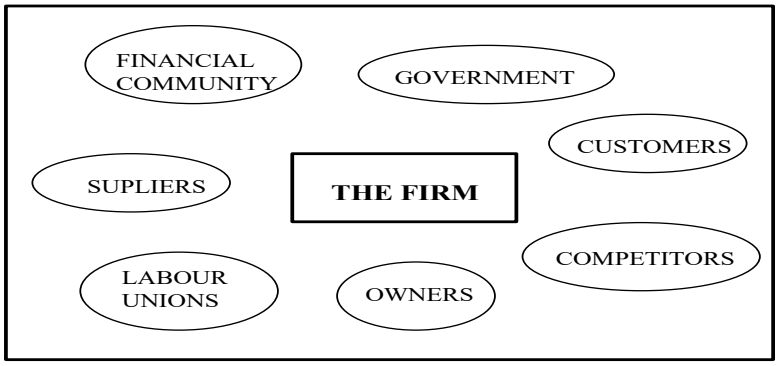
If we look at a company with a holistic view, we can say that a typical firm managers a number of resources for a achieving its objective. These resources can be categorized as:

- Personnel
- Material and Machines
- Utilities, e.g. energy
- Money
- Information

A typical management structure is as shown below. The management has different level of responsibility, control and decision making.

- Executives
- Senior managers
- Middle managers
- Line managers

At the same time when one looks at the world outside a company, that is the company's environment what does one see ?



SLIDE 1-3

All those elements in the figure are groups by whom the firm has to interact externally and internally. As a group has a given interest therefore each group tries to exert pressure and the company needs to develop strategies to deal with these external forces. In dealing with these groups the companies have to develop strategies to deal with the causes of competitive forces, e.g.

- Global markets
- International economical influences
- Worldwide competition
- Time constraints
- Social constraints
- New concepts in business

So companies constantly have to collect, analyse, and make decisions to meet the demand that are placed on them. To do this companies have devised a number of strategies and methods to deal with this. One such strategy is to concentrate on maximising the use of information for gaining competitive advantage.

- So to answer the question one can say that the companies can use information systems to better plan, organize and control their resources.

- They therefore need systems that allow them to collect data, internal and external and to convert such data into useful information

- A better use of information transaction within a firm or even with outside groups like customers and suppliers also gives rise to internal and external efficiencies and other benefits that ultimately translate into a higher profit and a stronger position for the firm in the market place.

What is Management Information System

In its simplest form it is any means of collecting data, storing it and manipulating it. Furthermore information systems are often associated with a strategy (reason) and therefore have a purpose for their existence.

There are two aspects to Information systems, one is Information System(IS) and the other Information Technology (IT). The companies look at their internal/external workings and having evaluated their needs they are the able to state what they require to meet their business objectives. For example, this could be a need for a word processing software, a stock control system, etc. This analytical activity is covered under the heading of Information Systems or IS.

On the other hand there has been explosion in the information processing machines, i.e. computers, networks, software technology, etc. The task of translating the identified systems into a physical reality is the domain of Information Technology or IT.

Therefore:

IS establishes the need for application i.e. a word processing software

and

IT satisfies the need for application i.e. deciding on buying MS WORD for instance, to satisfy the requirement for a word processing software package, and the necessary hardware for it.

So IS and IT can be thought of as demand and supply.

Obviously, information systems deal with information. In a firm data and information continuously flows around like blood in a body allowing giving life to various units to do what they have to do.

Let us now define Information.

What Is Information? How Can We Define Information?

Information is the smaller unit or the common denominator in the scheme of IS.

Information can be defined as data processed for a purpose.

Data

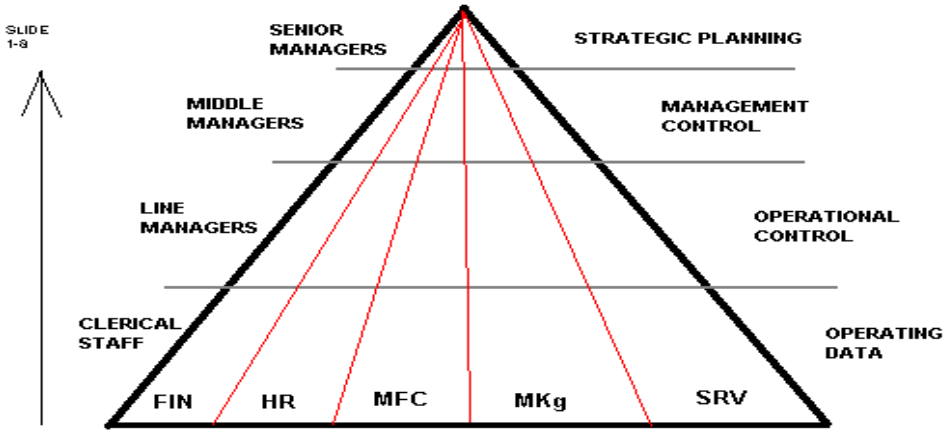
Today companies collect data that is necessary for their basic operational needs. One can find data stored about customers, prices, stock, suppliers, competitors etc. It should be noted that this data on its own is of no significant importance and until it becomes information. For instance, 540 is data but has no significance, but if 540 represents the number of employees in an organization, then it is significant.

To demonstrate the difference between data and information lets assume a company needs to know how much profits it has made. There are many data scattered all over the firm; but only when someone takes the total sales figure and subtracts the cost of sales and expenses that the company would know how much profit it has made. The profit figure is therefore a piece of information that can be used in decision making process, for instance, how to increase sales because the profit is low.

Having processed the data and obtain some information then the only reason for the existence of this information is its purpose, why it is needed. So the value of information is in its use. But use by whom ? The purpose of the information therefore depends as to who uses it and for what purpose

Information Users

The diagram below shows the role of various management levels and main operational unit of a firm. Where Fin stands for Finance, HR for Human Resources, MFC, Manufacturing, MKG Marketing and SRV Service units.



The first users of IS were clerical staff, processing basic data, writing notes and filling forms, etc.. Information systems were then used to assist in applications such as payroll, inventory, billing and so forth. They also provided (or input) much of the operating data, which often was converted to information for use by managers. Operating data are not needed for decision-making but must flow for the firm to operate according to its stated agenda.

Since data was available the next phase was to process it and use the derived information for ((line) management) control purposes. So line managers, that is the supervisors, use information for Operations Control.

The next stage was to further process the data available to use by the middle and senior managers for Management Control and Strategic Planning.

Remember that people outside the organisation also are the beneficiary of IS.

The Management Information System is a collection of IS for the provision of the information required by the Managers. The line managers need information to make decisions where the roles are well known and understood, middle managers: (division heads/regional managers/product directors, etc) need information for management control decision, for instance for comparing the monthly sales against the budget and that information that allows them to take a different course of action. This information is generally internally generated, historical and formal. Senior manages (executive) need information for strategic decision making. These are Long range- often not structured and complex information, e.g. company reorganisation, diversification, etc.

Information Characteristics

Every piece of information that flows in an organisation has characteristics:

- Its function - purpose for its existence
- Its use – the use it is put to by the user
- Its recipient - who it is intended for.

1. In general information is used to reduce uncertainty.
2. In a flight reservation system the use of information is to ensure customers have access to flight departure and arrival timetables. The system is expected to provide fast, accurate and up-to-date information.
3. It could be that information systems might not have (financially) quantifiable benefits, but benefits such as customer satisfaction due to speed of response and accuracy.
4. Internal use of information could be for monitoring for example, the performance of sales personnel.

The characteristics could also include details such as:

- Time: Immediate to long term
- Level of detail: Summary to highly detailed
- Source: External vs. internal
- Degree of Certainty: Uncertain to certain

- ❖ Frequency: Infrequent to frequent

Cost

Collection/manipulations/storage and presentation of data/information has an associated cost. To invest this cost the beneficiary of this information must see and perceive a value since her/his objective is often to maximize profit.

If a farmer plants cotton and in the month of September if there is 40cm of rain then he would make a profit of say 80$ per ton. If he plants corn and the rainfall in September is 100cm he makes 130$ per ton. So how much the farmer is willing to pay for an accurate whether forecast for September if he harvests 1000 tonnage of each crop?

$130 - $80 = $30 per ton x 1000 ton = $30000. This is the value of this information to him.

For another example see Curtis- Business Information System, page 12.

Because costs are involved whenever a company is to invest money on an IS, it must ultimately specify and measure the direct and/or indirect benefits derived as a result of the system introduction in monetary units.

What is System

Before defining the system, let us classify systems and cite some examples:

Physical systems	Central heating system, telephone system, fuel injection systems.
Abstract systems-	Logical system (numbers), philosophical systems
Social systems	Economical systems, Social security systems, Legal systems
Financial systems	Accounting systems

Definition of System

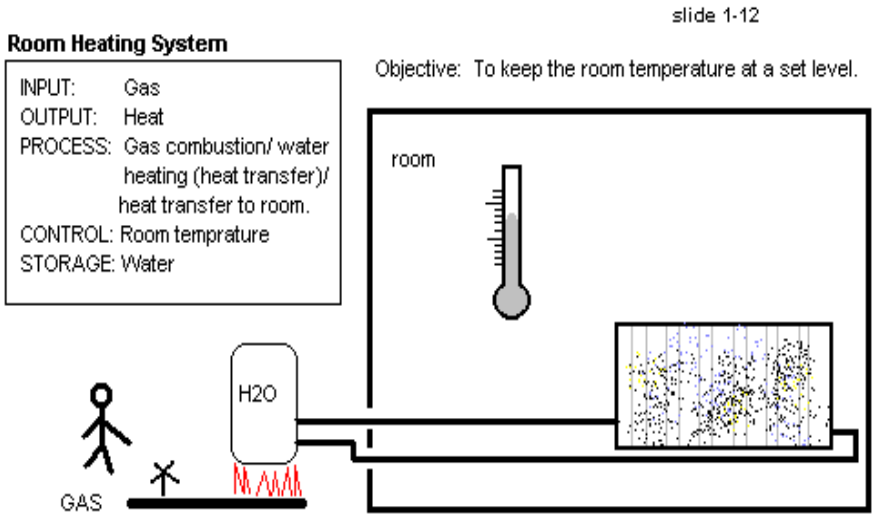
So what do all these diverse activities have in common so that they all are known as a system ?

- They are all made from parts
- Every part is related to every or some other part

Definition: A collection of interrelated parts (or tasks, components, etc.) that together has a purpose.

Why Interrelated? Because if there is a change in one part/task of the system, this change manifest itself as a change in another parts/tasks.

Example of a
system:



So now we are in a position to identify the component of a general system. These are:

Input, Output, Process, Control and Objective.

Why We Need to Study Systems?

The companies form themselves into functional groupings and/or departments and each of these conducts a well-defined process. Each of these departments therefore can be classified as a sub-system. To this end, the

general system theory serves as a useful tool for modelling each and every process in a given organization.

For a manufacturer therefore the input is raw materials, the output is finished good, the transformation is the manufacturing process, the control mechanism is the management, and the objective is to make profit.

In Bus ticket system, the objective is to have a full bus at all costs. The input is the total available seats and the output is the number of sold seats. The control is the 'error' i.e. the number of unsold seats and the process is adjust sales price

Notice how all the objectives are clear and measurable. There is also a need to draw a line around the system under consideration so as to separate it from its surrounding. This line is called the system boundary and its surround is called the system environment.

So a boundary is the line over which the input output flow.

Environment is the condition around the system that could influence (or be influenced) by the system.

One must not confuse the idea of the boundary and environment with geographical locations. Let's consider another example:

The store manager requires a system with the objective of speeding her access into and out of her store. For her therefore the system boundary is the perimeter of her store. Inclusion of sales or marketing systems into her boundary is not of interest to her. But suppose she is linked very closely with her suppliers therefore although her suppliers are physically and geographically displaced from her store, the store manager's boundary will now encapsulate them too.

System Classification

- Open-loop systems – no feedback or apparent control mechanism for adjustments
- Closed-loop systems – the systems where there is at least one feedback mechanism for controlling the output

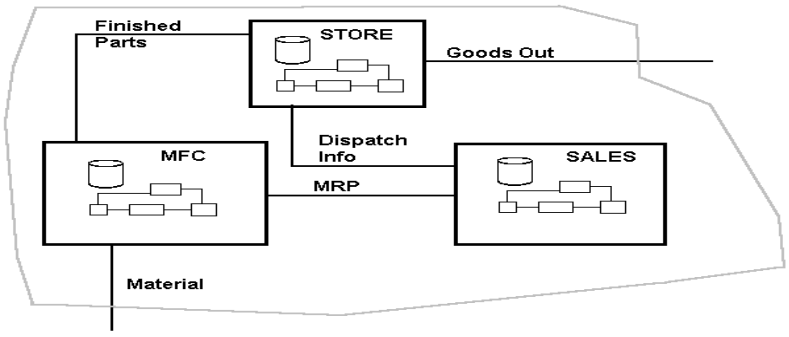
Open-loop example: companies that set on a particular course action and never change their direction.

Closed loop systems like a central heating system or a company that senses something and make the appropriate changes if necessary.

Sub-systems

Of course we can have systems within a system. This gives rise to the concept of sub-systems. As an example for the senior manager of a firm, the store is only a component of a larger system made-up from whole ranges of sub-systems such as, manufacturing, sales, accounts, etc. Thus the system environment for the Senior manager's the market place and clearly the boundary would perhaps encapsulate all the firms activities.

So when systems are composed of other systems that are interrelated by their input and outputs, the constituent systems are called sub-systems.



slide 1-16

Note that each sub-system is a system of its own.

So one can see that a larger system can be broken down (by the process of decomposition) into smaller sub-system, and each sub-system could be decomposed further into its constituent sub-systems. This process can be continued until reaching the basic elements, where the definition of system no longer applies.

Management Information Systems?

So if one incorporate information and system together for the purpose of decision making then a Management Information System (MIS) will emerge.

A system in this context is a framework for the provision of the information required.

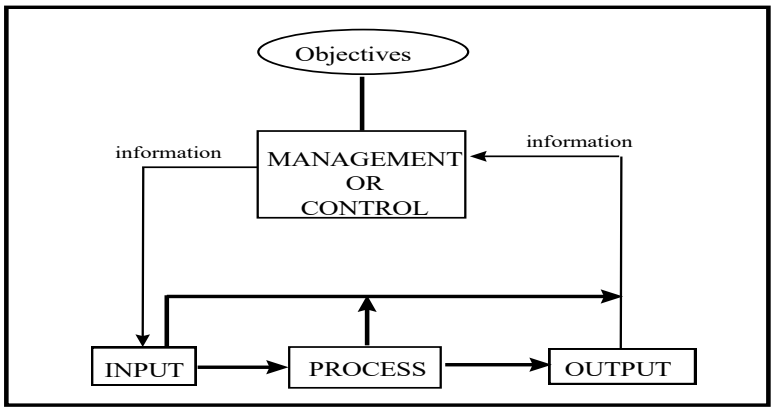
The management generally receives information about the output and having measured or evaluated against a value (benchmark/standard/set value) makes corrective action, which may affect the input. The system described many not necessarily resemble a closed-loop system.

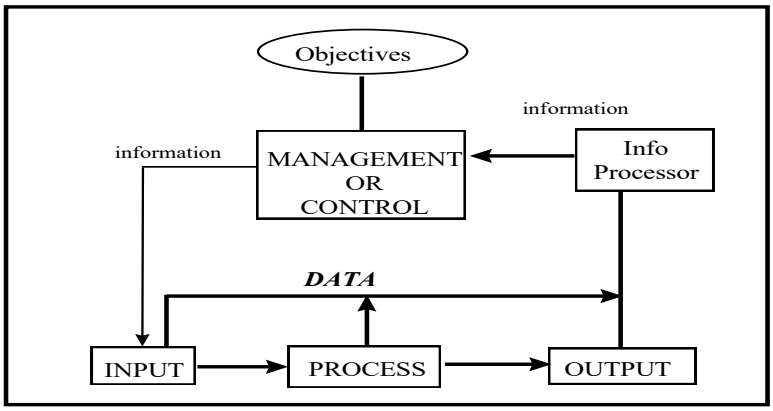
The output information could be a list of monthly sales (management decision), daily quality figures (operation decision).

slide 1-18

Management Information System (MIS)

A MIS is a frame work for the provision of the information required.





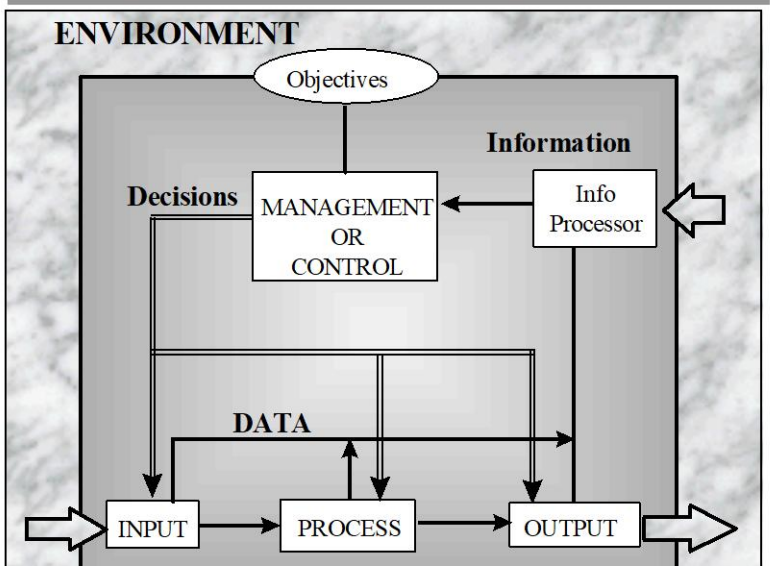
Sometimes the management needs also to collect data from the input and the process. For example, s/he might want to know how well the suppliers are performing or how the process is converting input to output. An MIS that produces a supplier delivery performance, for example, could be required.

Often information is not available directly from the processes in an organisation. In such scenarios then an information processor is needed to produce the required information. An example could be the monthly profit report.

slide 1-19

Management Information System (MIS)

A MIS is a frame work for the provision of the information required.



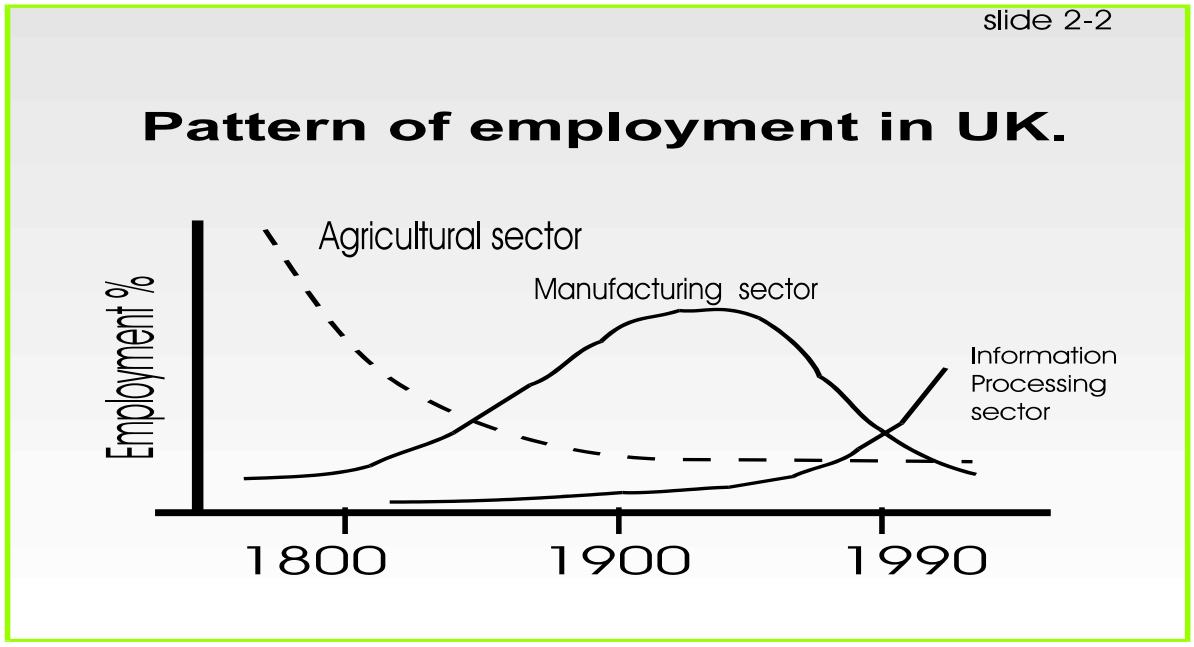
Management collects data (one way or other), having processed it against some objectives he then makes a decision and influences the input or process in the system. The whole system however is also subject to external forces from the environment it finds itself in.

So just like an architect produces a plan or a blue print that a builder can turn into a house, we use system theory to product the blue print of the required management system that an IT specialist can turn into what is names a computerised management information system.

Historical Development of MIS

Let us look at the structural changes in business environment.

The development of industrialization meant that in the early and mid 1900s there has been a shift from employment in the agricultural sector into manufacturing. Since the decline in manufacturing there has been a shift of emphases into service and information sector, such as banking, insurance, government agencies, etc.



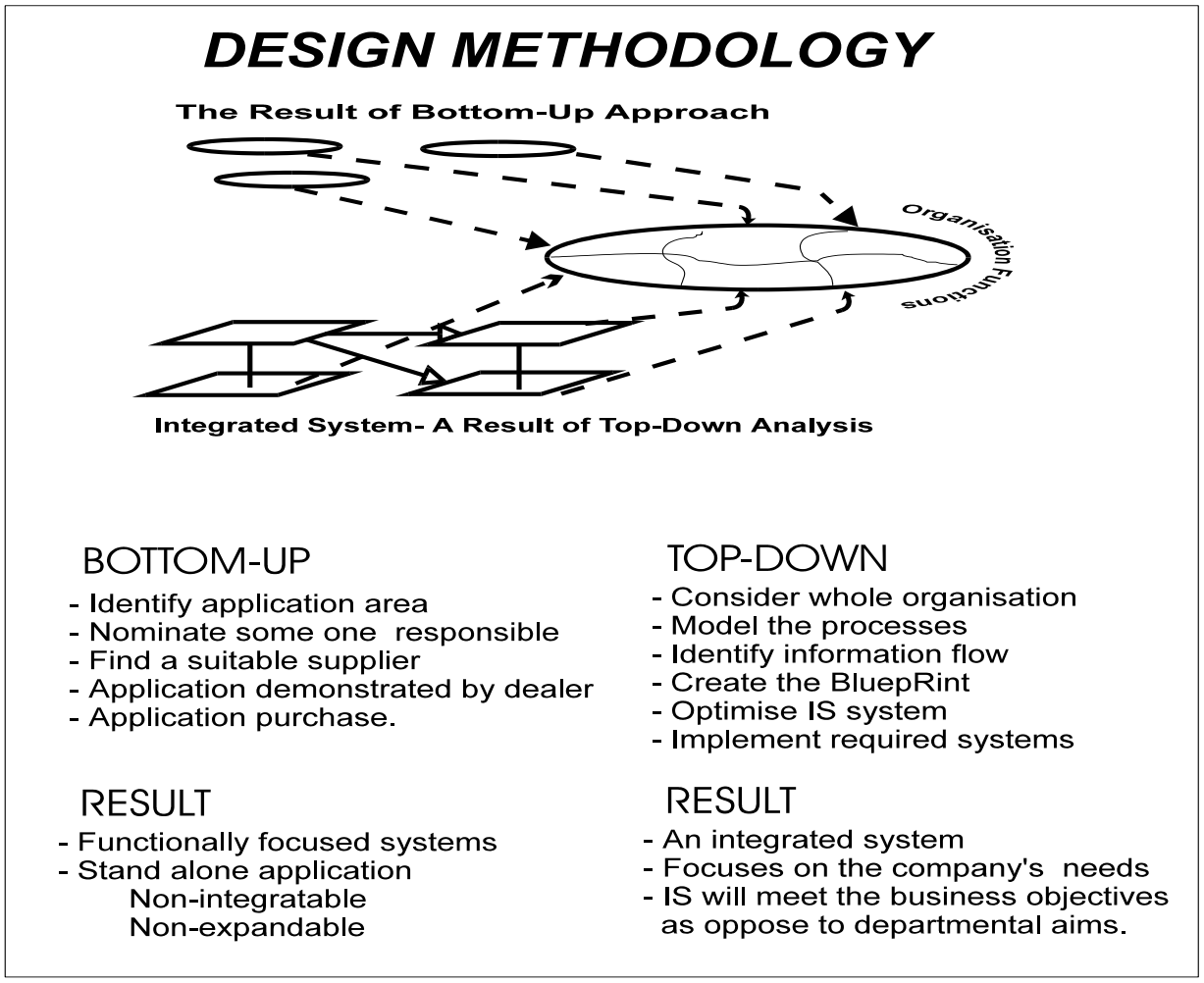
In this new era people are involved almost exclusively with handling and processing of information. There are other service areas such as tourism, travel, retail, police that are now dependent upon the information sector. In manufacturing, due to automation a greater number of people have moved from direct engineering work to processing information.

The processes in a firm can fall into those that can be formalised and those that are informal. These processes could also be conducted on a routine basis or non-routine basis. This gave rise to a matrix that set the development agenda for IS developments.

Non-routine	Database integration - systems can assist here i.e. budget analysis preparation, engineering cost analysis.	Electronic mail - It should provide the information structure, depends on the end-user having access to a computer. Merger and acquisition.
Routine	Ripe for automation e.g. order entry, inventory control, short term forecast, etc.	Lunch time meetings, team building exercises, etc.
	formal	Informal

Information System Modelling

It is important to set the scene prior to jumping in and talking about the tools used to develop or analyse information systems.



There are two terms that are generally used to define development strategy in many areas and not just IS/IT. They are Bottom up and Top down.

With the bottom-up approach, one starts at the lowest level and work upward towards a total integrated solution. Often this is not a satisfactory procedure for any but the smallest businesses because it fails to consider the business needs at the highest level. With this methods companies usually,

- identify their application areas, for instance, accounting, word processing, etc.,
- allocating a decision maker for each area i.e. a department manager,
- proposing a solution/purchase, and
- Implementing the optimum solution or purchasing and installing the product/equipment.

Application of the bottom-up approach is inappropriate in medium to large organizations. This is because they are often concerned with highly integrated data systems, strategic information systems, optimisation and specific business processes and business activities. For these organisations the top-down approach provides an opportunity for the entire organisation to be considered and the functions within it are then decomposed to their constituent components and they in turn are decomposed to reveal further sub-systems and detailed processes within them. This layering also identifies the system problems at various levels with increasing degree of detail.

The stages in the design and development an IS

The six generic design stages are also applicable to the design and development of an IS.

1- Determination of the Scope and Objective

This statement will indicate the scope and the area to be investigated. The example could be to speed up the sales order processing. The company may conclude that the orders are not processed quickly and this is why the company is losing customers.

2- Feasibility Study and System proposal

There are several major factors that influence the ability of the system to achieve the desired objectives. These factors could be related to any of the following reasons: technical, economical returns, non-economical returns, legal, operational, etc, and then to outline solutions with sufficient details to enable to make a decision as to whether it is worth going ahead with the project.

3- System Analysis

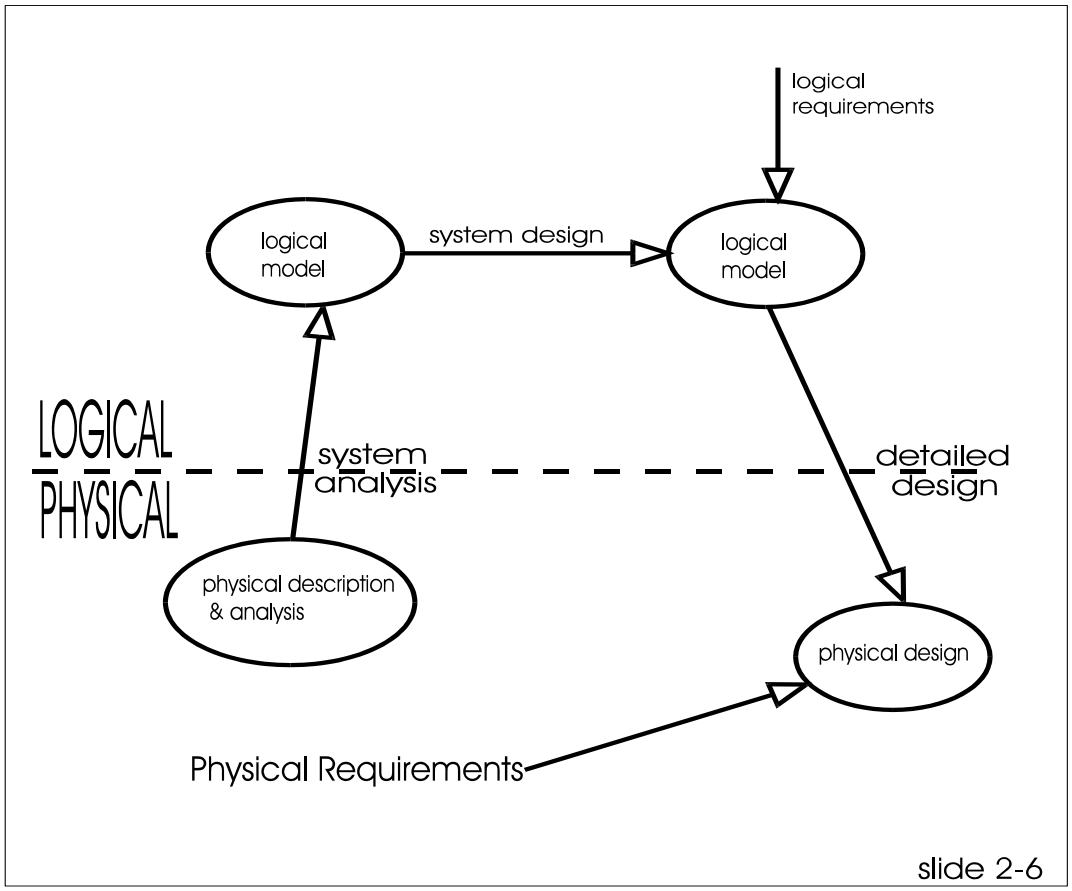
This concerns the study of an existing system for the purpose of designing a new improved system. This is where we apply system theory.

4 - system design

5- detailed design

6- implementation.

Now let's take a closer look at stages 3 (system analysis).



- Once an existing physical system is described then the analysis lead to a logical model of it as shown above. The secret here to keep away from details.

- The media on which data is stored is not important.

- The organisation of data store is not considered significant (just the type of data is important here, whether the cards indexed by surname or date of birth is irrelevant).

- Who or what carries out the process is ignored.

NB: Concentrate on what has to be done.

4- System Design

We should now know what our current system looks like. We apply the logical constraints and requirements of the new system and given that there are a number of different ways of implementing it on the physical level we could optimise our system. For example, where the computer should be located or where database must reside i.e. database vs. file or centralised vs. distributed arrangement.

5- Detailed Design.

At this stage detailed physical system specification, program functions/routines specification, project plans, file structures are produced.

6- Implementation.

The final stage involves:

- The purchase and installation of the hardware
- The programs (written and tested)
- Databases are created
- Operating procedures and manuals are written.

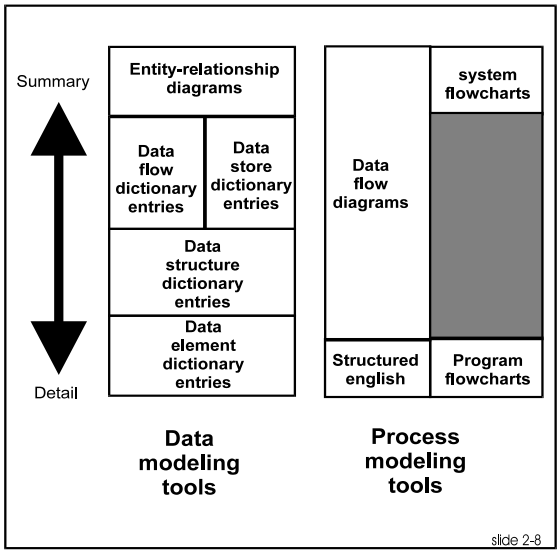
Analysis Tools

Just one more point about system design and analysis , remember that just as a tradesman or a builder uses tools to conduct her/his business the IS professional also needs to have tools to conduct her/his job. Analysis tools they use fall into two categories.

Process Modelling Tools - describe the processes and flows that are carried out by a system.

Data Modelling Tools - describe the data and its relationships, for example. What data (on employees and departments) and work relationships between the entities (Employee and Department).

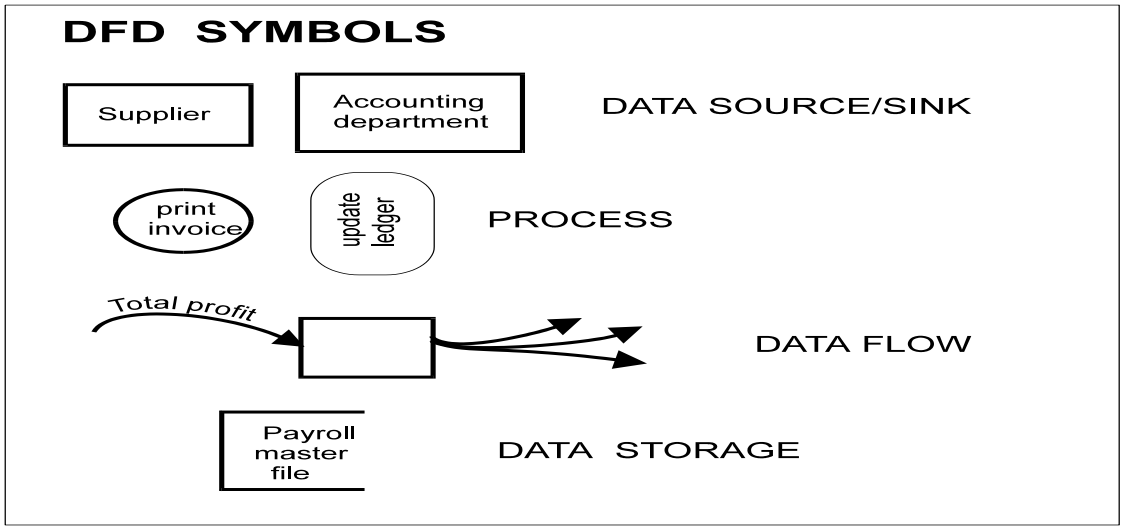
The specific tools used are:



Data Flow Diagram

DATA FLOW DIAGRAM

- Graphical presentation of a system



DFD is graphical representation of a system that uses a small number of symbols and shapes to illustrate how data flows through interconnected processes.

DFD is used to translate a physical system into a logical domain. This is done by decomposition of the physical systems into their constituents and the establishment of the data flow between them. That is to say we are conducting a system process analysis.

Data in a business can undergo complex processing prior to presentation. But it does not matter how complex the processes are, as they can be broken down into a number of basic elements/steps. These processes could be categorised as described below:

- Classification of data – for example into invoice data, order data, payment data, etc.
- Rearranging/sorting/filter data.
- Summarising/aggregating data.
- Performing calculations on data
- Selecting and query data

The process that data go through is specified using structured English/decision tables or logical flow charts.

The output of the DFD analysis therefore will be a document consisting of documents/charts/dictionaries. At the implementation level the data stores could become files and Databases. The processes could become sub-routines and functions and computer programs.

DFD Symbols:

Data source/sink - these are like windows between the system and its environment where data originates from or terminates to. Each terminator is labelled by the name of the environment element. It can be a person, an organisation of another system.

Process - this is the action that changes an input data/information to an output data/information. The process could be printing an invoice, computing net pay. Note that not much detail is required. It is pertinent to point out that taking a particular good from a basket in storeroom is not a process that is only data processes are considered and not physical processes.

Data flow – this refers to the path that data/or a group of data will have to travel. The structure of data is not specified so data can be a single data element or one or more files. Data flow can be divergence/convergence.

Data storage - to store data a data store is used. Think of a data store as data at rest.

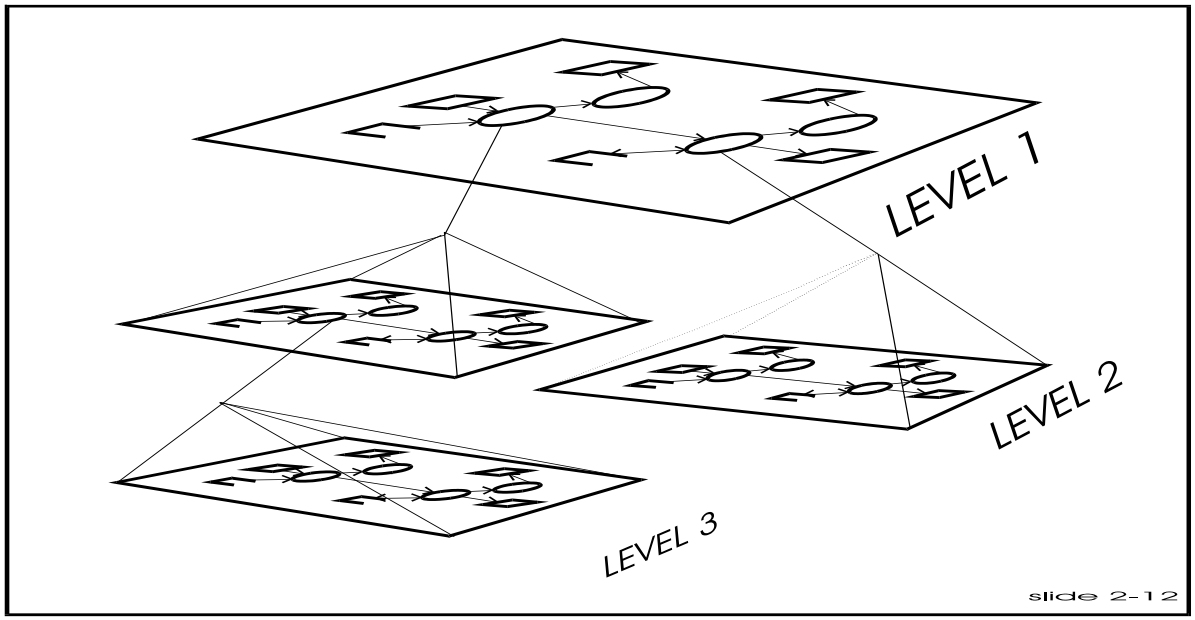
Process - the labels used must reflect the process being undertaken. Each process is numbered like 4.1, where the 1st number represents the level and the second number shows the process number at that level.

Data flow – this refers to the flow of data it often is easy to confuse data flow with control or material flow.

Data storage - the physical implementation of a store is not relevant at this stage.

To see another example of DFD please refer to MIS, Fig 8.8, Page 226.

Levelled DFD Diagram



To Draw a DFD Diagram:

1- Identify major processes
2- Identify major data source/sinks
3- Identify data flows
4- Label all necessary items
5- Conduct data modeling next.

One of the major uses of DFD is to act as means of communicating a particular system. Therefore large DFDS become confusing. It is for this reason that a level DFD is used.

DFD also lends itself well to the hierarchy and often the processes within the organisation specially where a number of large and/or complex processes are at work within various functionally organised firms. This lends itself well with the Top-down approach which allows a layering/levelling and decomposition of model. Furthermore, DFDs can become large and unmanageable. To overcome this a diagram is limits the number of process to 7, and not exceeding 9.

- Most modelling often starts at the highest level. This is called a context diagram.
- As the model is development leads to more levels and later more details are added. These level diagrams are called Figure n Diagrams.
- Limitation due to the physical system must not be allowed to effect this stage of the analysis, for instance, the process of distributing copies of invoice, should not be limited by the physical limitation due to the network configuration, protocol and so forth.

Structured English

_All the necessary processes are identified as the result of a DFD analysis. The next task to

define the events that occur within each process. A number of methods are available such as Flow charts/Structure English/Structuregrames etc. These techniques take over what DFD does not do, since the details of a process is not expressed within DFD.

Structured English documentation of a salesperson commision program.

```
START

Initialize storage
        TOTAL.SALES,  TOTAL.COM =0

Process sales data
        DO WHILE  (more  records)
              PERFORM  read data
              PERFORM  process data
              PERFORM  print data
         END DO

Finaltotals
        PRINT TOTAL.SALES,  TOTAL.COM

STOP


Read Data
        READ SALES.RECORDS

Process data
        IF (SALES.AMT  >  1000)
           THEN
                 COMMISSION = 100 + (SALES.AMT - 1000)* 0.15
           ELSE
                  COMMISSION = SALES.AMT *0.10
        END IF
        Accumulate  TOTAL.SALES,  TOTAL.COM

Print data
        PRINT  detail line
```

slide 2-13

Remember that the processes are established by talking to end users who often use descriptive language to define what actually happens at the physical level. This lead to a structured approach of using key words to define the objective called Pseudo code.

The following shows the C++ codes for the above programme written in Structured English:

```
#include <iostream.h>

int main()

{

        int s=0,c=0,ts=0,tc=0;
```

```
cout<<"This program is to calculate the commissions to be paid according to the sales amount."<<endl;

cout<<"The amount are shown in MTL."<<endl;

cout<<"Enter the amount of sales."<< endl;

cout<<"When you are finished enter -1 to quit"<<endl;

cout<<"Sales = ";

cin>>s;

cout<<endl;

while (s!=-1) {

        if (s>1000)

                c=100+(s-1000)*0.15;

        else

                c=s*0.1;

        cout<<"The commision to be paid for "<<s<<" MTL. sales is: "<<c<<" MTL."<<endl;

                ts=ts+s;

        tc=tc+c;

cout<<endl;

cout<<"Enter the amount of sales."<< endl;

cout<<"When you are finished enter -1 to quit"<<endl;

cout<<"Sales = ";

cin>>s;

cout<<endl;

}

cout<<endl;

cout<<"Total sales amount is: "<<ts<<" MTL."<<endl;

cout<<"Total commision paid is :"<<tc<<" MTL."<<endl;
```

```
    cout<<endl;

    return 0;

}
```

It is clear from the above the advantages of a language based on Natural Language than a High-Level language such as C++.

Process Documentation – this is a process dictionary compiled for all the identified processes that show the process inputs/outputs and the process itself. It links the structured English to a DFD and the data dictionary.

The field process has the process number and the name of that in the DFD. INPUT and OUTPUT specify the data that flows in and out of the process.

For another example relating to process dictionary refer to MIS, TMV.2, P 724

Flowcharts

This was one of the oldest structured approaches to system and program design analysis. Although still used extensively, its symbols, originating from 1960s, still represent the technology of that time and have not changed. Today flowcharting is used to cover two separate activities:

- documenting the system - these are referred to as system flow charts
- documenting programs - these are referred to as program flow charts

For an example of sales commission flow chart refer to MIS, TMV1.2, P 729.

Data Dictionary

As one develops the DFD much of the information about the data and/or the structure of data can be copied into a separate document called the data dictionary.

A written description of the data contained in the databases. There are four dictionaries representing Data Dictionary.

- Data flow dictionary entry - describes each data flow in the DFD diagram.

Refer to Slide MIS, TMII.2, P697.

Note that the named data structure is the content or the fields of this data flow.

- Data store dictionary entry - describes each unique data store in the DFD diagram.

Refer to MIS, Slide TMII.3, P699.

Note that the Activity Field lists how active certain number of records are and the Access Field lists

items such as security access, etc.

- Data structure dictionary entry - this is data on data and is provided for every structure listed on

both the data store and data flow forms.

Refer to MIS, Slide TMII.4, P700.

Note that Element Field lists each data element in the structure.

- Data element dictionary entry – this is the detail of every data element included in all of the structures in data flow and data stores of the DFD. One form per data element.

Refer to MIS, Slide TMII.5, P701.

NB: Alias - perhaps an invoice number is called a bill number.

Specific value identifies individual digits in a number, for example 1=UK country code

For further information follow the Kismet LTD case in Business Information System, Curtis.

Page 339 and pages 366-375.

For further study and for selecting a topic for your project, the following could be useful:

Presentation topics:

1. Office Automation P435, Ch15
2. Decision Support Systems P407, Ch14
3. Accounting Information Systems P357,P589, Ch12, Ch20
4. Expert Systems P459, Ch16
5. Marketing Information systems P529, ch18
6. Manufacturing Information system .P561, Ch19
7. HR Information systems P617, C21

Ref. Management Information System, McLeod

Chapter

2

Introductory Statistics

Principle Objectives

- *To introduce students to the fundamental aspects of statistics which are the prerequisites knowledge for the Course.*

- *To ensure students are aware that simple terms and concepts are sometimes not fully understood and their full comprehension is a prerequisite for the subject of data and information analysis.*

Introduction

An entire or total possible group that may be available for study is called a population. In practice we can only look at a small part of that population, called a sample. In general data may be of two types: 1) continuous data obtained by measuring 2) discrete data obtained by counting. Here are the two examples of continuous data and discrete data respectively:

- Of the total student population of 1500 at this University it may only be practical to measure heights of a sample of 100 of them.

- To discover the spread of the students from different parts of the country it may only be practical to count from a sample of 200.

Grouping the Data

The following is the of weight of 100 students in kilograms:

61 63 65 66 67 68 72 71 69 66 67 62 64 63 63 65 69 70 74 65

72 66 66 67 68 69 70 70 73 64 65 64 63 71 71 68 68 68 67 66

68 66 66 73 61 68 64 64 64 70 71 70 69 66 67 66 65 71 78 66

62 69 69 66 66 69 66 64 67 67 67 71 68 68 72 69 70 71 62 67

71 71 72 68 67 68 71 67 67 68 69 68 69 65 67 70 68 67 66 67

As it stands, it is difficult to readily extract any information from the above data. Therefore, we should group or classify it.

1) Find the range of data (max – min) = (74 – 61) =13 (by omitting 78 as this may be an error hence the figure is recorded but ignored for the time being)

2) The range is (usually) divided into a number of equal width classes (the number is between 5 and 20) say:

60-62 63-65 66-68 69-71 72-74

3) To avoid ambiguities we form boundaries by going to the next 0.5 of decimals to ensure no observation can fall on the division between two classes. These are called class boundaries.

4) Class limits are defined as the maximum and minimum possible observable values that can occur between two class boundaries.

Class Limits (kg)	**Class Boundaries (kg)**	**Class Mark**	**Tally**	**Frequency**
60-62	59.5-62.5	61	////	5
63-65	62.5-65.5	64	//// //// //// ///	18
66-68	65.5-68.5	67	//// //// //// //// //// //// //// //// //	42
69-71	68.5-71.5	70	//// //// //// //// //// //	27
72-74	71.5-74.5	73	//// ///	8
				Total: 100

Frequency Table of Weights of Students

5) Find the mid-point of each class called the class mark and form a tally to find the frequency of each class.

6) When tabulated such information is called a frequency table, and should always be labelled such.

7) The class width is defined as the difference between two consecutive class boundaries.

8) The relative frequency is defined as the frequency of each class divided by total frequency.

Note: It is not always convenient to use equal size classes e.g. earning in a firm:

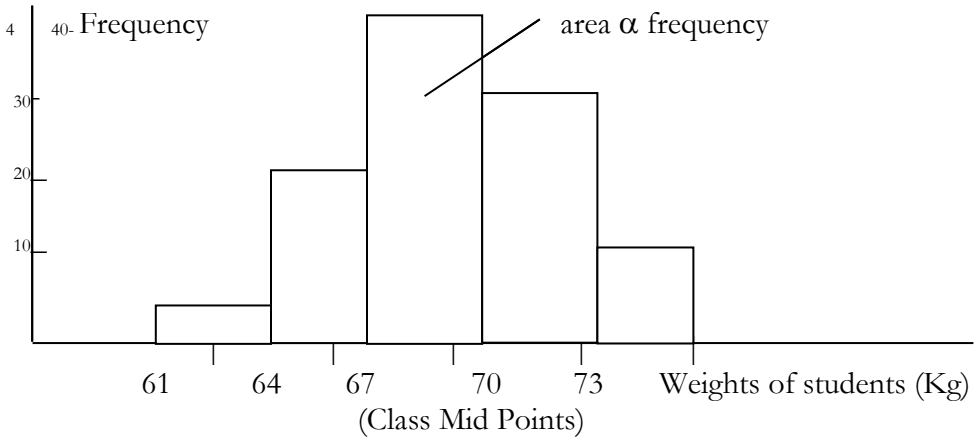
120 people at	£10000 or less
20 people between	£10000-£20000
1 person at	£50000

Here we use the concept of the open class i.e. "£10000 or less" or 20000 and above.

Presenting Data

Histogram: The frequency table of weights can be transferred to a graphical outline where the horizontal axis forms the observation of interest. The vertical axis forms the frequency measurement. A rectangle is erected on each class interval with its centre at each respective class mark and the area proportional to the frequency of each class.

(If the classes are all of equal width, the heights will be proportional to frequency, this being the most usual case).



Histogram of Weights of students

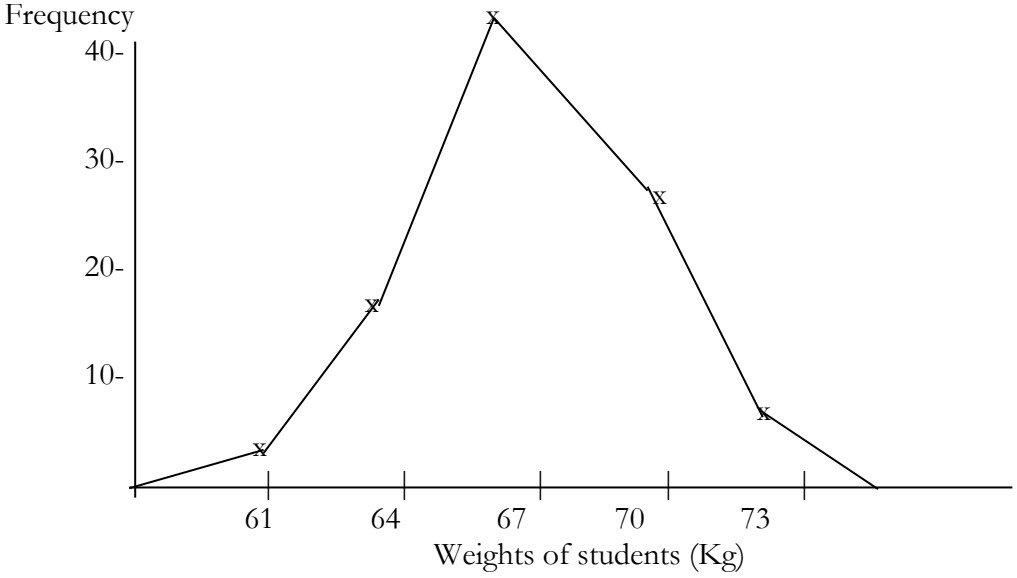
Note:

- The "picture" gives an idea of how the weights are distributed among the students.
- The total area under the histogram represents the total frequency of students.
- For all future purposes all observations in a class are considered to be "at the class mark". This is inaccurate: the accumulation may be on the left-hand side or right-hand side of the

given class mark. If the accumulation is on the left-hand side of the class mark, then the contribution is smaller than the class mark and if the accumulation is on the right-hand side of the class mark, then the contribution is larger than the class mark.

The Frequency Polygon: This is a plot of frequency against class mark (it may be obtained from the histogram by joining the midpoints of the tops of the rectangles). Again, it gives an idea of how the weights are distributed.

Frequency Polygon



Note: The smoothing of the frequency polygon to give a frequency curve relates to the population from which our sample of 100 students is drawn.

Later on we will need to become familiar with the idea of relative frequencies. A relative frequency histogram is a histogram with the vertical axis representing relative frequencies of each class (obtained by dividing each frequency by the total frequency).

Class Mark	Frequency	Relative frequency
61	5	0.05

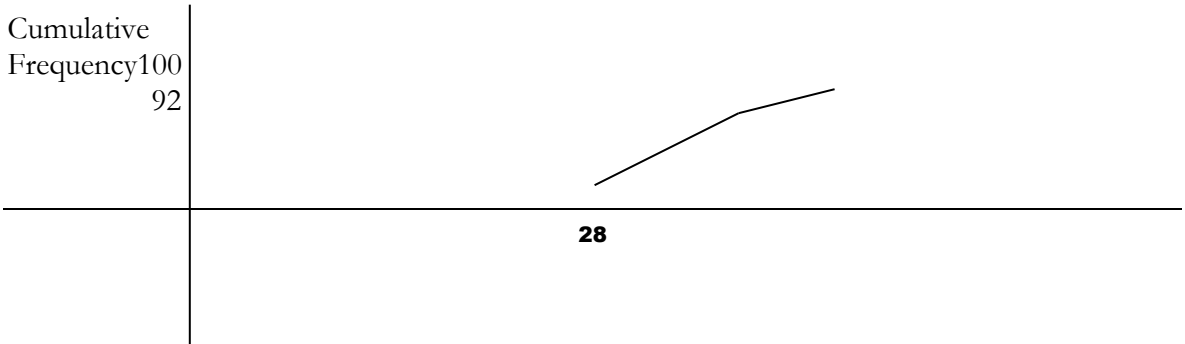
64	18	0.18
67	42	0.42
70	27	0.27
73	8	0.08
	Total: 100	Total: 1

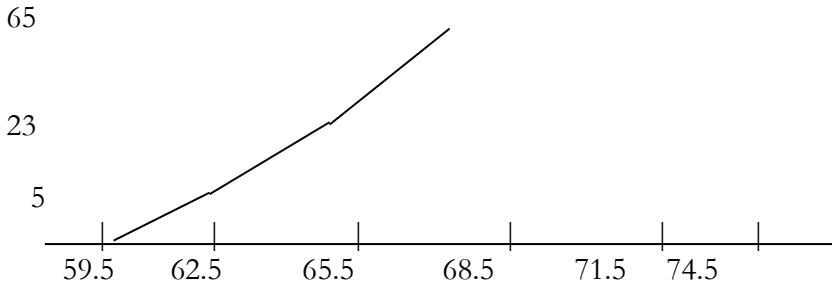
The total area under a relative frequency histogram represents the total of all relative frequencies, which is 1.

Cumulative Frequency Polygon (Ogive): As the name implies, the Cumulative Frequency is the sum of all frequencies below a particular Upper-Class Boundary.

Weight (kg)	"Less than"
<62.5	5
<65.5	23 = 5+18
<68.5	65 = 5+18+42
<71.5	92 = 5+18+42+27
<74.5	100 = 5+18+42+27+8

A Cumulative Frequency Polygon is a plot of Cumulative Frequency against Upper Class Boundary.





Cumulative Frequency Polygon of weights of students

Descriptive Measures

Measures of Location (Central Tendency) are parameters that represent the magnitude of the data.

1) The Arithmetic Mean or Mean of a set of data is denoted μ (for population mean) or $\bar{x}$ (for a sample mean) and is defined as

$$\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum x}{n}$$

where the data consists of the n numbers $x_1, x_2,\ldots,x_n$.

If the data is in a frequency table with k classes and $x_1, x_2,\ldots,x_k$ being midpoints and $f_1,f_2,\ldots\ldots,f_k$ being the corresponding frequencies then:

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + ... + f_kx_k}{f_1 + f_2 + ... + f_k} = \frac{\sum fx}{\sum f}$$

Note that,

$$\sum f = n = \text{total number of items in the data.}$$

Example on weights of students (mean of grouped data):

X	f	Fx
61	5	305
64	18	1152

67	42	2814
70	27	1890
73	8	584
	Total: 100	Total: 6745

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{6745}{100} = 67.45 \text{ Kg}$$

The mean is not always a good measure of central tendency because it is affected by extreme values, but it is used frequently in statistical theory and so is important.

2) The Median is the middle value of the data after it has been arranged in an array. If there is no middle value it is the mean of the middle two. Here is the example:

3 4 4 5 6 8 8 8 10 Median = 6

5 5 7 9 11 12 15 18 Median = (9 + 11) / 2 = 10

Median of grouped data

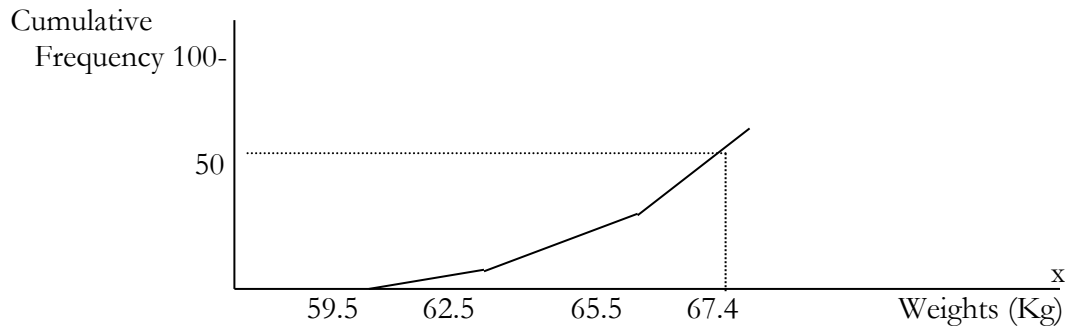
Example:

Class	Frequency, f
59.5–62.5	5
62.5–65.5	18
65.5–68.5	42
68.5–71.5	27
71.5–74.5	8
	100

We need to count up 50 (= 100 / 2, because we are looking at the middle item) items from top (or bottom). There are 5 + 18 = 23 in the first two classes and 5 + 18 + 42 = 65 in the first three classes. 50th student falls on third class we require another 50–23=27 from third class i.e. median class. Assuming that the 42 weights are equally spread between 65.5 to 68.5 kilograms, we require 27/42 of the way from 65.5 to 68.5.

i.e. median $= 65.5 + \frac{27}{42}(68.5 - 65.5) = 65.5 + \frac{9}{14} \cdot 3 = 67.4$ kg

Alternatively we may obtain the median of the grouped data from the cumulative frequency polygon:



(65.5,23) and (68.5,65) are the end points of the line we want.

$$\frac{y-23}{23-65}=\frac{x-65.5}{65.5-68.5}, \quad \frac{y-23}{-42}=\frac{x-65.5}{-3}, \quad \frac{y-23}{14}=x-65.5.$$

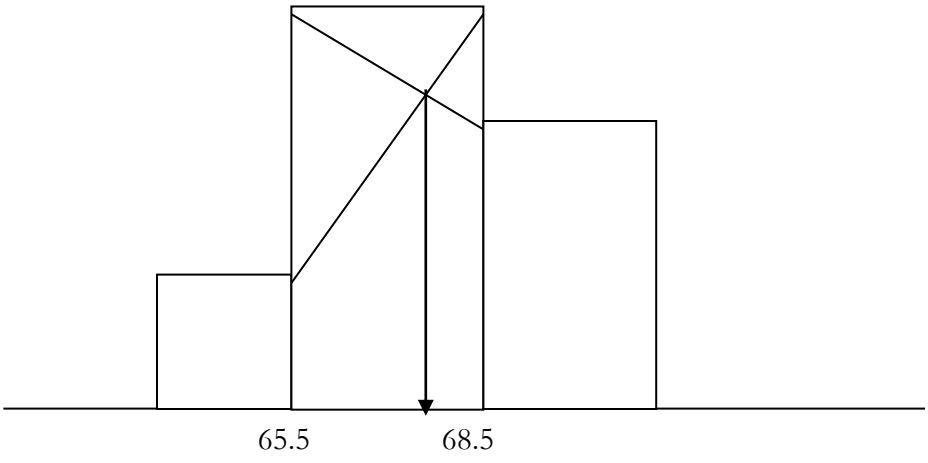
Take $y=50$ to get $\frac{50-23}{14}=x-65.5$ or $\frac{27}{14}=x-65.5$ $x=65.5+\frac{27}{14}$ =67.4 (as expected).

The median is a good representative of the data because it is not affected by extreme values.

3) The Mode of a set of data is the most common value. It may not be unique and it may not exist. Mode is not affected by extreme values, e.g.

2	2	5	7	9	9	9	10	10	11	12	Mode = 9
1	2	4	4	6	9	9	11	14			Two modes (bimodal)
5	6	8	10	12	14						No mode

Mode of grouped data may be found graphically by geometrical construction. Draw highest rectangle in the histogram (that of modal class) and the two either side:



(65.5,18) and (68.5,42) are the end points of the positively sloped line. (65.5,42) and (68.5,27) are the end points of the negatively sloped line. Abscissa of the intersection gives the mode (=67.3).

Measures of Dispersion

Measures of dispersion are parameters that measure the dispersion, spread or scatter of the data.

The range = highest value – lowest value. Advantage: Easy to calculate and understand. Disadvantages: Affected by extremes. Uses: In quality control: control charts.

Data - Sample vs Population

	Sample	Population
Mean	$\bar{x} = \frac{\sum x}{n}$	$\mu = \frac{\sum x}{n}$

Standard Deviation for non-grouped data is given by:

$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}} \qquad \sigma = \sqrt{\frac{\sum (x-\mu)^2}{n}} = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n}}$$

Groped Data – Sample vs. Population

	Sample	Population
Mean	$\bar{x} = \frac{\sum fx}{\sum f}$	$\mu = \frac{\sum fx}{\sum f}$

Standard Deviation for group data is given by:

for k classes with midpoints $x_1, x_2, \ldots, x_k$ and corresponding frequencies $f_1, f_2, \ldots, f_k$ is given by:

$$s = \sqrt{\frac{\sum f(x-\bar{x})^2}{(\sum f)-1}} = \sqrt{\frac{\sum fx^2 - \frac{(\sum fx)^2}{\sum f}}{(\sum f)-1}} \qquad \sigma = \sqrt{\frac{\sum f(x-\mu)^2}{\sum f}} = \sqrt{\frac{\sum fx^2 - \frac{(\sum fx)^2}{\sum f}}{\sum f}}$$

Using the data from our earlier example (and the formula for the sample), therefore:

x	f	fx	x^2	fx^2
61	5	305	3721	18605
64	18	1152	4096	73728
67	42	2814	4489	188538

70	27	1890	4900	132300
73	8	584	5329	42632
	Total: 100	6745	-	455803

Hence the standard deviation can be calculated:

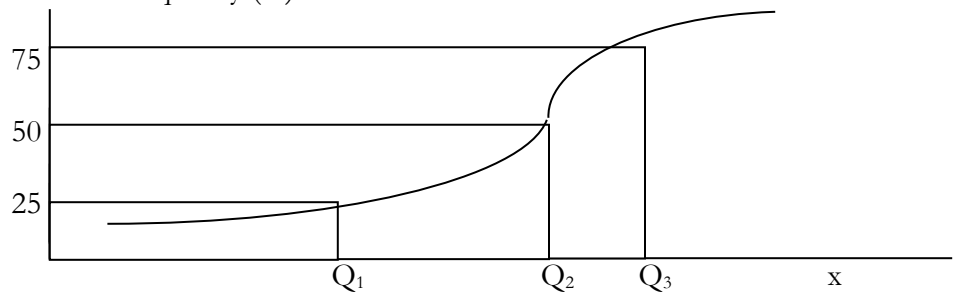
$$s = \sqrt{\frac{\sum fx^2 - \frac{(\sum fx)^2}{\sum f}}{(\sum f) - 1}} = \sqrt{\frac{455803 - \frac{6745^2}{100}}{100-1}} = \sqrt{\frac{455803 - 454950.25}{99}}$$

$$= \sqrt{\frac{852.75}{99}} = 2.935$$

Note that Variance= (standard deviation) ^{2}

Interquartile Range is obtained from Cumulative Frequency Polygon

Cumulative Frequency (%)



Q_1, Q_2 and Q_3 are known as Quartiles. Note that Q_2 = Median.

Interquartile Range = $Q_3 - Q_1$ i.e. range that contains the middle 50% of the data. Sometimes we use the Semi-Interquartile Range = 1/2 ($Q_3 - Q_1$) as a measure of dispersion.

The first and third quartiles may be calculated from grouped data in a similar way to the median:

Q_1: We want the 25th item from top. 5 + 18 = 23 (sum of first two classes). 25 – 23 = 2. Assuming 42 weights have equally distributed.

$$Q_1 = 65.5 + 2\frac{68.5 - 65.5}{42} = 65.5 + \frac{2}{14} = 65.642.$$

Q_3: We want the 75th item from top. 5 + 18 + 42 = 65 (sum of first three classes).75 – 65 = 10 Assuming 27 weights have equally distributed.

$$Q_3 = 68.5 + 10\frac{71.5 - 68.5}{27} = 68.5 + \frac{30}{27} = 69.611$$

The advantage of the Inter-Quartile Range is that it is unaffected by extreme values.

Remember:

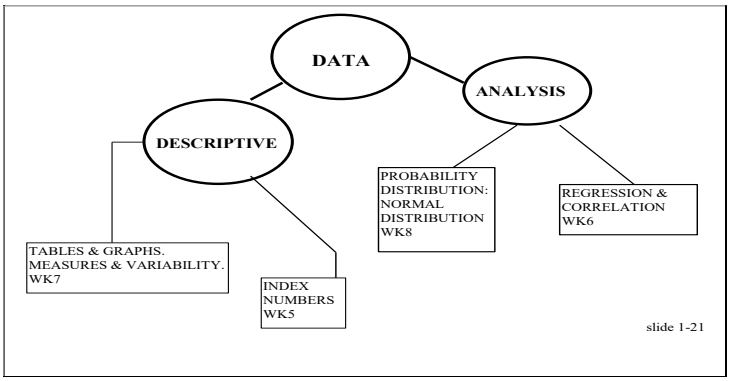
- Descriptive statistics is the numerical method used to tabulate, present (table/graphs), to summaries data sets.

- Analytical statistics covers the core of the statistics theories and techniques. These could be classified as:

The concept of probability (assessing the likelihood of events in a context of uncertainty example: likely hood of a machine braking down).

Statistical inference and decision making (making decision about population parameters based on information from sample)

Tests of association: "goodness of fit" measurement of relationships between variables.



Slide 1-6 Notes: Data can be sampled data or population data.

Tutorial

Introduction to Spreadsheets using Excel.

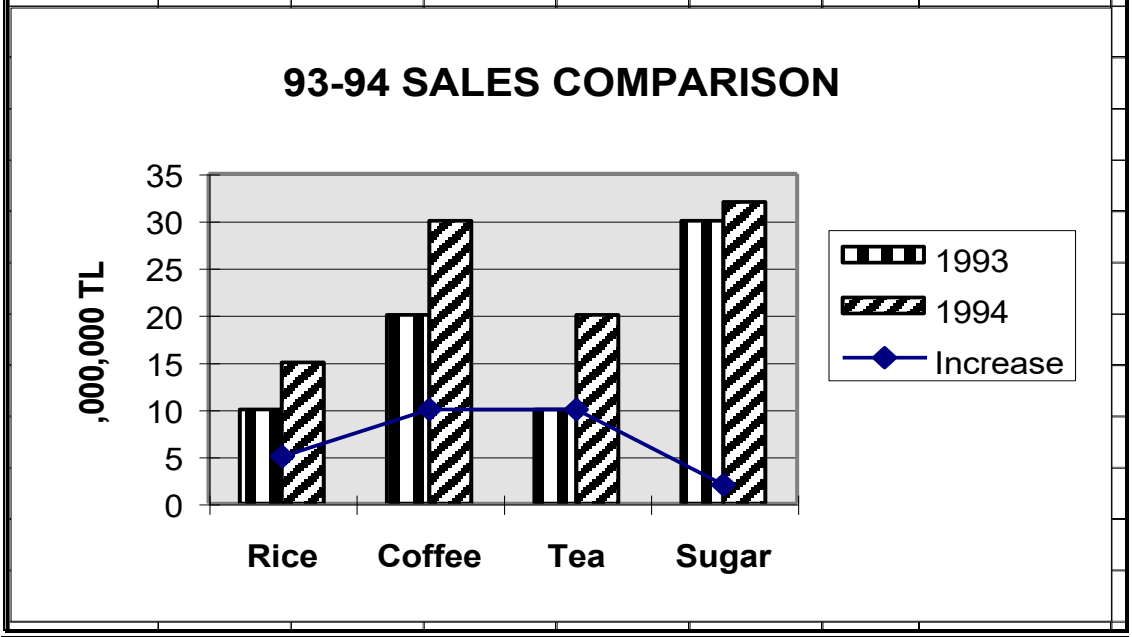
Your task: To produce the following spread sheet (noting all its details such as shading, graph titles) utilising the given formulae.

Sales Increase = The difference in the sales value of any two given years

VAT = 17.5% of Sale

Total Sales = Sum of all items listed

Sales (,000,000TL)							
ITEMS	1993	1994	Increase		94 VAT		VAT Rate:
Rice	10	15	5		2.625		17.50%
Coffee	20	30	10		5.25		
Tea	10	20	10		3.5		
Sugar	30	32	2		5.6		
total sale	70	97	27		16.975		



- Save your worksheet with a name EXER1 (make sure to save it in your own directory)
- Type your name in a free cell and print your worksheet (to include in your portfolio)

Example 1 The numbers shown below are the times in seconds (to the nearest second) for 40 children to complete a length of a swimming pool. The swimmers were divided into heats as the pool had eight lanes.

	Lane number							
Heat	1	2	3	4	5	6	7	8
I	40	49	43	35	42	43	46	36
II	42	36	37	44	39	41	31	45
III	38	48	44	51	38	53	35	32
IV	30	43	41	52	46	43	50	40
V	39	41	48	47	32	52	47	42

Display the data in the form of a grouped frequency table using intervals 30-34, 35-39etc.

Example 2 The figures below show part of a frequency distribution. State the lower and upper class boundaries for the second class.

Lifetime (hours)	Frequency
400-449	22
450-499	38
500-549	62

Example 3 Five coins were tossed 100 times and after each toss the number of heads was recorded. The table below gives the number of tosses: during which 0, 1, 2, 3, 4 and 5 heads were obtained. Represent this data in a suitable diagram.

Number of tosses	Number of heads (frequency)
0	4
1	15
2	34
3	29
4	16
5	2
	Total: 100

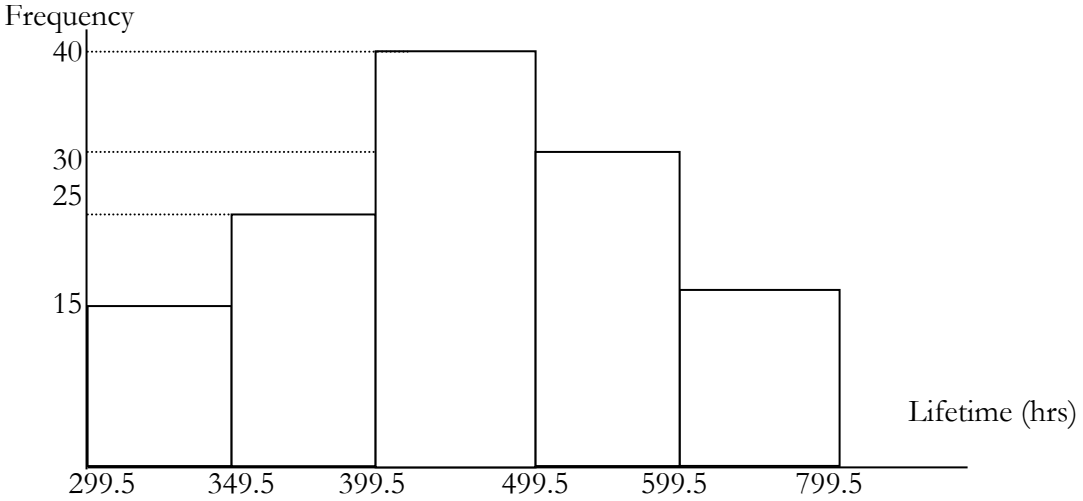
Example 4 The table below gives the age distribution of the workers in a certain factory.

Age group	16-20	21-25	26-30	31-40	41-50	51-70
Number of workers	20	18	14	18	16	24

Draw a histogram of this information.

Example 5 Five packets of chemical have a mass of 20.0l grams, 3 have a mass of 19.98 grams and 2 have a mass of 20.03 grams. What is the mean mass of the packets?

Example 6 Following figure shows a histogram for the lifetime of electric light bulbs. Draw up a frequency distribution.



Example 7 Draw a frequency polygon to represent the information given below

Age of Employee	Frequency
15-19	5
20-24	23
25-29	58
30-34	104
35-39	141
40-44	98
45-49	43
50-54	19
55-59	6

Example 8 The heights of 5 men were measured as follows: 177.8, 175.3, 174.8, 179.1, 176.5 cm. Calculate the mean height of the 5 men.

Example 9 Each of 200 similar engine components are measured correct to the nearest millimetre and recorded as follows:

Length (mm)	Frequency
198	8
199	30
200	132
201	24
202	6

Calculate the mean length of the 200 components.

Example 10 Using a coded method calculate the mean of the frequency distribution given in Example 9.

Example 11 The marks obtained by 50 students in a test were as follows:

Marks	10-19	20-29	30-39	40-49
Number of students	2	8	31	9

Calculate the mean mark obtained.

Example 12 The heights of a group of boys are measured to the nearest centimetre with the following results:

Height (cm)	157	158	159	160	161	162	163	164	165	166	167
Frequency	20	36	44	46	39	30	22	17	10	4	2

Find the mode of the distribution.

Example 13 Obtain a cumulative frequency distribution for the data below.

Height (cm)	Frequency
150-154	8
155-159	16
160-164	43
165-169	29
170-174	4

Example 14 The table below shows the annual rents of people in 1972.

Rent (£ per annum)	Under40	40-	80-	120-	160-	200-	240-280	280 and over
Percentage Of households	2	15	25	27	18	8	3	2

Draw an Ogive and from it estimate the median-rent.

Example 15 The weekly wages of five office workers are: £142.5, £155, £210, £171 and £200. Find the range of wages.

Example 16 An examination of the wages paid by a certain company showed that the upper quartile was £256 per week whilst the lower quartile was £192 per week. Find the semi-inter-quartile range.

Example 17 Find the standard deviation of the numbers 3, 5, 8, 9 and 10 (regard as population)

Example 18 The diameters of 200 bearings were measured with the following results:

Diameter (mm)	5.94-5.96	5.97-5.99	6.00-6.02	6.03-6.05	6.06-6.08
Frequency	8	37	90	52	13

Calculate the mean diameter and the standard deviation

Assumed mean = 6.01 mm Unit size = 0.03 mm.

Chapter

3

Standards and Errors

Principle Objectives

- *To introduce the students to the concept of standards.*
- *To enable students to distinguish between the primary and secondary standards.*
- *To remind students of the main and supplementary international system of units.*
- *To introduce the student to different type of errors and their causes.*
- *To understand terms such as accuracy, resolution, linearity, sensitivity and calibration.*
- *To identify and analyse the sources of uncertainty.*
- *To apply numbers and understand their representation.*

Introduction

Two major systems of units have been concurrently used in recent years, the metric and imperial systems, both in existence. The move towards "The International System" of units (SI) was intended by 1975. It is now accepted and agreed and the "Six Base" units in SI are as follows: -

Quantity	Unit	Symbol
1. length	the metre	m
2. mass	the kilogram(me)	Kg
3. time	the second	s
4. temperature	the Kelvin	K
5. electric current	the ampere	A
6. luminous intensity	the candela	cd

In addition there are two supplementary units:

phase angle	the radian	rad
solid angle	the steradian	Sr

The latter two units are not regarded as base units as yet

Primary Standards

These are the ultimate standards for the units. The nature of the primary standards depends upon the nature of the quantity to be measured.

In SI units, the primary standards are as follows:

1. Length -meter: The length equal to 1,650,763.73 wavelengths (of the orange line in the spectrum of an internationally-specified Krypton discharge lamp) in vacuum of the radiation corresponding to the transition between level $2P_{10}$, and 5 ds of the Krypton 86 atom (reproducible).

2. Mass -Kilogram: The mass of a platinum-iridium cylinder preserved at the International Bureau of Weights and Measures at Sèvres, near Paris (Fixed)

3. Time - second: The interval occupied by 9,192,631,770 cycles by the radiation corresponding to the transition of the caesium-133 atom (reproducible)

4. Temperature -Kelvin: The Kelvin is the fraction 1/273.15 of the thermodynamic temperature of the triple point of water. On the Celsius scale, the temperature of the triple point of water is 0.0l°C. Therefore, l° C = 273.15 °K.

Note: A temperature interval of l° C = a temperature interval of 1°K.

5. Electric current -ampere: The ampere is defined as that current which, if maintained in two straight parallel conductors of infinite length, of negligible circular x-section, and placed 1 meter apart in a vacuum, would produce between these conductors a force of 2×10^{-7} N per meter of length (reproducible).

6. Luminous intensity - Candela: The luminous intensity in the perpendicular direction of a surface of 1/600,000 of black body at the temperature of freezing platinum under a pressure of 101,325 N/m² (producible).

The definition of the supplementary units is:-

1. phase angle-radian: The radian in the phase angle between 2 radii of a circle which cuts off on the circumference ~ are equal in length to the radius;

2. solid angle - steradian: the solid angle which, having its vertex in the centre of a sphere, cuts off an area on the surface of the sphere equal to that of a square with sides of length equal to the radius of the sphere.

In the past primary standards were of a fixed nature, i.e. a physical object, e.g. the standard 'meter' was a metal bar with lines engraved upon it, likewise the standard yard.

Recently there has been a preference for reproducible primary standards. The reproducible standards can then be acquired. The expertise and equipment required for these are highly specialised and can only be acquired from established laboratories such as the National Physical Laboratory (N.P.L.) in the U.K.

For industrial purposes, more practical forms of standard are required for routine checking. These standards, in decreasing order, (i.e. less accurate) are:-

1. Reference

2. Laboratory

3. Working

Whilst these have no legal definition, a valid system of standards must conform to the principle of traceability, i.e. any standard must be traceable back to the primary standards.

In the U.K. the organisation charged with care and manufacture of primary standards is the National Physical Laboratory (N.P.L.) The British Calibration Service (B.C.S.) deals with calibration at lower levels relieving the N.P.L. of minor workloads.

Error

The True Error is the actual true value of the measurand. Because of inaccuracies in the measuring system it is never obtained except by accident:

The Indicated Error is the magnitude of the measurand indicated or recorded by the measuring system.

The Total Error (usually referred to as simply error) is the difference between the 'indicated value' and the 'true value' taking account of the sign.

Error = Indicated Error – True Error = $V_i - V_t$

The classification of errors may be undertaken in a number of ways. Consider the following:

A. Observer Errors

1. Visual Errors_- These are errors resulting from parallax using analogue display devices. they may be minimised by good design of the readout devices, i.e. indicator close to scale, mirror backing, etc., or eliminated by the substitution of a digital display.

2. Individual's Errors_- Some individuals constantly read high or jump the gun when synchronized readings are to be taken. They may be reduced by experience and training.

3. Computational Errors - These often arise due to incorrect mathematical techniques or "rounding up".

4. Blunders - These do occur, and as long as they can be identified they can be disregarded. They are identified by repeating measurements and plotting results.

B. Measuring System Error

1. Systematic Errors_- That part of an error which is due to a consistent cause and is therefore of constant magnitude when repeated measurements are made to an unchanging quantity under unchanging conditions. There are three components of systematic errors:

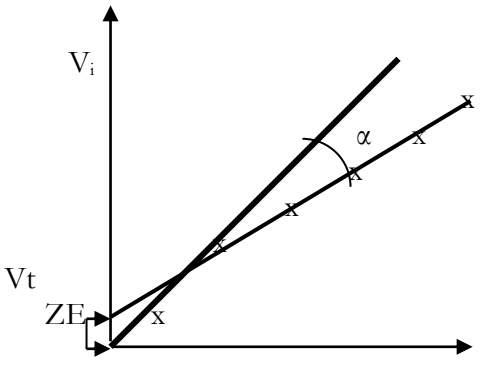
   - Zero Error - The error of the system when the measurand is at zero or datum value.

   - Sensitivity Error - The error decreases or increases progressively as the measurand increases. Also called GAIN or SLOPE error.

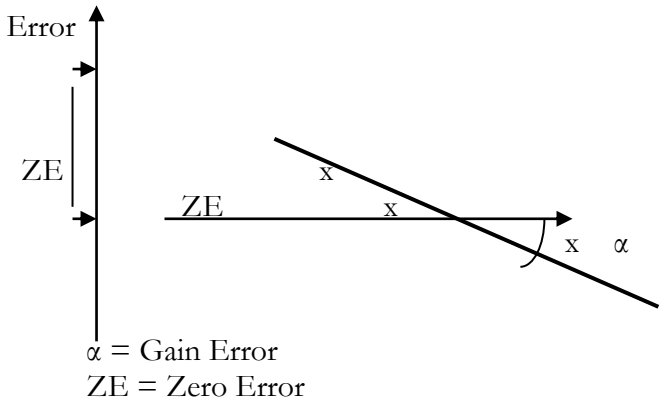
   - Linearity Error_- The indicated value departs from a proportional relationship with the measurand.

These errors are identified by plotting 'calibration graphs' or 'error graphs' as shown below:

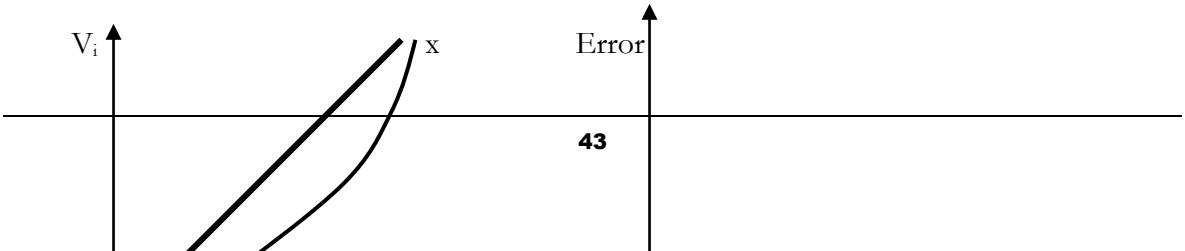
Calibration Graph

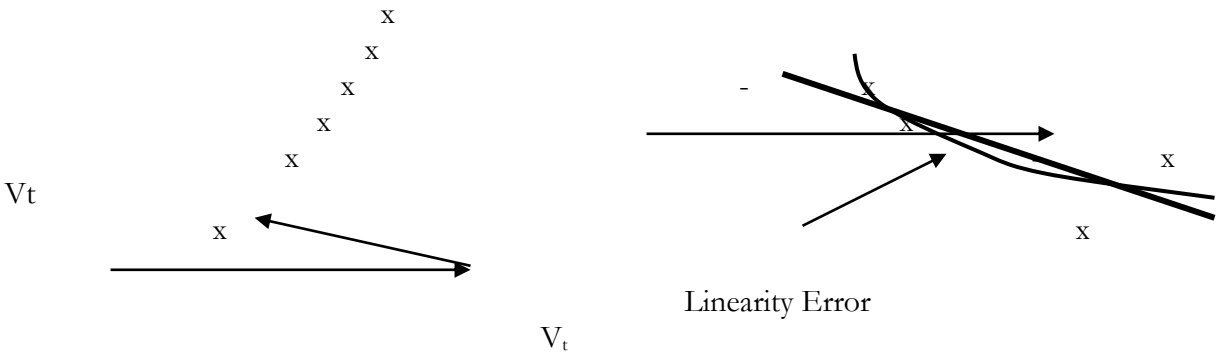


Error Graph



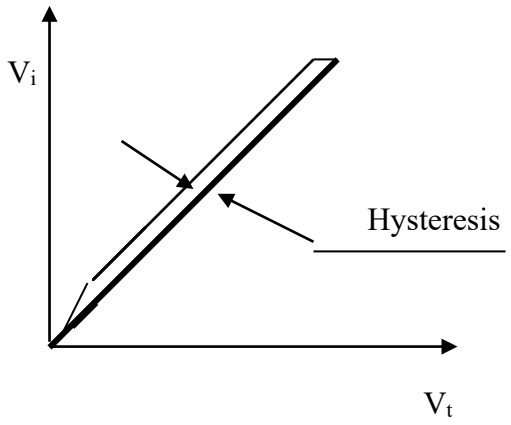
α = Gain Error
ZE = Zero Error





2. Random Errors – That part of an error which is unaccounted for by the systematic error, the magnitude of which varies in an unpredictable manner with a distribution about an average value when repeated measurements of an unchanging quantity is made under the same conditions.

Random errors are susceptible to statistical treatment. They may arise due to a number of factors including friction, backlash and hysteresis; the latter being of a mechanical or electrical nature and is identified by the typical characteristic shown below.



3. Environmental Errors – The most troublesome and most common environmental error is that due to changes in temperature, causing dimensional changes and changes in the physical properties both elastic and electrical. These errors may introduce both zero and sensitivity errors. DRIFT is the name given to the effect of changing environment.

4. Resolution Errors – These result from the inability of the operator to read the output to an infinite degree, due to the fact that the readout may operate in steps e.g. a slider moving over a potentiometer. In the case of a digital readout a resolution error is inherent.

5. Dynamic Errors - The errors previously mentioned are all applicable to the measurement of static parameters. A measuring system will still experience these errors when subjected to a rapidly changing input, and in addition there will be errors due to the inability of the system to respond at the required speed. These additional errors are called dynamic errors

and may be far more serious than the static errors, and may invalidate a particular measuring system for a particular application.

6. Application Errors - These errors arise from the act of introducing a measuring system into a situation, examples are:

    - Placing a mercury-in-glass thermometer in a thimble of water will alter the temperature of the water.
    - A hand-held tachometer forced onto a rotating shaft may lower the speed of the shaft.
    - A bonded strain gauge mounted on a foil test piece will alter the stiffness.
    - The mass of an accelerometer placed on a small beam will alter the characteristics of the system.

Instrument Specifications

Accuracy

Accuracy is the closeness of the 'indicated value' to the 'true value' of the quantity being measured and may be specified in a number of ways.

- Percentage of true value. If the accuracy of an instrument is expressed in this way, then the error is calculated thus:

Error = $[(V_i - V_t) / V_t] \times 100\ \%$

The percentage error stated is the maximum for any point in the range of the instrument.

- Percentage of full-scale deflection (f.s.d.). Here the accuracy is calculated on the basis of the maximum value of the scale, thus:

Error = $[(V_i - V_t) / V_{f.s.d}] \times 100\ \%$

It will be seen that an accuracy specified as a percentage of f.s.d. implies a less accurate instrument than one having the same accuracy as a percentage of true value. For example, an error of +/- 1% of f.s.d. on a pressure gauge having a range of 1000 kN/m² would mean that a true pressure of 100 kN/m² could read from 90 to 110 kN/m²; as a percentage of true value it would read from 99 to 101 kN/m².

Sensitivity

Sensitivity is the ratio of the change in the output to the corresponding change in the input and is usually required to be as high as possible,

Sensitivity = Change in output signal/Change in input signal

The sensitivity of an instrument will have units according to the system used and the form of the measurand, e.g. mercury-in-glass thermometer - mm/deg. C, and a thermocouple -mV/deg. C.

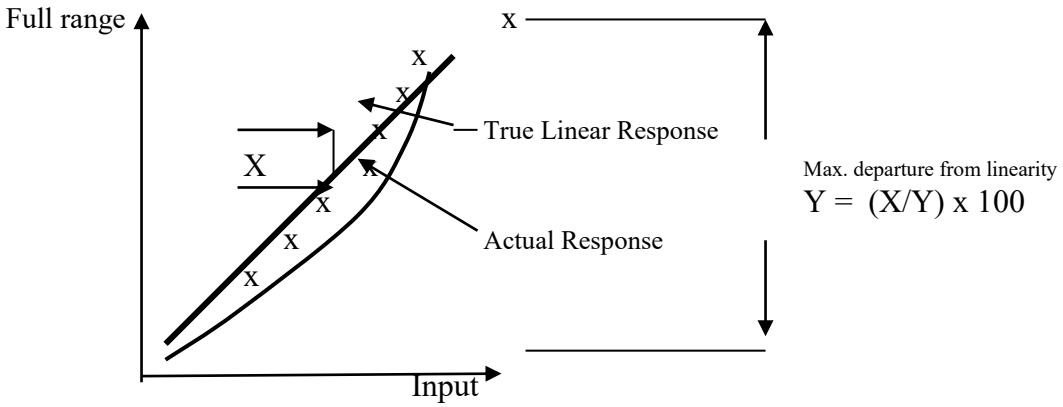
When the input and output of an instrument are or the same form e.g. a voltage amplifier, the words Magnification or gain ~re sometimes used instead or sensitivity.

Resolution

This is the smallest change in the input signal which can be detected by the instrument. It may be expressed as an actual value or as a fraction of the full-scale value.

Linearity

Linearity defines the relationship between the output and the input as the input spans the full measuring capability of an instrument. In ideal instruments the outputs are directly proportional to the inputs. See graph below:



Calibration

Calibration is the process of determining the relationship between the 'indicated value' and the 'true value' over the full range of an instrument.

Thus calibration is the process of checking a measuring system against a standard, or a standard against a higher-grade standard. (See also section on 'Standards'.)

Additional remarks

True value in any parameter does not exist. If the true value was known there would be no need for the measurement after all! But a "true value" is assumed after carrying out many measurements and ensuring consistency. Even then, the "true value" includes some uncertainty. In a philosophical sense, we would never know the exact extent of an error in any measurement:

- Error is defined as the deviation between the true value and the value at hand

- Uncertainty is the deviation between the assumed "true value" and the value at hand

The words error and uncertainty are used interchangeably, but what is really meant is the uncertainty. That is why in the rest of this chapter whenever the word error is used what is really meant is uncertainty.

C. Statistical Error

This error often occurs due to statistical variations relating to recording of data and methods used. The statistical uncertainty decreases as the number of measurements increases. The statistical uncertainty affects the precision of the results. If you are including an extra few mm due to incorrect placement of the ruler, when measuring a length, the result will be inaccurate by the same amount no matter how many times the measurement is repeated

Accuracy vs. Precision

The words precision and accuracy tend to be used synonymously. However, in scientific measurements, there is a clear distinction. Precision of a result means that next time the measurement is repeated using the same setup, the result will be very close to the previous measurement. The scatter of the results obtained from this setup will be limited to a very narrow interval. Accuracy, on the other hand, is a measure of how close the result is to the true value.

Note that a precise result is not an accurate result even though the measuring instruments might be giving results within a very narrow range, the calibration might be faulty or there might be some other systematic effects preventing the resulting value to get closer to the "true value". For instance, a digital thermometer may be designed to read to within 0.01 degree, but if there are any calibration problems then although one may think the instrument is accurate within the stated accuracy the reality is very different. In a 20 ^{o}C room a precise reading of 22.59 ^{o}C temperature may be observed but the reality is that the actual temperature is 20 ^{o}C.

Note also that an accurate result is not necessarily a precise result. There may not be any significant effect in the system so that the results obtained are accurate but the precision of the measurements is not adequately high. However, in practice, an accurate result is usually a precise result.

Parent vs. Sample Population

The set of measurements taken in any experiment can usually be considered as a sample of a bigger hypothetically infinite distribution called the parent population. In principle, the parent population contains the outcomes of all the possible measurements. If this parent population were to be known, the value sought in the measurement could be determined with the utmost accuracy. Unfortunately, this is not possible. Instead, a small set of measurements is taken as a sample of the parent population. This sample population should be of a size that would represent the parent population as best as it is possible. The desired values, which are the parameters of the parent population, can be estimated from the parameters of the sample population.

Significant Figures

The most significant figure is the left-most non-zero digit in a number

The least significant figure is, depending on the existence of a decimal point,

$$\begin{cases} \text{the right - most digit, if there is no decimal point} \\ \text{the right - most digit , if there is a decimal point} \end{cases}$$

The significant figures are all the digits between the least and the most significant figures, inclusively.

Suppose that you are measuring the length of a table with a ruler and the smallest division that the ruler has is the millimetre. Reporting a 2322 mm value after the measurement is meaningful and the number of significant figures is four. On the other hand, reporting a 2322.1234 mm value is meaningless, because nothing can be determined less than a mm with the ruler mentioned. So the significant figures are determined by the precision of the measurement instrument.

Number	Signif. Figs.	No. of signif. Figs	rounding to 5 signif.figs.
256.8746	2,5,6,8,7,4,6	7	256.87
2322	2,3,2,2	4	Not possible
255.465	2,5,5,4,6,5	6	255 (3 Signif. Figs.) only)
150.0000	1,5,0,0,0,0,0	7	150 (3 Signif. Figs.only)

Note that a number without any decimal point cannot be rounded.

Chapter

4

Index Numbers

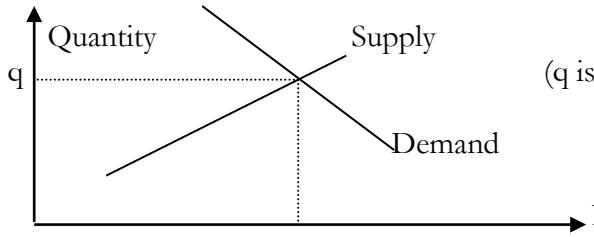
Principle Objectives

- *To explain the concept of the index number*
- *To give examples of their use*
- *To explain how they are calculated*
- *To show how simple, aggregate and weighted aggregate indices are calculated.*
- *To understand and calculate price indices and Retail Price Index (RPI)*

Introduction

Let us look at an example of how a price for an item such as shoes is arrived at.

Supply and Demand Curve



(q is the equilibrium quantity)

In a free market economy the price is a function of the intersection of demand and supply curve- giving rise to the Equilibrium Quantity.

There are a number of factors that can affect this equilibrium, these are:

- ❖ Increase in income
- ❖ Reduction in supply
- ❖ Increase in demand
- ❖ Inflation
- ❖ Deflation
- ❖ Technology

Note: Inflation is the changes in price increasing (a central controlling variable). Deflation is the changes in price decreasing.

Index number helps us explain how a variable changes over a period of time.

For example how a computer price has changed over a period of time, or how computer engineers pay compares to industrial engineers over the past 3 years. To do this we would ideally like to have one number.

For instance the price of a computer before and after could have been:

Before	After
1000	2000

Difference = 2000 –1000 =1000

At first it looks the difference is the number we are looking for! However, there are more appropriate methods as outlined below.

- Simple Difference Method

So by stating that the price of a PC has changed by 1000 unit (Simple Difference Method) alone the information does not indicate whether the PC price has doubled, trebled, only slightly gone up or has jumped up substantially.

This is because we do not know what they have been before. Therefore we need a starting point.

- Percentage Method

The next improvement is to divide the new number by the old number and find the percentage change as shown below:

2000/1000 x 100 = 200 %

This means that the price of the PC has doubled from whatever it was before. We have now a relative price (an index number for PC). This is known as the Simple Index Number - defined as a number that measures the new price (or indeed any other variable as the case may be such as a given quantity) relative to what it was before.

PR (Price Relative, %) = P_n/P_b x 100

Where P_n and P_b are price now and price before respectively.

Example (percentage method)

Year	PC Price	PR (Simple Indexes)	Price Change (%)
2000	1000		
2001	1500		
2002	2000		

Note:

- Base Year is 2000.
- PR values are simple indexes.
- An index of 100% means no change.

PR is fine for one item. What if there are many items that are purchased together i.e. printers, scanners, speakers, etc.

Example

Item	2000	2001
PC	1000	1500
Printer	500	600
Monitor	750	1000

How are we going to find an index representing all these items (explain the price of these items all together with one number)? That is to say an aggregate (all together) price index.

Aggregate means all items together.

There are 2 options.

Option 1:

a) Add all expenditure in 2000 b) Add expenditure in 2001. Then divide b) by a) (again take 2000 as the base year) to find the simple aggregate index.

ITEM	2000 (price)	2001(price)
PC	1000	1500
Printer	500	600
Monitor	750	1000
	2250	3100

Simple Aggregate Index = 3100/2250 = 137.7 %

Option 2:

Work out the PR of individual items. Then average all PRs.

ITEM	2000(price)	2001(price)	PR
PC	1000	1500	150%
Printer	500	600	120%
Monitor	750	1000	133.3%
	403.3%		

Averaged Aggregate Index = 403.3/3 = 134.4 %

How are we going to find an index that takes into account the quantity of individual items and gives an equal importance to them all?

The answer to this question is: Include quantities in our index by using the product "price×quantity".

Weighted Index

This index take into account the importance of each item i.e. if the quantities of PCs are larger, then the effect of the price of PCs will be amplified. The index can be calculated as follows:

Weighted Index = [Total value (cost) at present/Total value (cost) at the base year]x100

$$= \frac{\sum P_n Q_?}{\sum P_b Q_?} \cdot 100$$

The difficulty here is whether to include the quantity now or quantity before.

There are 2 methods, these are:

1. LASPEYRE

Base Weight Index $= \frac{\sum P_n Q_{\boldsymbol{b}}}{\sum P_b Q_{\boldsymbol{b}}} \cdot 100$

2. PAASCHE

Current Weighted Index $= \frac{\sum P_n Q_{\boldsymbol{n}}}{\sum P_b Q_{\boldsymbol{n}}} \cdot 100$

Example

ITEM	P_b (2000)	Q_b (2000)	P_n (2001)	Q_n (2001)
PCs	1000	8	1500	6
Printers	500	2	600	1
Monitors	750	6	1000	4

LASPEYRE	
P_nQ_b	P_bQ_b
1500.8=12000	1000.8=8000
600.2=1200	500.2=1000
1000.6=6000	750.6=4500
∑=19200	∑=13500
= (19200/13500) x 100 = 142.2%	

PAASCHE	
P_nQ_n	P_bQ_n
1500.6=9000	1000.6=6000
600.1=600	500.1=500
1000.4=4000	750.4=3000
∑=13600	∑=9500
= (13600/9500) x 100 = 143.1%	

Which method should we use?

- LASPEYRE: which measures the change in the cost of purchasing the same products using the base year quantities.
- PAASCHE:which measures the change in the cost of purchasing the same products using the current year quantities.

Which method is more appropriate? Compare the drawbacks and the advantages of LASPEYRE and PAASCHE.

Average Weighted Index

- Base Averaged Weighted Price Relatives
- Now (Current) Averaged Weighted Price Relatives

Q_n	ITEM	2000	2001	PR	Weight	PR×Weight
6	PCs	1000	1500	150%	9000	13500
1	Printers	500	600	120%	600	720
4	Monitors	750	1000	133.3%	4000	5332
					∑=13600	∑=19552

Index (I) $= \frac{\sum(\text{PR} \times \text{W})}{\sum \text{W}} = \frac{19552}{13600} \cdot 100 = 143.7\%$

Note:

In the above table the Weight (W) $= Q_n P_n$ (Q_n was used). If Q_b were used then Weight (W) = $Q_b P_b$.

$$\text{I}_i = \frac{(P_n / P_b)_i W_i}{\sum W_i} \cdot 100$$ for individual items.

$$\text{I} = \frac{\sum(PR \times W)}{\sum W} \cdot 100$$ for all items.

Retail Price Index (RPI)

This is the base averaged weighted price (relative) index. In England, since 1914, it measures the changes in the prices of a basket of some 350 or so goods and services bought by the average family from one month to another. The family samples are selected (7000) to cover a good representation of the population and for example excludes some group of pensioners and high income families whose spending pattern is not believed to be a reflection of their income.

Prices are measured at a cross-section of stores from selected hypermarkets to small shops across the country.

The content of the basket and their weights from time to time is changed to represent the changes in the consumer purchasing behaviour like the addition of colour TV or leisure spending.

The base year is also changes as appropriate. In UK the base years have been adjusted as 1956, 1962, 1974 and 1987 (RPI 394.5). If base years are not adjusted then the Index will become large and the items become out of date where comparisons are being made with a far distance point in time.

US RPI Components (August 1998)

Category	Weights
Food	152
Catering	47
Alcoholic Drinks	80
Tobacco	36
Housing	172
Fuel and Light	47
Household Goods	77
Household Services	48
Clothing and Footwear	59
Personal Goods and services	40
Motoring Expenditure	143
Fares and Other Travel Costs	20
Leisure Goods	47
Leisure Services	32
All Items	1000

Application of RPI

Actual Vs IPR Adjusted

	1987	1988	1989	1990	1991	1992	1993
RPI	100%	121%	129%	135%	138%	140%	143%
WAGES ACTUAL	430	470	490	520	530	540	550
WAGES IPR ADJUSTED	430	520	555	581	593	602	615

Summary:

Simple Difference Method and Percentage Method [PR index] [one item]

Simple Aggregate Index and Averaged Aggregate Index [many items]

Weighted Index [LASPEYRE (base) and PAASCHE (now)] [many items + equal importance]

Base or Now Averaged Weighted Index [many items + varying importance]

Tutorial

Example 1 Find the average weighted index in page 23 if Q_b was used in Weight ($W = Q_bP_b$).

Chapter

5

Correlation

Principle Objectives

- *To explain the concept of correlation by an example*
- *To derive a coefficient for two sets of data*
- *To interpret the coefficient of correlation*
- *To the introduce related terms and their significance*

Introduction to Correlation

Our interest here is to learn about extracting information about lists of numerical data that might or might not be related. For example suppose in a Bar manager that primarily sells beverages has to place his next day order with his supplier the day before. He has got the idea that his sales volume generally depends on the weather. So for some times he has been compiling a list of daily average temperatures (as given by the news paper weather forecasts) and the volume of the beverage sales to see if he could more accurately predict his daily needs.

Average Temperature (Degree C)	Liters of Cola (x1000)
23	58
17	50
24	54
35	64
10	40
16	43
15	42
24	50
18	53
30	62

Fig. 4.1 Bar Manager's Data

Note:

- Principally we would like to know whether these two variables that the bar manager has chosen are in any way related to each other. For if they are not he is wasting his time.
- Technically variables that are related are often said to be correlated. Even when two sets of data are known to be related to each other it is useful to be able to express the strength (closeness) of their relationship.
- The technique used to find their strength of their relationship (or their association) is called Correlation Analysis.

Other examples, where correlation analysis is used to establish as to whether a relationship or an association between two sets of figures exist, and how strong is the relationship, are as follows:

- Advertising expenditures vs. volume of sales
- Maintenance (quality) expenditures vs. the quality of parts
- Income vs. sales of luxury goods
- Family income vs. university education
- Student attendance vs. performance.

If there is a relationship between these numbers, what could it possibly be, and how are we going to discover it?

The method used to find the form of the relationship is called Regression Analysis. This method is fully explained in the next Chapter.

Once the form of a relationship is established it is then possible to predict or extrapolate (near) future outcomes based on the past data as long as the pattern of future events follow the past.

Variables could be related to each other either functionally or statistically. For example if y = mx + c represents the equation of a straight line, we can say the y is function of x as long as we know the values of m and c. That is to say, we can determine the exact value of y if we are given a given value of x. Another word for Exact is Determined hence the term deterministic. By the way, for the time being, x is the independent variable and y is the dependent variable.

This exact relationship, unlike many scientific relationships, does not often exist in statistical analysis of socio-economic data. Instead the data when plotted represents scatter. The plot of the Bar manager's data is the case in point. Temperature and cola consumption are statistically related.

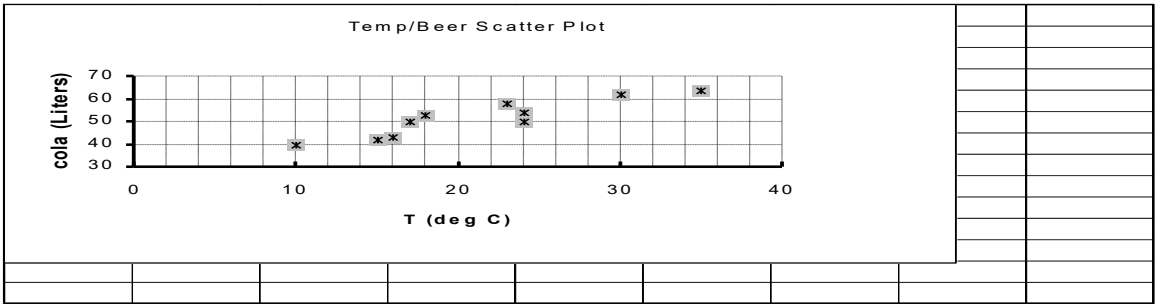


Fig. 4.2 The plot of Cafe manager's data

The first step in such an analysis is then to plot the points to see how the picture looks. Each pair of data is plotted on the two suitably selected axes.

If one then looks at the scatter one can generally see that as (forecasted) temperature raises the cola consumption also increases. The points do not exactly line up in a straight line but an approximate "linear" relationship seems to exist. This relationship however is statistical and not functional since there is not a unique cola consumption level for each degree in temperature. For example at 24 degree Centigrade there had been two different level of beer consumption.

Not all regressions are linear or near linear. The graph below shows the effect of increasing labour on the production unit costs. This describes the concept of the diminishing returns.

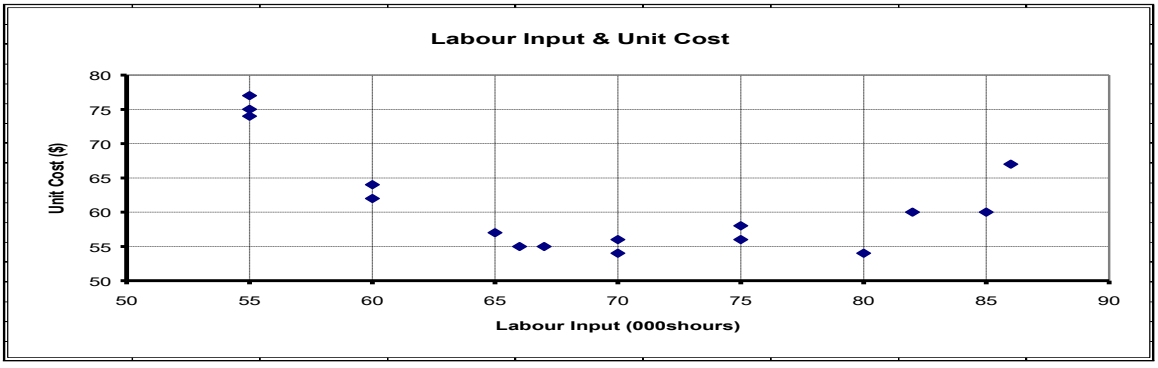


Fig. 4.3 A nonlinear situation

Correlation Coefficient

We have already said that the association between variables is called correlation. In simple terms a correlation could be Strong or Weak or a level in between the two.

If the two variables change in the same direction (for example the temperature and cola, then if temperature increases so does the cola consumption) then this is called a positive correlation.

If the variables change in the opposite direction (if your income increases your consumption of potato might decrease because you are now richer) then this is called a negative correlation.

Note:

A negative correlation does not mean "no correlation" but it indicates the direction of correlation.

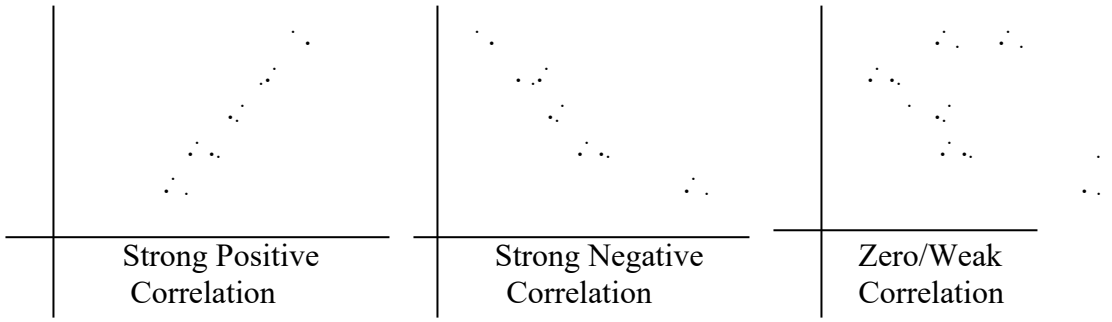


Fig. 4-4 Various forms of correlation

Class discussions:

What can be observed about these graphs?

It looks as if we might be able to draw straight lines through the points on the first and second plots. The correlation of the first one would be a positive one since the graph has a positive slope and the second has a negative correlation since it has a negative slope.

Although one can look at a scatter plot and judge whether there is a positive, negative, strong, weak, or no correlation it would be desirable to be able to calculate a numerical indicator of the degree of correlation.

Deriving an expression for correlation coefficient

One way to develop a numerical correlation indicator is as follows:

For the bar manager example if we work out the averages for the two columns we shall have:

Average of the (average) temperatures:

$$\bar{x} = \frac{\Sigma \text{average temperatures}}{\text{N}} = \frac{\Sigma \text{T}}{\text{N}} = 21.2$$ and,

Average of Liters of cola: $$\bar{y} = \frac{\Sigma \text{Liters of Cola}}{\text{N}} = \frac{\Sigma \text{L}}{\text{N}} = 51.6$$

where N is the number of entries or data points, in this case 10. We then can plot the lines $x = \bar{x}$ and $y = \bar{y}$ on our scatter plot.

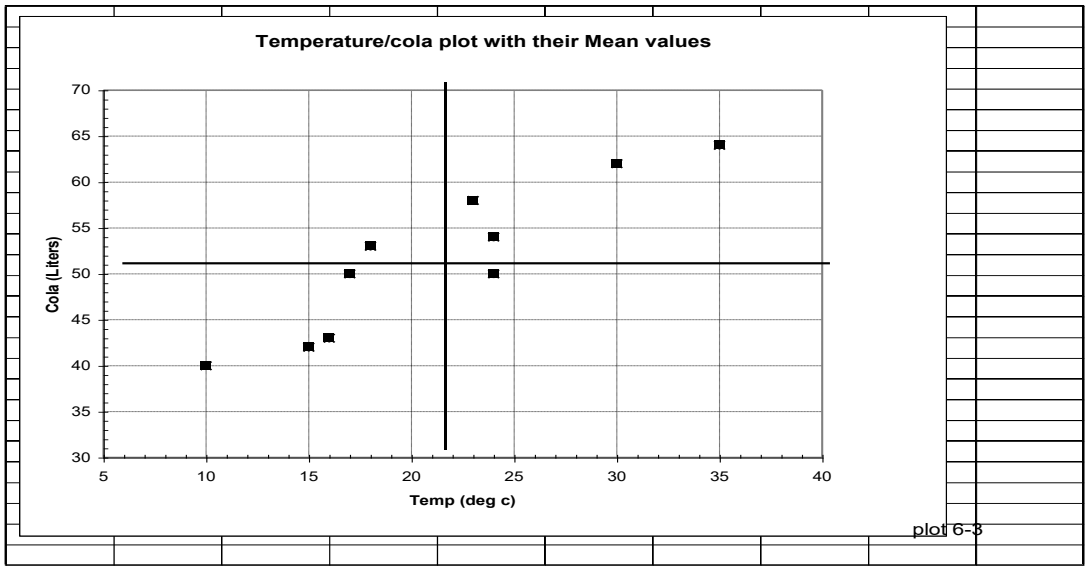
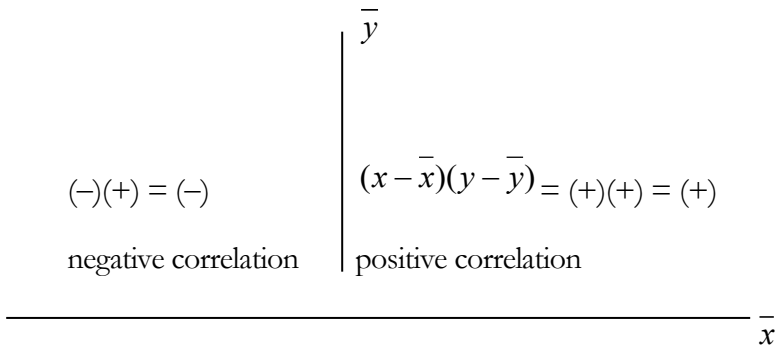


Fig. 4-5 Cola - Temperature graph with superimposed mean values

If for a given data point, say (x, y), we obtain $(x-\bar{x})(y-\bar{y})$ we can see where the data point is located.

If on:



(–)(–) = (+) (+)(–) = (–)

positive correlation negative correlation

So if we then take the contribution of all the data pair points into account the formula becomes:

$\sum (x-\bar{x})(y-\bar{y})$

So if our data points lead to a line with positive slope then the result of this formula would be >0 and if our data points indicate a line with a negative slope then our formula's result would be <0. So perhaps this could be our formula for a numerical correlation figure.

$\sum(x-\bar{x})(y-\bar{y})>0 \Rightarrow$ positive correlation

$\sum(x-\bar{x})(y-\bar{y})<0 \Rightarrow$ negative correlation

$\sum(x-\bar{x})(y-\bar{y})=0 \Rightarrow$ no correlation

But this formula has some drawbacks and needs to be modified. The reasons for amendments are as follows:

- An increase in our sample size, N, would affect the magnitude of this correlation figure without necessarily implying a stronger correlation. So to overcome this problem we modify our formula to:

$$\frac{\sum(x-\bar{x})(y-\bar{y})}{N}$$

The name for this formula is the Covariance of x and y. Therefore,

Cov (x,y) = $\dfrac{\sum(x-\bar{x})(y-\bar{y})}{\mathrm{N}}$.

We however face another problem:

- Our formula $\sum(x-\bar{x})(y-\bar{y})$ for the correlation is also dependent on the units we employ to represent x and y or Temperature and Cola. For example Cola could be represented in Liters, Gallons, Pounds etc., and temperature in degrees Centigrade, Fahrenheit or Kelvin.

This problem can be resolved if we divide a given number (or unit to be precise) by another number that has the same unit(s). In this process the resultant number will have no dimension(s).

In this case we divide our covariance formula by the product "(standard deviation of x) (standard deviation of y)" and obtain the correlation coefficient known as Pearson's Coefficient or simply as the correlation coefficient) denoted by r:

$$r = \frac{\mathrm{Cov}(x,y)}{\sigma_x \sigma_y} = \frac{\frac{1}{\mathrm{N}}\sum(x-\bar{x})(y-\bar{y})}{\sigma_x \sigma_y}.$$

Remember that standard deviation gives a measure of spread of the scattered points about the mean and is given by the formulas:

$$\sigma = \sqrt{\frac{\sum x^2}{\mathrm{N}} - \left(\frac{\sum x}{\mathrm{N}}\right)^2}$$

Note that $\dfrac{\sum x}{\mathrm{N}}$ is the formula for the mean or $\bar{x}$.

Therefore,

$$r = \frac{\frac{1}{\mathrm{N}}\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{\sum x^2}{\mathrm{N}} - \left(\frac{\sum x}{\mathrm{N}}\right)^2}\sqrt{\frac{\sum y^2}{\mathrm{N}} - \left(\frac{\sum y}{\mathrm{N}}\right)^2}} = \frac{\frac{1}{\mathrm{N}}\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\left[\frac{\sum x^2}{\mathrm{N}} - \left(\frac{\sum x}{\mathrm{N}}\right)^2\right]\left[\frac{\sum y^2}{\mathrm{N}} - \left(\frac{\sum y}{\mathrm{N}}\right)^2\right]}}.$$

Multiplying top and bottom by N, we get,

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\mathrm{N}\left[\frac{\sum x^2}{\mathrm{N}} - \left(\frac{\sum x}{\mathrm{N}}\right)^2\right]\mathrm{N}\left[\frac{\sum y^2}{\mathrm{N}} - \left(\frac{\sum y}{\mathrm{N}}\right)^2\right]}} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{\mathrm{N}}\right]\left[\sum y^2 - \frac{(\sum y)^2}{\mathrm{N}}\right]}}.$$

Expanding the numerator and disassociating the Σ,

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

Taking out constants $\bar{x}$ and $\bar{y}$, leads to,

$$r = \frac{\sum xy - \bar{y}\sum x - \bar{x}\sum y + \bar{x}\bar{y}\sum 1}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{N}\right]\left[\sum y^2 - \frac{(\sum y)^2}{N}\right]}} = \frac{\sum xy - \bar{y}\sum x - \bar{x}\sum y + N\bar{x}\bar{y}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{N}\right]\left[\sum y^2 - \frac{(\sum y)^2}{N}\right]}}.$$

By substituting for $\bar{x} = \frac{\sum x}{N}$ and $\bar{y} = \frac{\sum y}{N}$, we can write,

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

After simplifying the correlation coefficient can be expressed as:

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

Example

T(deg C) x	Cola(L) y	x^2	y^2	xy
23	58	529	3364	1334
17	50	289	2500	850
24	54	576	2916	1296
35	64	1225	4096	2240
10	40	100	1600	400
16	43	256	1849	688

15	42	225	1764	630
24	50	576	2500	1200
18	53	324	2809	954
30	62	900	3844	1860
Total:212	516	5000	27242	11452

$$r = \frac{11452 - \frac{212 \times 516}{10}}{\sqrt{\left(5000 - \frac{212^2}{10}\right)\left(27242 - \frac{516^2}{10}\right)}} = \frac{11452 - 10939.2}{\sqrt{(5000 - 4494.4)(27242 - 26625.6)}}$$

$$= \frac{512.8}{\sqrt{505.6 \times 616.4}} = \frac{512.8}{\sqrt{311651.84}} = \frac{512.8}{558.258} = 0.918$$

Interpreting the Coefficient

A value of +1 or −1 for r indicates that the data points lie on a perfect straight line, and a value of zero indicates that there is no correlation. Note that r cannot exceed a value greater than 1 and Always $-1 \leq r \leq 1$.

The second point to be realized is that if our sample number is high then there is a greater chance that a lower correlation value should be acceptable to when our sample numbers are low. Therefore, it is possible to produce a table that gives the minimum value of r for a given sample size.

Fig. 4-6 below gives us the value of r which must be exceeded for a given sample size if we are to be able to deduce the existence of a correlation between x and y which is significant at a given percentage level. 5% significant means that there is a 95 % chance that the correlation exists.

The Correlation Coefficient

The table gives the values of the correlation coefficient for different levels of significance; v = number of pairs in sample (N) −2.

Confidence	%90	%95	%98	%99	%99.9
Significance	0.100	0.0500	0.0200	0.0100	0.00100
v = 1	0.98769	0.99692	0.999597	0.999877	0.9999
2	0.90000	0.95000	0.98000	0.990000	0.9990
3	0.8054	0.8783	0.93433	0.95873	0.9911
4	0.7293	0.8114	0.8822	0.91720	0.9740
5	0.6694	0.7545	0.8329	0.8745	0.9507

6	0.6215	0.7067	0.7887	0.8343	0.9249
7	0.5822	0.6664	0.7498	0.7977	0.8982
8	0.5494	0.6319	0.7155	0.7646	0.8721
9	0.5214	0.6021	0.6851	0.7348	0.8471
10	0.4973	0.5760	0.6581	0.7079	0.8233
11	0.4762	0.5529	0.6339	0.6835	0.8010
12	0.4575	0.5324	0.6120	0.6614	0.7800
13	0.4409	0.5139	0.5923	0.6411	0.7603
14	0.4259	0.4973	0.5742	0.6226	0.7420
15	0.4124	0.4821	0.5577	0.6055	0.7246
16	0.4000	0.4683	0.5425	0.5897	0.7084
17	0.3887	0.4555	0.5285	0.5751	0.6932
18	0.3783	0.4438	0.5155	0.5614	0.6787
19	0.3687	0.4329	0.5034	0.5487	0.6652
20	0.3598	0.4227	0.4921	0.5368	0.6524
25	0.3233	0.3809	0.4451	0.4869	0.5974
30	0.2960	0.3494	0.4093	0.4487	0.5541
35	0.2746	0.3246	0.3810	0.4182	0.5189
40	0.2573	0.3044	0.3578	0.3932	0.4896
45	0.2428	0.2875	0.3384	0.3721	0.4648
50	0.2306	0.2732	0.3218	0.3541	0.4433
60	0.2108	0.2500	0.2948	0.3248	0.4078
70	0.1954	0.2319	0.2737	0.3017	0.3799
80	0.1829	0.2172	0.2565	0.2830	0.3568
90	0.1726	0.2050	0.2422	0.2673	0.3375
100	0.1638	0.1946	0.2301	0.2540	0.3211

Fig. 4-6 values of the correlation coefficient for different levels of significance

In our example of the bar manager we have 10 points so v = 10 –2 = 8. The table tells us that a value not less than ($\mp$) 0.6319 must exist if we are to be 95% certain that a correlation exists between the next day-forecasted temperature and the cola consumption. The reason for + or – sign is that the line may have a positive or negative slope.

Points to note

- The correlation coefficient measures the degree of linear association between two variables. If we found a value of r close to zero this means that a straight line cannot be fitted to the scatter but it does not mean that other relationship that could well model the data points does not exist.

- The correlation coefficient does not explain cause or effect of change x and y. That is to say, it does not answer the question "is x related to y". It simply indicates the direction of change of the two variables with respect to each other. So a degree of correlation does not imply that x depends on y just that the data (measured outcome) could be modelled by a linear relationship. This is referred to as Casualty.

- In our example, the excess consumption could be due to the closure of a pub nearby so that when the weather is hot customers of the closed Bar would come for drinks. So one has to be aware of "spurious or doubtful correlation" and that a third (hidden) factor may be responsible. There may be a number of simultaneous causes.

Important: Note that r for the Temperature - Cola example is = 0.918. Therefore.

$r^2 = 0.84$ - what this means is that 84% of the reason for increased Cola consumption is due to changes in temperature. That is say that only 84% of the variance in y is statistically explained by knowing x. Meaning that 84% of the variation in sales is explained by variation in temperature. Hence, 100–84 = 16% is due to other factors.

A value of r^2 that is equal to 1 means that (r = 1 and) the result is 100% explainable by relationship between the two variables. In our example therefore 84% of the predicted values of cola is a result of the change of temperature and 100–84=16% is due to other factors.

Summary

$$\mathrm{Var}(x) = \frac{\sum (x-\bar{x})(x-\bar{x})}{\mathrm{N}} = \frac{\sum (x-\bar{x})^2}{\mathrm{N}}$$

$$\mathrm{Cov}(x,y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{\mathrm{N}} = \frac{\sum xy}{\mathrm{N}} - \bar{x}\bar{y}$$

$\sum (x-\bar{x})(y-\bar{y})$ (size depends on the number of points i.e the value of N)

$\frac{\sum (x-\bar{x})(y-\bar{y})}{\mathrm{N}}$ (size does not depend on the number of points N but is scale-dependent (depends on the size of the variables)

$r = \frac{\mathrm{Cov}(x,y)}{\sigma_x \sigma_y}$ (size does not depend on the number of points N and is scale-independent).

Tutorial

Example 1 The following table consists of the Come 202 grades of the 23 students in 2000-2001 Semester.

	No.	Attendance	Homework	Final	Overall
1	9706	8	4,67	7	24
2	9719	8,5	11,33	9	40
3	9725	6	4,67	17	35
4	9727	6	5,33	6	32
5	9800	10	17,33	25,5	82
6	9806	10	21,33	30	91
7	9810	7	3,33	21	42
8	9811	7	17,33	18,5	63
9	9820	7,5	10,67	10	42
10	9822	7	18	9,5	52
11	9825	8	16,67	21	66
12	9826	5	0	0,5	1
13	9827	11	19,33	37	100
14	9836	8,5	17,33	26,5	75
15	9837	5,5	3,33	5,5	21
16	9842	6,5	4	1	10
17	9845	7	6,67	8	36
18	9850	10	21,33	27	85
19	9854	7,5	17,33	21	69
20	9856	7,5	11,33	17	44
21	9865	6	0	8,5	24
22	9901	10	21,33	33,5	98
23	9917	9	18,67	30,5	79
	Avr.	7,76	11,796	16,97	53

1. What are the standard deviations of the last four columns of the table?
2. What is the correlation between the students' attendances and their final grades?
3. What is the correlation between the students' attendances and their overall grades?
4. What is the correlation between the students' homework and overall grades?
5. What is the correlation between the students' final and overall grades?

Discuss the results.

Chapter

6

Regression

Principle Objectives

- *To find the form of the relationship between two variables.*
- *To derive expressions for Normal Equations.*
- *To carry out regression analysis.*
- *To analysis the implications for pair of regression lines.*
- *To consider non-linear cases.*

Introduction to Regression

The method used to find the form of the relationship between two variables is called Regression Analysis. Once the form of a relationship is established it is then possible to predict or extrapolate (near) future outcomes based on the past data as long as the patterns of the future events follow the past. Note that variables could be related to each other either functionally or statistically. For example if y = m . x + c represent the equation of a straight line, we can say the y is function of x as long as we know the values of m and c. That is to say, we can determine the exact value of y if we are given the value of x.

Remember that the correlation coefficient was expressed as:

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

Regression Analysis

This is the name given to a technique used to find a mathematical function to represent the data points. These points could be closely correlated or not as the case may be but correlation only represents the accuracy of the predictability of the model.

So if it is thought that a relationship between the data points exists then it is useful to develop a "predictive" function expressing their relationship. Here we will limit ourselves to two variables linear regressing. Therefore note that there could be more two variables involved and that the relationship may not be linear.

If we look at the scatter plot of the temperature-cola (of chapter 4) we could perhaps draw a number of lines through the points to approximately give us what looks like the best straight line but how do we know that we have drawn the best line. The method used to find the best line fit is called the Least Squares (Minimum Square error). In this method one obtains the error or the difference between the y and the y of the best line. Since this error could be positive or negative one then squares it to convert all the errors to have a positive sign. The method then sums all the errors and finds the best estimate of m and c of a straight line that minimises this total error.

When all mathematical manipulations have been done the formula for the best straight-line estimate through the data points is given by:

$$y = mx + c$$

Where,

$$m = \frac{\sum xy - \bar{x}\sum y}{\sum x^2 - \bar{x}\sum x} \quad \text{and}$$

$$c = \bar{y} - m\bar{x}.$$

Full verification of the above formulae will be given in pages 41-43.

Let's find the best line fit for the bar managers scattered points:

Example

T(deg C) x	Cola(L) y	x^2	xy
23	58	529	1334
17	50	289	850
24	54	576	1296
35	64	1225	2240

10	40	100	400
16	43	256	688
15	42	225	630
24	50	576	1200
18	53	324	954
30	62	900	1860
Total: 212	516	5000	11452

$$m = \frac{11452 - 21.2 \times 516}{5000 - 21.2 \times 212} = 1.01$$ and

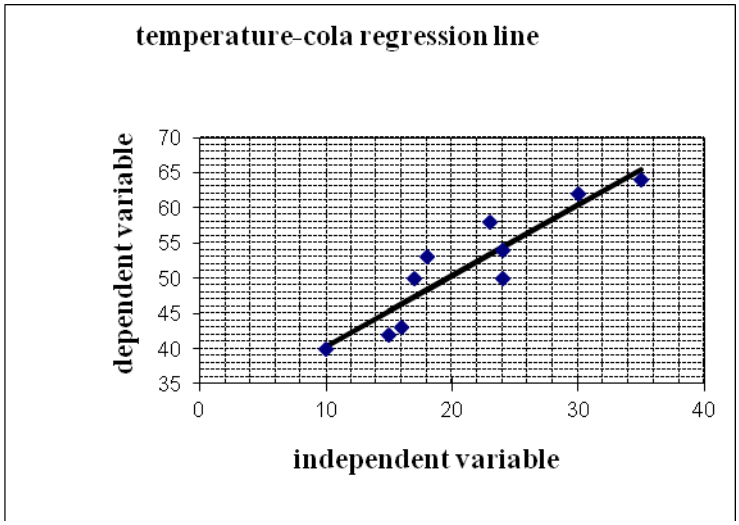
$$c = \bar{y} - m\bar{x} = 51.6 - 1.01 \times 21.2 = 30.19$$

Hence, we have the regression line

$$y = mx + c = 1.01x + 30.19$$.

Temp (C)	Cola (L)	Cola (L)	Error
X	y	y estimate	y−y estimate
23	58	53.42	4.58
17	50	47.36	2.64
24	54	54.43	-0.43
35	64	65.54	-1.54
10	40	40.29	-0.29
16	43	46.35	-3.35
15	42	45.34	-3.34
24	50	54.43	-4.43
18	53	48.37	4.63
30	62	60.49	1.51

Regression plot Temperature vs. Cola:



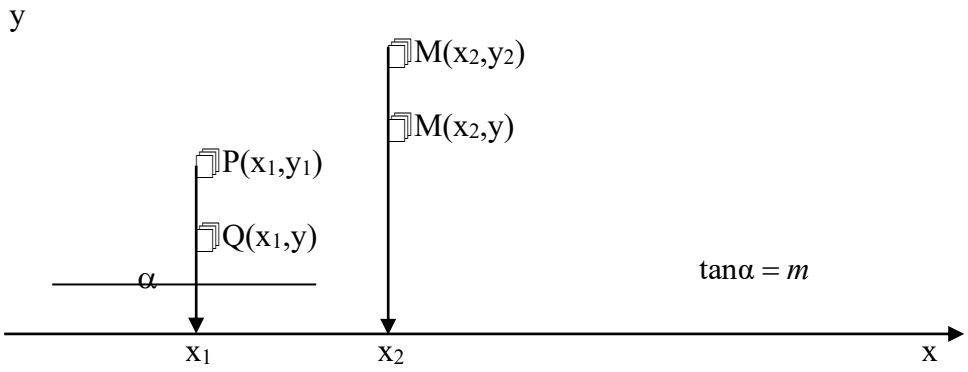
Remarks

We constructed our regression line from a data set. This line is one amongst many possible lines so a line distribution is possible where our regression line is located between the best and worst fit line. For a given value of x therefore we could construct a distribution for y using these differing lines. Therefore, it is unlikely that all the variation of y obtained by the use of our regression equation is due to variation of x (unless our line is the best line).

When we use this equation to estimate a particular outcome say amount of cola at temperature 34 ^{o}C, part of the answer, that is the litters of cola, is therefore due to change of the temperature (this can be called the explained variation) and part is due to other factors (unexplained variation).

The closeness of the relationship between x and y could be indicated by the Coefficient of Determination. This coefficient is simply the square of the coefficient of correlation (i.e. r^2) and measures the proportion of the total variation that can be explained. Note that r squared should not be taken as a statement of causation i.e. the cause of increase drinking is higher temperature, but that a statistical relationship exists between the two variables and that part of the variation can be explained and part of it cannot be explained.

Best line of fit - Slope of Regression Line



Consider the n points (x_1,y_1), (x_2,y_2), ….. ,(x_n,y_n) shown in above figure. To find an accurate best line through the points (for y on x regression line), we can take the equation of the line as $y = mx + c$. Where m and c are unknowns and can be determined as sown below.

Note that errors are represented by PQ, MN and so forth. These errors can be written as:

$$\text{PQ} = y - y_{estimate} = y_1 - (mx_1 + c) \quad \text{(positive)}$$
$$\text{MN} = y - y_{estimate} = y_2 - (mx_2 + c) \quad \text{(negative)}$$

Therefore, it can be safely concluded that the general expression for the error is:

$$y_i - (mx_i + c)$$

To remove the sign of the error (as we are only interested in the magnitude of the errors) the error terms can be squares and then we find the sum of the errors as shown below:

$$s \equiv \sum(\text{error})^2 = \sum_{i=1}^{n}(y_i - mx_i - c)^2$$

m and c are variable in the above formula and y_i and x_i the constants.

For s to be minimum, therefore,

$$\frac{\partial s}{\partial m} = \frac{\partial s}{\partial c} = 0$$

$$\frac{\partial s}{\partial m} = \sum_{i=1}^{n} 2(y_i - mx_i - c)(-x_i) = 2\sum_{i=1}^{n}(y_i - mx_i - c)(-x_i) = 0$$

$$\sum(-x_i y_i + mx_i^2 + cx_i) = 0$$

Generalising, therefore leads to,

$$\sum xy = m\sum x^2 + c\sum x \qquad (1)$$

Similarly,

$$\frac{\partial s}{\partial c} = \sum 2(-1)(y_i - mx_i - c) = -2\sum(y_i - mx_i - c) = 0$$

$$\sum y - m\sum x - nc = 0$$

Therefore,

$$\sum y = m\sum x + nc \qquad (2)$$

Equations (1) and (2) are known as the Normal Equations. Dividing equation (2) by n, gives

$$\frac{\sum y}{n} = \frac{m\sum x}{n} + c \quad \text{or} \quad \bar{y} = m\bar{x} + c.$$

$\therefore$ The point $(\bar{x}, \bar{y})$ lies just on regression line. $c = \bar{y} - m\bar{x}.$

Derivation of m and c:

Substituting $c = \bar{y} - m\bar{x}$ in equation 1 gives,

$$\sum xy = m\sum x^2 + c\sum x = m\sum x^2 + (\bar{y} - m\bar{x})\sum x = m(\sum x^2 - \bar{x}\sum x) + \bar{y}\sum x$$

$$\boldsymbol{m} = \frac{\sum xy - \bar{y}\sum x}{\sum x^2 - \bar{x}\sum x} = \frac{\sum xy - \frac{\sum y \sum x}{n}}{\sum x^2 - \bar{x}\sum x} = \frac{\sum \boldsymbol{xy} - \bar{\boldsymbol{x}}\sum \boldsymbol{y}}{\sum \boldsymbol{x^2} - \bar{\boldsymbol{x}}\sum \boldsymbol{x}} = \frac{\mathrm{Cov}(x, y)}{\sigma_x^2}.$$

Now since m is known, c can also be found as shown below?

$$\boldsymbol{c} = \bar{\boldsymbol{y}} - \boldsymbol{m}\bar{\boldsymbol{x}}.$$

Pair of regression lines

The formula used to determine a value of x from a given value of y is different from the formula used to determine y from given value x. To see why this is so, consider the heights and weights of 10 men given below:

Heights(cm)x	160	160	165	170	170	170	170	175	180	180
Weights(kg)y	55	55	55	55	55	60	70	65	70	80

$$r = \frac{\sum xy - \frac{\sum x \sum y}{N}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{N}\right]\left[\sum y^2 - \frac{(\sum y)^2}{N}\right]}} = \frac{105850 - \frac{1700 \times 620}{10}}{\sqrt{\left(289450 - \frac{1700 \times 1700}{10}\right)\left(39150 - \frac{620 \times 620}{10}\right)}}$$

r = 0.796.

Observe that since r is symmetric in x and y. Using the Normal Equations, we can find m and c as follows:

$$m = \frac{\sum xy - \bar{x}\sum y}{\sum x^2 - \bar{x}\sum x} = \frac{105850 - 170 \times 620}{289450 - 170 \times 1700} = \frac{450}{450} = 1$$

and

$$c = \bar{y} - m\bar{x} = 62 - 1 \times 170 = -108.$$

So the equation of the regression line (Weight on Height) is,

$$y = mx + c = x - 108.$$

If weights were taken as x values and heights as y values, what would be the values of m and c?

$$m = \frac{\sum xy - \bar{x}\sum y}{\sum x^2 - \bar{x}\sum x} = \frac{105850 - 62 \times 1700}{39150 - 62 \times 620} = \frac{450}{710} = 0.634$$

and

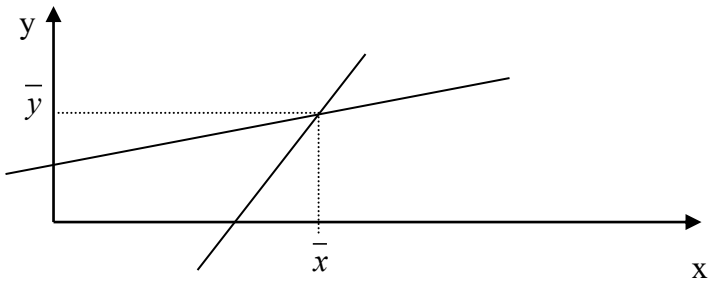
$$c = \bar{y} - m\bar{x} = 170 - 0.634 \times 62 = 130.704$$

So the equation of the regression line (Height of weight) is,

$$y = mx + c = 0.634x + 130.704.$$

Therefore, the line for y on x as shown is different to x on y line.

It should be noted that the point (170,62) [or (62,170) according to the chosen reference point] is on both of the regression lines (rounding may cause a little distortion). This elucidates that the two lines intersect at x and y mean values.



Chapter

7

Time Series Forecasting

Principle Objectives

- *To understand the concept of time series.*
- *To derive an expressions for time series.*
- *To carry out time series analysis.*
- *To project trend for a given variable.*
- *To consider non-linear cases.*

Introduction

In the preceding chapters we discussed the topic of regression analysis as a tool for prediction. In this respect, regression analysis provides a useful guide to managerial decision-making. In this Chapter we develop the concept of Time-Series analysis and demonstrate how forecasting methods in business assist in the process of managerial planning and control.

A time series is a set of numerical data that is obtained at regular periods over time. Quarterly production figures, annual sales, etc. are examples of Time Series. Time Series is useful for systematic movements of technical and business data that look random but needed for making decisions. Time Series forecasting methods involve the projection of future values of a variable based entirely on the past and present observations of that variable.

There are four underlying influences or components making up time series data:

1) Trends (T): Overall or persistent, long term upward or downward pattern of movement.

Reason for influence: Changes in technology, population, wealth and value.

Duration: Several years.

2) Seasonal variation (S): Fairly regular periodic fluctuations that occur within each 12-month period year after year.

Reason for influence: Weather conditions, social customs, and religious customs. Duration: Within 12 months (or monthly or quarterly data).

3) Cyclical variation (C): Repeating up-and-down swings or movements through four phases: from peak (prosperity) to contraction (recession) to trough (depression) to expansion (recovery or growth).

Reason for influence: Interactions of numerous combinations of factors influencing the economy.

Duration: Usually 2–10 years with differing intensity for a complete cycle.

4) Random or Irregular (R): The erratic or "residual" fluctuations in a series that exist after taking into account the systematic effects –trend, seasonal, and cyclical.

Reason for influence: Random variations in data or due to unforeseen events such as strikes, hurricanes, floods, political assassinations, etc.

Duration: Short duration and non-repeating.

Note that seasonal variation and cyclical variation are almost same in Engineering, but different in business: just consider the increase in sales during the Christmas periods.

Time series data (Y), observed data values, can be represented by:

1. Additive Model $Y = T + S + C + R$
2. Multiplicative Model $Y = T.S.C.R.$

In the Multiplicative Model any observed value in a Time Series is the product of these influencing factors; that is, when the data are obtained annually, an observation Y_i recorded in the year i may be expressed as in equation $Y = T_i.C_i.R_i$. When the data values are obtained either quarterly or monthly, an observation Y_i recorded in time period i may be expressed as in equation $Y = T_i.S_i.C_i.R_i$ where T_i, C_i, and R_i are the values of the trend, cyclical, and random components, respectively, in time period i, and S_i is the seasonal component in time period i.

The Multiplicative Model is most appropriate and easier model to use. To this end, the Additive Model is left for future considerations.

Smoothing the Annual Time Series

When we examine annual data, our visual impression of the overall long-term tendencies or trend movements in the series is obscured by the amount of variation from year to year. It then becomes difficult to judge whether any long-term upward or downward trend effect really exists in the series. In such situations, the Moving Averages Method may be used to smooth a series and thereby provide us with an overall impression of the pattern of movement in the data over time. Moving Averages for a chosen period of length consist of a series of arithmetic means computed over time

such that each mean is calculated for a sequence of observed values having that particular length. The method of moving averages is used only for smoothing and is used when variations are not nearly linear.

Example 1 - (Moving Averages):

Year/Quarter		Petrol Sales (Seasonally unadjusted y	Centred Moving Averages T_y	Ratio (y/Ty) ×100 %	seasonal index S %	Petrol sales (seasonally adjusted)
1999	Q1	9	–	–	88	10
	Q2	11	–	–	109	10
	Q3	15	12	125	115	13
	Q4	12	13.125	91	88	14
2000	Q1	11	14.625	75	88	13
	Q2	18	15.75	114	109	17
	Q3	20	18	111	115	17
	Q4	16	21.25	75	88	18
2001	Q1	25	24.25	103	88	28
	Q2	30	27.5	109	109	28
	Q3	32	29.625	108	115	28
	Q4	30	30.625	98	88	34
2002	Q1	28	32.25	87	88	32
	Q2	35	34	103	109	32
	Q3	40	–	–	115	35
	Q4	36	–	–	88	41

Four Quarter Moving Averages

				Trend Value of Petrol Sales Centred Moving Averages
1999	Q1	9		–
	Q2	11		–
			(9+11+15+12)/4=11.75	
	Q3	15		(11.75+12.25)/2=12
			(11+15+12+11)/4=12.25	
	Q4	12		(12.25+14)/2=13.125
			(15+12+11+18)/4=14	
2000	Q1	11		14.625
	Q2	18		15.75

Calculation of seasonal index: Ratios ($\frac{y}{T_y} \times 100$)

Year	Q1	Q2	Q3	Q4
1999	–	–	125	91
2000	75	114	111	75
2001	103	109	108	98
2002	87	103	–	–

Average Seasonal Effect:

$$\frac{75+103+87}{3} \quad \frac{114+109+103}{3} \quad \frac{125+111+108}{3} \quad \frac{91+75+98}{3}$$

$$=88(\text{rounded}) \quad =109(\text{rounded}) \quad =115(\text{rounded}) \quad =88$$

*The four quarterly averages should sum to 400. Do you know why?

Example 2 - (Moving Averages): The following data represent the annual sales (in millions of constant 2002 dollars) for a food-processing company for the years 1990–1997.

Year	Coded Year	Sales Y	Centred Moving Averages T_y	Ratio (y/Ty)×100 %	Sales (adjusted)
1995	0	36.4			
1996	1	38.4			
1997	2	42.6			
1998	3	34.8			
1999	4	28.4			
2000	5	23.9			
2001	6	27.8			
2002	7	42.1			

a) Plot the data on a chart.
b) Fit a 4-year (Centred) Moving Average to the data and plot the results on your chart.

Regression and Forecasting

The component factor of a time series most often studied is trend. We study trend for predictive purposes; that is, we study trend as an aid in making intermediate and long-range forecasting projections. The linear trend method is used when variations are nearly linear. Hence regression line method is applied here for measurement of trend (T).

When using the regression line method for fitting trends in time series, our interpretation of the coefficients is simplified if we code the X values so that the first observation in our time series is selected as the origin and assigned a code value of X=0. All successive observations are then assigned consecutively increasing integer codes: 1,2,3…

Model Selection - First and Second Differences

1) If a linear trend model were to perfectly fit a time series, then the first differences would be constant. That is, the differences between consecutive observations in the series would be the same throughout.

$Y_2 - Y_1 = Y_3 - Y_2 = \ldots = Y_n - Y_{n-1}$.

2) If a quadratic trend model were to perfectly fit a time series, then the second differences would be constant. That is,

$(Y_3 - Y_2) - (Y_2 - Y_1) = (Y_4 - Y_3) - (Y_3 - Y_2) = \ldots = (Y_n - Y_{n-1}) - (Y_{n-1} - Y_{n-2})$.

Measurement of Seasonal variation (S):

Business data are usually presented in two forms:

1) Seasonally adjusted
2) Seasonally unadjusted.

The purpose of seasonal adjustments is to allow the underlying movements in data series (y) to be revealed, undistorted by any seasonal variation.

Measurement of Cyclical Variation (C):

Series data are affected by periodic upturns and downturns. These variations are referred to as cyclical (Business cycles, trade cycles, etc.) and hence only of interest for long term planning. These variations are only of interest to government and economists and not to businesses. Energy sector for instance is an exception since long term cyclical variations are important.

Measurement of Random Variation(R):

The term random variation is used to cover all types of variation other than trend, seasonal and cyclical movements. It covers all unpredictable movements such as the effects of strikes, natural disasters such as floods and famines, the effects of the Stock Exchange Crash in October 1987, etc. It can be measured as the residual after the regular (trend, seasonal and cyclical) factors have been removed from a series, but there is a little or no interest in such residuals in their own right and we devote no further attention to them.

To establish cyclical variation the following analysis is carried out:

1. Find the trend values of the observed y data, i.e. T (using one of the methods explained earlier)

2. "De-trend" the data i.e. calculate $\frac{y}{T}$.

3. Where appropriate, "de-seasonalise" the series by dividing by the seasonal index values (S) – normally this is not necessary because annual data only would be used for cyclical analysis.

4. Finally, remove as much of the random variation (R) as possible by a process of smoothing. This is achieved by taking moving averages.

Tutorial

Example 1 (regression line method): The following table shows the number of cars sold for the years 1981–1989. Find the regression line.

Year	Time Period (x)	No. of cars (000) (y)	x^2	xy
1981	1	21	1	21
1982	2	25	4	50
1983	3	24	9	72
1984	4	29	16	116
1985	5	35	25	175
1986	6	38	36	228
1987	7	40	49	280
1988	8	52	64	416
1989	9	63	81	567
Total:	45	327	285	1925

$$m = \frac{\sum xy - \bar{x}\sum y}{\sum x^2 - \bar{x}\sum x} = \frac{1925 - 5 \times 327}{285 - 5 \times 45} = 4.83 \text{ and}$$

$$c = \bar{y} - m\bar{x} = 36.3 - 4.83 \times 5 = 12.15$$

So the equation of the regression (trend) line is $y = mx + c = 4.83x + 12.15$. Now, regression (trend) line $\mathrm{T}y = 4.83x + 12.15$ is used in forecasting. In this example, our initial code value (time period) 1 can be corrected to 0.

Example 2 (regression line method): The following table shows the revenues for Eastman Kodak Imaging Company for the years 1984–1996.

Year	Coded Year	Revenue
1984	0	10.6
1985	1	10.6
1986	2	11.5
1987	3	13.3
1988	4	17
1989	5	18.4
1990	6	18.9
1991	7	19.4
1992	8	20.2
1993	9	16.3
1994	10	13.7
1995	11	15.3
1996	12	16.2

- Find the regression line equation.
- What are the coded years for the years 2003, 2004 and 2005.
- Forecast the revenue of the company for the years 2003, 2004 and 2005.

Chapter

8

Quality Management

Principle Objectives

- *To define quality terms*
- *To understand dimensions of quality and its grades*
- *To analyse the concepts of Total Quality Management.*
- *To appreciate the cost of non-conformance.*
- *To identify the component of ISO 9000 for specific applications.*

Introduction

Quality as defined by the Oxford concise dictionary is "the degree of goodness or worth". However, quality is understood by many professional managers to be "fitness for purpose". The following two definitions together offer an understanding of what quality means:

- Quality is a measure of conformance of a product or service to certain specifications or standards, fitness for use.
- Quality is about adding value for the customer. This suggests that non-value adding activities are a legitimate quality target. Quality therefore is not a separate activity; it concerns all activities in an organization.

While quality is sometimes defined in absolute terms in business and industrial activities, the following meanings are used:

- Fitness for purpose
- Conformance with stated requirements/specifications

- ❖ Satisfying a particular need or function

Quality terms:

Quality dimensions, Quality assurance, Quality Control, Quality Management, TQM, Quality Costs, Quality Systems, BS5750/EN29000, ISO 9000.

Dimensions of Quality

- Performance For a car, fuel consumption or acceleration

For a hotel or a restaurant prompt service.

- Features These are often elements that the customers do not expect but are highly observable, i.e. complimentary newspapers for hotel guests, immobilizer in a new car given to the customer as a gift.

- Reliability To ensure that the product functions effectively - for a car, that it starts every time and for a restaurant, that it can provide the full service on the menu every time.

- Conformance This refers to the design specifications and functional characteristics. The product or a service is expected to conform to a given set of standards.

- Durability This is the period which a product or service is expected to last without any deterioration of performance.

- Serviceability The product and service can be serviced with ease, and there is a prompt, courteous and efficient customer support.

- Aesthetics Aesthetics concern product looks, feels, sounds, tastes or smells. These usually relate to personal judgment or preference.

Quality Control is the function of testing, inspecting, and taking corrective and preventive action in order to maintain quality standards.

Quality Assurance is the system of policies and procedures established by an organization in order to achieve and maintain quality.

Quality Engineering is the process of introducing quality into design and anticipating quality problems prior to production stages.

Total Quality Management (TQM)

TQM is a strategy to improve quality within an entire organization. To implement TQM the senior management must determine quality priorities and establish systems for:

1. Quality management practice

2. Procedures to be followed

3. Allocation of resources

As stated, TQM is a strategy/philosophy hence is not a system or a technique to be defined. However, TQM incorporates the following:

- An integrated managerial approach focusing on the needs of the customer.
- A quality plan offering a disciplined and structured approach to quality improvement.
- A set of procedures (sometimes standardized), techniques and tools.
- A system of collecting and analyzing of information.
- Employee participation and team work approach and encouragement of creative thinking.
- Continuous improvement arising from having a closed loop from quality planning through into execution and back to planning. This requires an integrated data approach that can support work teams. It is important that quality plans include measures and performance targets which are meaningful to managers and operators in such a way that they focus on causes rather than faults.
- Costs of poor quality. Quality is about adding value for the customer and the quality of normal business process should be considered important. The support for quality has to be imbedded in quality systems supporting a TQM approach/strategy.

The Characteristics of a TQM

The three main characteristics of a TQM approach are as follows:

- Focus on conformance i.e. the production (or service) process produces components of finished goods (or service) which conform to the specifications.
- Has a system referring to the original design which the customer had expressed a preference for.

Organization culture which, encourages initiative and creative thinking and where responsibility is localized throughout the whole organization.

Establishing TQM

Step1 Establish the TQM management and cultural environment

- Vision
- Long term commitment
- People Involvement
- Disciplined Methodology

- ❖ Support Systems
- ❖ Training

Step 2 Define mission of each component of the organization

Step 3 Set performance improvement opportunities, goals and priorities

Step 4 Establish improvement projects and action plans

Step 5 Implement projects using improvement methodologies

Step 6 Evaluate

Step 7 Review and cycle (improved performance)

- ❖ Reduce cycle time
- ❖ Lower cost
- ❖ Innovation

Gurus of quality

Edward Deeming - worked for several years in USA and Japan. His work primarily concerned the following:

- ❖ Continuous improvement of production processes.
- ❖ Encouraging employment commitment (employee participation)
- ❖ Aiming for highest standards (customer focus)

Deming Vision ("out of the crisis")

1. Create constancy of purpose toward improvement of product and service, with the aim to become competitive and to stay in business.
2. Adopt the new philosophy. We are in a new economic age. Western management must awaken to the challenge, must learn its responsibilities, and take on leadership for change.
3. Drive out fear so that everyone may work effectively for the company.
4. End the practice of awarding business on the basis of price tags. Instead minimize total cost. Move toward a single supplier for any one item, on a long-term relationship of loyalty and trust. This does not mean that there must always be sole sourcing. Three vendors instead of twelve may be the answer.
5. Improve constantly and forever the system of production and service, to improve quality and productivity, and thus constantly decrease costs.

6. Institute training on the job. (Never use OJT as a sole training tool.)

7. Institute leadership. The aim of supervision should be to help people and machines and gadgets do a better job. Supervision of management is in need of an overhaul, as well as supervision of production workers.

8. Cease dependence on inspection to achieve quality. Eliminate the need for inspection on a mass basis by building quality into the product in the first place. No after-the-fact acceptance sampling.

9. Break down barriers between departments. People in research, design, sales, and production must work as a team to foresee production and use problems.

10. Eliminate slogans, exhortations, and targets for the work force asking for zero defects and new levels of productivity. Such exhortations only create adversarial relationships, as the bulk of the causes of low quality and low productivity belong to the system and thus lie beyond the power of the work force.

11. a) Eliminate work standards (quotas) on the factory floor. Substitute leadership.
    b) Eliminate management by objectives. Eliminate management by the numbers, numerical goals. Substitute leadership.

12. Remove barriers that rob the hourly worker of his or her right to pride of workmanship. The responsibility of supervisors must be changed from sheer numbers to quality.

13. Institute a vigorous program of training and retraining.

14. Create a top management structure to accomplish the transformation.

Joseph Juran - he made major contributions in improving management practices. He believed management is responsible for 80% of failures. His solutions were always long term.

Kaoru Ishikawa - he encouraged the instigation and development of quality circles. He also introduced the concept of just in time (JIT) in many manufacturing companies.

Cost of Quality

Prevention costs - quality assurance costs associated with planning and implementing the program.

Appraisal costs - direct costs of measuring quality.

Internal failure costs - costs associated with defects found during quality inspection.

External failure costs - costs associated with defective products sold to consumers.

Cost of getting it right	Vs	Cost of getting it wrong
(Conformance) e.g. prevention,		(Non-conformance) e.g. Failure costs

Appraisal costs, etc. internal, external costs, etc.

Components of Quality Manual

1. Quality Manual, Policy, Objectives, QA principles, Organization Outlines, Procedures Outlines, Standard Procedures Index
2. General Standard Procedures
3. Nonstandard Procedures

Standardize Quality Systems

- ISO 9000 Part 1-3
- EN 29000 Part 1-3
- BS 5750 Part 1-3

Part 1: contains 20 items

Part 2: contains 18 items

Part 3: contains 12 items

Proposal Analysis – Purchasing (Supply Chains)

1) Conventional method:

The most popular method of selecting a supplier is by competitive bidding –i.e. Competitive on price therefore wide supplier base (no commitment to a few quality suppliers) but note that the flexibility is greater.

2) Quality Approach:

How suppliers intend to comply with quality requirements

Or

State explicitly how they plan to achieve quality level with consistency

Capability Survey:

The purchasing and supply personnel of a buyer firm usually carry out a survey of a company wishing to supply it. The survey considers the following:

- Supply quality policy
- Attitude (Quality Circle, for instance)
- Design of experiments to reduce process variation

- Experience (production/service)
- Quality assurance/Quality control: Technique, tools, etc.
- Repeating structure

Chapter

9

Quality Tools

Principle Objectives

- *To understand how Tool Selector Chart works*
- *To discuss the applications of the PDCA Cycle*
- *To apply a Pareto Chart in a real situation*
- *To understand the how Cause & Effect (Fishbone) Diagram is constructed*
- *To construct and apply Gantt chart*
- *To apply Control charts*

Tool Selector Chart

This chart organises the tools by typical improvement situations, such as working with numbers, with ideas, or in teams. Abbreviations in the following chart are given below the chart.

Working with Ideas	Generating/Grouping	Deciding	Implementing
AND			●
Affinity	●		
Brainstorming	●		
C & E / Fishbone	●	●	
Flowchart	●	●	●
Force Field	●	●	
Gantt			●
ID	●	●	
Matrix			●
NGT/Multivoting		●	
Prioritisation		●	
PDPC			●
Radar		●	
Tree	●		●

AND ≡ Activity Network Diagram

Scheduling sequential and simultaneous tasks to find the most efficient and realistic path.

Affinity ≡ Gathering and Grouping LDEA to allow a team to creatively generate a large number of ideas/issues and form national groupings.

Force Field ≡ Positives and negatives of change –to identify pros and cons of a solution.

ID ≡ Interrelationship Diagraph

Looking for drivers and outcomes –to identify analyse and classify. C&E.

Gantt ≡ A simple tool that uses horizontal bars to show which tasks can be done simultaneously – A good scheduling/monitoring tool for projects.

NGT ≡ Nominal Group Technique

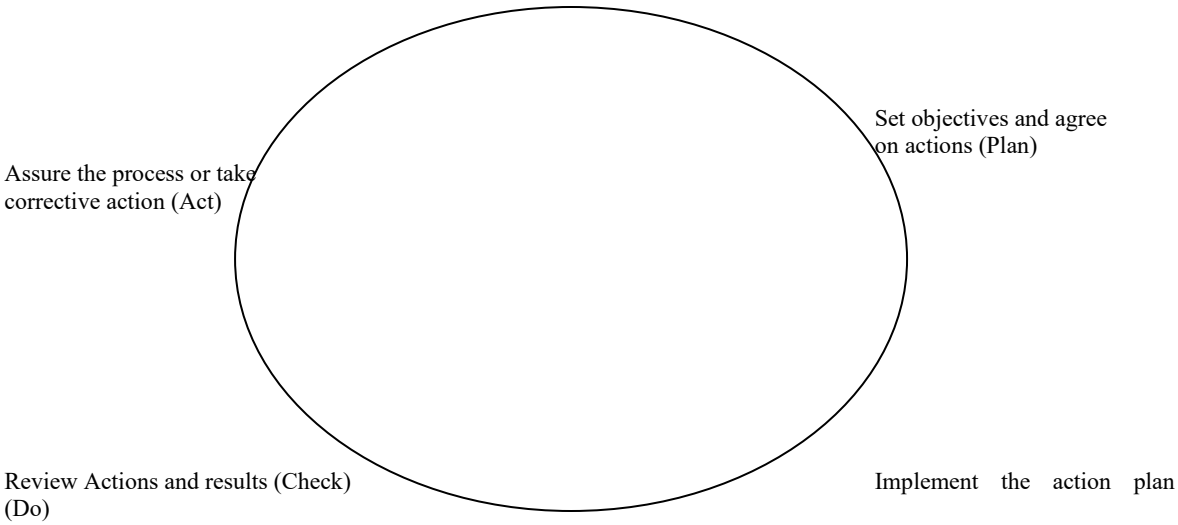
Ranking for consensus to form a consensus on relative importance of issues, problem, solutions, etc. –useful technique for evaluating responses to a questionnaire.

Radar Chart ≡ Rating organization performance issues a single graphics where the current performance is measured against ideal.

Tree Diagram ≡ Mapping the tasks for implementation used to break a goal into small elements for implementation (i.e. means or steps needed).

PDCA Cycle

The PDCA (Plan, Do, Check, Act) Cycle



P Plan where you want to be and what you need to do to get there

D Do what you planned on doing

C Check to see if the objective was achieved

A Act on the information. Insofar as you were successful, standardize; where

you see room for improvement, recycle.

Pareto Chart

Why use it?

To focus efforts on the problems that offer the greatest potential for improvement by showing their relative frequency or size in a descending bar graph.

What does it do?

- ❖ Helps a team to focus on those causes that will have the greatest impact if solved.
- ❖ Based on the proven Pareto principle: 20% of the sources cause 80% of the problem.
- ❖ Displays the relative importance of problems in a simple, quickly interpreted, visual format.
- ❖ Helps prevent "shifting the problem" where the "solution" removes some causes but worsens other.
- ❖ Progress is measured in a highly visible format that provides incentive to push on for more improvement.

How do I do it?

1. Decide which problem you want to know more about.

Example: Consider the case of HOTrep, an internal computer network help line: Why do people call the HOTrep help line; what problems are people having?

2. Choose the causes or problems that will be monitored, compared, and rank ordered by brainstorming or with existing data.

   (a) Brainstorming

Example: What are typical problems that users ask about on the HOTrep help line?

   (b) Based on existing data.

Example: What problems in the last month have users called in to the HOTrep help line?

3. Choose the most meaningful unit of measurement such as frequency or cost.

- ❖ Sometimes you don't know before the study which unit of measurement is best. Be prepared to do both frequency and cost.

Example: For the HOTrep data the most important measure is frequency because the project team can use the information to simplify software, improve documentation or training, or solve bigger system problems.

4. Choose the time period for the study.

   - ❖ Choose a time period that is long enough to represent the situation. Longer studies don't always translate to better information. Look first at volume and variety within the data.

   - ❖ Make sure the scheduled time is typical in order to take into account seasonality or even different patterns within a given day or week.

Example: Review HOTrep help line calls for 10 weeks (May 22 – August 4)

5. Gather the necessary data on each problem category either by "real time" or reviewing historical data.

   - ❖ Whether data is gathered in "real time" or historically, check sheets are the easiest method for collecting data.

Example: Gathered HOTrep help line calls data based on the review of incident reports (historical).

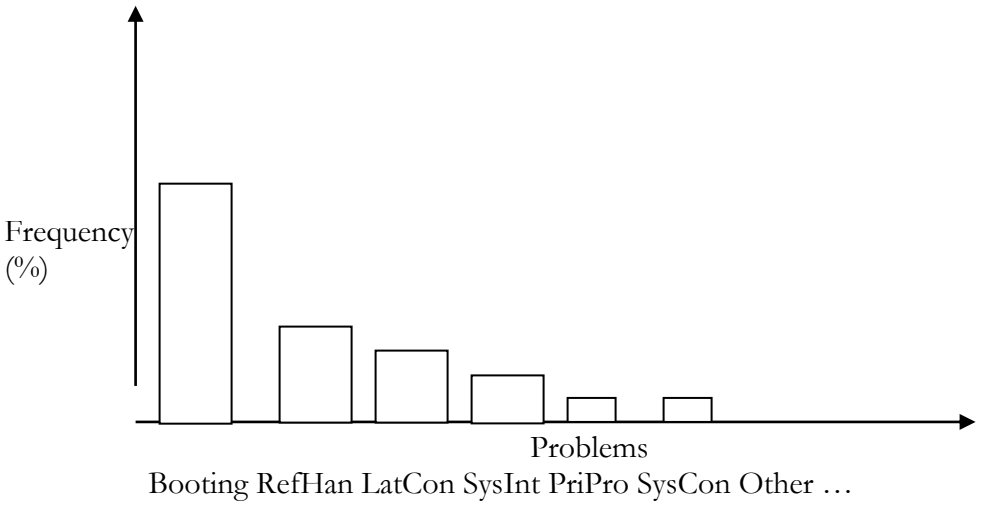
Tip Always include with the source data and the final chart, the identifiers that indicate the source, location, and time period covered.

6. Compare the relative frequency or cost of each problem category.

Example (Pareto Chart):

Problem Category	Frequency	Percent (%)
Bad configuration	3	1
Boot problems	68	33
File problems	8	4
Late connection	20	10
Print problems	16	8
Reflection hang	24	12
Reflection sys. integrity	11	5
Reflection misc.	6	3
System configuration	16	8
System integrity	19	9
Others	15	7
Total	206	100

A Pareto chart



7. List the problem categories on the horizontal line and frequencies on the vertical line.

   - ❖ List the categories in descending order from left to right on the horizontal line with bars above each problem category to indicate its frequency or cost. List the unit of measure on the vertical line.

8. (Optional) Draw the cumulative percentage line showing the portion of the total that each problem category represents.

   (a) On the vertical line, (opposite the raw data, #, $, etc.), record 100% opposite the total number and 50% at the halfway point. Fill in the remaining percentages drawn to scale.

   (b) Starting with the highest problem category, draw a dot or mark an x at the upper right-hand corner of the bar.

Add the total of the next problem category to the first and draw a dot above that bar showing both the cumulative number and percentage. Connect the dots and record the remaining cumulative totals until 100% is reached.

9. Interpret the results

   - ❖ Generally, the tallest bars indicate the biggest contributors to the overall problem. Dealing with these problem categories first therefore makes common sense. But, the most frequent or expensive is not always the most important. Always ask: What has the most impact on the goals of our business and customers?

Variations

The Pareto Chart is one of the most widely and creatively used improvement tool. The variations used most frequently are:

- ❖ Major Cause Breakdowns in which the "tallest bar" is broken into sub-causes in a second, linked Pareto.

- ❖ Before and After in which the "new Pareto" bars are drawn side by side with the original Pareto, showing the effect of a change. It can be drawn as one chart or two separate charts.

- ❖ Change the Source of Data in which data is collected on the same problem but from different departments, locations, equipment, and so on, and shown in side-by-side Pareto Charts.

- ❖ Change Measurement Scale in which the same categories are used but measured differently. Typically "cost" and "frequency" are alternated.

Cause & Effect Diagram (Fishbone Diagram)

Why use it?

To allow a team to identify, explore, and graphically display, in increasing detail, all of the possible causes related to a problem or condition to discover its root cause(s).

What does it do?

- ❖ Enables a team to focus on the content of the problem, not on the history of the problem or differing personal interests of team members.

- ❖ Creates a snapshot of the collective knowledge and consensus of a team around a problem. This builds support for the resulting solutions.

- ❖ Focuses the team on causes, not symptoms.

How do I do it?

1. Select the most appropriate cause & effect format. There are two major formats:

   - ❖ Dispersion Analysis Type is constructed by placing individual causes within each "major" cause category and then asking of each individual cause "Why does this cause (dispersion) happen?" This question is repeated for the next level of detail until the team runs out of causes. The graphic examples shown in Step 3 of this tool section are based on this format.

   - ❖ Process Classification Type uses the major steps of the process in place of the major cause categories. The root cause questioning process is the same as the Dispersion Analysis Type.

2. Generate the causes needed to build a Cause & Effect Diagram. Choose one method:

   - ❖ Brainstorming: without previous preparation

   - ❖ Check Sheets based on data collected by team members before the meeting.

3. Construct the Cause & Effect/Fishbone Diagram

   - ❖ Place the problem statement in a box on the right-hand side of the writing surface.

❖ Allow plenty of space. Use a flipchart sheet, butcher paper, or a large white board. A paper surface is preferred since the final Cause & Effect Diagram can be moved.

Tip Make sure everyone agrees on the problem statement. Include as much information as possible on the "what", "where", "when", and "how much" of the problem. Use data to specify the problem.

(a) Draw major cause categories or steps in the production or service process. Connect them to the "backbone" of the fishbone chart.

Production	Services
Manpower	People
Materials	Policies
Methods	Process
Machines	Procedures

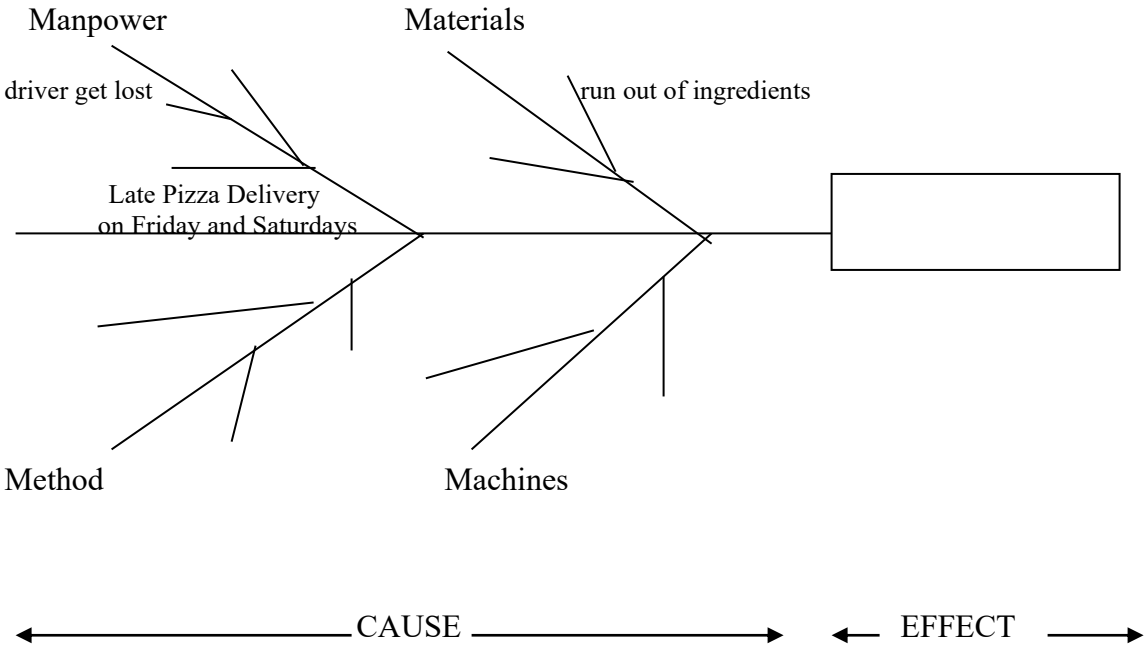


Illustration Note: In a Process Classification Type format, replace the major "bone" categories with: "Order Taking", "Preparation", "Cooking" and "Delivery".

❖ Be flexible in the major cause "bones" that are used. In a Production Process the traditional categories are: Machines (equipment), Methods (how work is done), Materials (components or raw materials), and People (the human element). In a Service Process the traditional methods are: Policies (higher-level decision rules), Procedures (steps in a task), Plant (equipment and space), and People. In both types of processes, Environment

(buildings, logistics, and space), and Measurement (calibration and data collection) are also frequently used. There is no perfect set or number of categories. Make them fit the problem.

(c) Place the brainstormed or data-based causes in the appropriate category.

- In brainstorming, possible causes can be placed in a major cause category as each is generated, or only after the entire list has been created. Either works well but brainstorming the whole list first maintains the creative flow of ideas without being constrained by the major cause categories or where the ideas fit in each "bone".

- Some causes seem to fit in more than one category. Ideally each cause should be in only one category, but some of the people causes may legitimately belong in two places. Place them in both categories and see how they work out in the end.

Tip If ideas are slow in coming, use the major cause categories as catalysts, eg, "What in 'materials' is causing….?"

(d) Ask repeatedly of each cause listed on the "bones", either:

- "Why does it happen?" For example, under "Run out of ingredients" this question would lead to more basic causes such as "Inaccurate ordering", "Poor use of space", and so on.

- "What could happen?" For example, under "Run out of ingredients" this question would lead to a deeper understanding of the problem such as "Boxes", "Prepared dough", "Toppings" and so on.

Tip For each deeper cause, continue to push for deeper understanding, but now when to stop. A rule of thumb is to stop questioning when a cause is controlled by more than one level of management removed from the group. Otherwise, the process could become an exercise in frustration. Use common sense.

(e) Interpret or test for root cause(s) by one or more of the following:

- Look for causes that appear repeatedly within or across major cause categories.

- Select through either an unstructured consensus process or one that is structured, such as Nominal Group Technique or Multi-voting.

- Gather data through Check Sheets or other formats to determine the relative frequencies of the different causes.

Variations

Traditionally, Cause & Effect Diagrams have been created in a meeting setting. The completed "fishbone" is often reviewed by others and/or confirmed with data collection. A very effective alternative is CEDAC®, in which a large, highly visible, blank fishbone chart is displayed prominently in a work area. Everyone posts both potential causes and solutions on Post-it™ notes in each of the categories. Causes and solutions are reviewed, tested, and posted. This technique opens up the process to the knowledge and creativity of every person in the operation.

Gantt Chart

Another widely used, schedule-monitoring method is the Gant chart. It is a simple tool that uses horizontal bars to show which tasks can be done simultaneously over the life of the project. Its primary disadvantage is that it cannot show which tasks are specifically dependent on each other.

Example (Gantt chart):

Final Year Project = Developing a programmable robotic arm.

Activity\Period	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
Literature search	XXXXXXXXX							
Documentary studies		XXXXXXXXXX						
Case study			XXXXXXXXXXXX					
Action Research or Survey								
Develop Supervisor W.								
Develop H. W.						XXXXXX		
Write up								XXX

Control Charts

Why use it? To monitor, control, and improve process performance over time by studying variation and its source

What does it do?

- Focuses attention on detecting and monitoring process variation over time
- Distinguishes special from common causes of variation, as a guide to local or management action.
- Serves as a tool for ongoing control of a process.
- Helps improve a process to perform consistently and predictably for higher quality, lower cost, and higher effective capacity.
- Provides a common language for discussing process performance.

How do I do it?

There are many types of Control Charts. The Control Chart(s) that your team decides to use will be determined by the type of data you have. Two handouts have been prepared on this very important topic to ensure students are about to apply these charts to real situations.

Common Questions for Investigating an Out-of-Control Process

☐Yes ☐No Are there differences in the measurement of accuracy of instruments/methods used?

☐Yes ☐No Are there differences in the methods used by different personnel?

☐Yes ☐No Does the environment, e.g. temperature, humidity, affect the process?

☐Yes ☐No Has there been a significant change in the environment?

☐Yes ☐No Is the process affected by predictable conditions? Example: tool wear.

☐Yes ☐No Were any untrained personnel involved in the process at the time?

☐Yes ☐No Has there been a change in the source for input to the process? Example: raw materials, information.

☐Yes ☐No Is the process affected by employee fatigue?

☐Yes ☐No Has there been a change in policies or procedures? Example: maintenance procedures.

☐Yes ☐No Is the process adjusted frequently?

☐Yes ☐No Did the samples come from different parts of the process? Shifts? Individuals?

☐Yes ☐No Are employees afraid to report "bad news"?

☐ A team should address each "Yes" answer as a potential source of a special cause.

Chapter

10

Basic Probability

Principle Objectives

- *To understand the basic principles of probability*
- *To discuss the applications of basic probability*
- *To apply more advanced concepts probability to the solution of real problems.*

Introduction

Probability is a measure of how likely an event is to occur. The lowest value is zero which means it cannot happen. The highest value is 1 which means it is certain to happen. Probability can be expressed as a fraction

P(E), the probability that event E occurs

no. of ways that E can occur

=

no. of different events that can occur

Total Probability = 1 so that

P (E) + P(not E) = 1

Example: A bag holds 5 blue balls, 3 red balls and 2 black balls.

The probability that a ball selected at random is blue = 5/10 = 0.5.

The probability of selecting a red ball

P(R) = 3/10 = 0.3.

The probability of selecting a non black ball = 1 – P(B)

= 1 - 2/10

= 8/10 = 0.8.

This is a theoretical approach to probability. Sometimes we cannot analyse a situation to work out probabilities but have to perform an experiment.

To work out the probability that a thumb tack (drawing pin) will land point up you have to do an experiment to estimate the probability. The more times you repeat the trial, the closer the estimate will be to the probability.

Possibility Space

Sometimes when more than one event takes place it is useful to draw a diagram to illustrate all the different possible outcomes. For example to find the probability of scoring a total of 10 with 2 dice we can make a table to illustrate the possibility space.

		1	2	3	4	5	6	First die
	1	2	3	4	5	6	7	
	2	3	4	5	6	7	8	
Second die	3	4	5	6	7	8	9	
	4	5	6	7	8	9	10	
	5	6	7	8	9	10	11	
	6	7	8	9	10	11	12	

There are 36 equally likely outcomes; of these, 3 result in a score of 10. Hence the probability of obtaining a score of 10 is 3/36.

Combining Probabilities

Suppose we roll a die two times. The probability of getting a '6' on the first throw is simply 1/6. To find out the probability of obtaining say two sixes on the two throws is more complicated. To help us we can draw a tree diagram. Each throw will be shown as follows with the probabilities written on the diagram.

a six
1/6
not a six
5/6
First throw

For the second throw, we have to add more branches

six
six 1/6 1/6 x 1/6 = 1/36 two sixes
not a six 1/6 x 5/6 = 5/36 one six
1/6 5/6

not a six six 5/6 x 1/6 = 5/36 one six
1/6 5/6 x 5/6 = 25/36 no sixes
5/6 not a six
5/6

First Second
throw throw

To obtain the probability of rolling two sixes, you go along the top branch and multiply the probabilities;

i.e. 1/6 x 1/6 = 1/36.

To obtain the probability of rolling one six there are two routes. Each of them has probabilities 1/6 x 5/6.

To find the probability of one six you have to add the end point probabilities, i.e. 5/36 + 5/36 = 10/36.

Example

A box contains 5 red balls, 4 green balls and 3 yellow balls. If two balls are drawn out what is the probability that they are both red?

As before you can draw a tree diagram where each split represents a different selection.

Note: in writing the probabilities on the diagram, the second branch has its probability values changed because the number of balls is reduced.

red
4/11
red
5/12 not red
7/11
red
not red 5/11
7/12
not red
6/11

First Second
Ball Ball

The probability of getting two red balls is therefore 5/12 x 4/11 = 20/132

Probability Distribution

Before we look at the normal distribution we must understand probability distributions.

Where frequency distribution record a physical count of how many times a particular value of the variable has been seen to occur, probability distribution record the degree of likelihood with which it will occur. So in table 8-1 the lifetime of Sunrise bulbs column 1&2 represent the frequency distribution whereas the other columns represent the probability distribution.

so Pr(400< lifetime <500) = 0.165+0.135+0.1+0.06+0.03 = 0.49

Note that the table shows tabulation of continuous data. So a probability 400<X<500 makes sense and a probability X=400 is only meaningful in tables where the variables are a discrete.

Discrete variables can only be certain fixed values. Examples are the size of a family, number of rainy days in September, these variables are countable.

Continuous variables can be any value, some ones weights since it can be measured to any desired accuracy or number of decimal places. These variables are not countable but measurable.

Just as we were able to draw histograms to represent frequency distributions we can represent probability distributions graphically too.

Notice that the horizontal scale of our bar chart shows the scale interval since our data items are continuous. It however conventional to draw a continuous line or curve and to use mid. class intervals for the horizontal scale.

Also our bar chart representing discrete variables is appropriate since the data items can be measure in "whole" units of measure. The lines joining our frequency distribution however can be smooth curves since our data can take any value on the horizontal scale of the graph because our data is a continuous value.

The vertical axis on the probability distribution does not have the usual x.y relationship expected from mathematical function (dependent, independent variable) for continuous data but the area under the graph is significant. So the area between the horizontal axes 450-390 will be proportional to the probability that a randomly selected Sunrises Bulb will last in between those two hours. From table 8-1

Pr (390 < X <450) = 0.5*0.145+0.165+0.135+0.5*0.100 =0.4225 = 42.25%

Normal Distribution

This is an important continuous probability distribution. Many measurements in the nature and industrial world such as height, weight, time in random events can be described by the normal distribution.

Parameters of the NDn.

Note that μ and σ are the population parameters. If we use sample parameters Xbar and s then to estimate the population the N in the denominator in some formulas is changed to (N-1). To avoid student confusion we therefore deal with population parameters here and use sample estimate in future lessons.

Normally distributed events are those that have a bell shape distribution graph and are symmetrical about its central peak. (draw an example on the board). Since it is symmetrical the mean, median and mode are identical.

There are two parameters that completely specify any normal curve. These are the central measurement variable-mean μ and the spread measurement variable-SD σ .

So the mean is the value that the distribution is centred around. = $(1/N).\sum x$ = Xbar & μ

The SD measured the spread around the mean.

PROPERTIES OF THE ND.

The bell shape of ND has precise mathematical properties which hold for all normal distributions. These are:

- Approximately 68% of all observations fall within the range of 1σ on both side of the mean.

$P\{ (\mu-1\sigma) \subseteq X \subseteq (\mu+1\sigma) \} = 0.68$

- App. 95% fall within +- 2σ

$P\{ (\mu-2\sigma) \subseteq X \subseteq (\mu+2\sigma) \} = 0.95$

- App. 9.7% fall within 3σ

$P\{ (\mu-3\sigma) \subseteq X \subseteq (\mu+3\sigma) \} = 0.997$

It therefore is possible to define the standard normal distribution as a bell shape probability distribution that has $\mu=0$ and $\sigma=1$. Note that the graph is symmetrical aroundμ. A table of the areas under the standard normal curve (thus probabilities) have been prepared.

The area under the curve bounded by the mean and Z on the table shows the probability of a random event. The -ve value of Z is obtained by symmetry.

In real life most data sets do not have a mean of zero and an SD of 1. For example the seems to have a mean of around 410 and SD of around 25 (use the graph plot 8-2 to approximate these). By normalising the problem at hand this table can be used to solve non standard normal distribution problems

Using the Table:

So how can we find out the probability of a random purchase of sunrise bulb to last longer than 500 hours, that is Pr(X>500). {given mean= 400.90 and SD=49.23}

This is like having to find the area under the graph that lies to the +ve side of the Horizontal axis at the 500 mark.

We then change this non-standard problem into a standard probability distribution by taking the variation of X from the mean μ and dividing it by the SD σ. That is to say how many SD from the mean is the point of interest:

$Z=(X-\mu)/\sigma$

This transforms our X into a standard normal variable Z.

So Z= (500-400.90)/49.23

Z=2.013

Pr(X>500) = 0.5- Pr (0<Z<2.013)

0.5- {0.4778+(0.4783-0.4778)*3/10}

0.5- {0.4778+0.00015}

0.5-0.47795

0.02205 = 2.21%

EXAMPLE: Probability of a light bulb to last between 450 and 370 hours?

Pr (370 < X < 450)

Pr{ (370-400.90)/49.23 < Z < (450-400.90)/49.23 }

Pr(-0.62766 < Z < 0.99736)

(0.2324+(0.2357-0.2324)*766/1000) + (0.3389+(0.3413-0.3389)*736/1000))

(0.2324+0.00253) + (0.3389+0.00177)

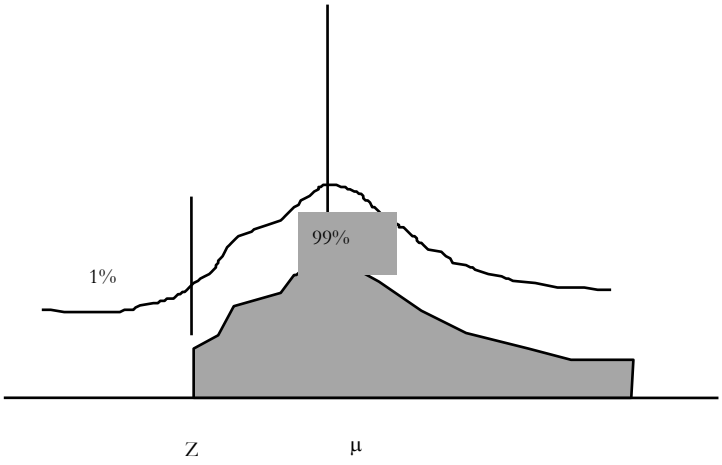
0.23493 + 0.34067 = 0.5756 = 57.56%

.....

Practical Exercise

Machinery packs goods into packages advertised as being nominally 9 KG in weight. In fact because of slight variability in machine performance, the actual pack weight may be above or below 9 Kg. However, it is known that the machine produces packs whose weight is normally distributed with a SD of 0.06 Kg. (The mean is normally set by the operator). Because of regulations governing short weights, the firm producing the packages must ensure that no more than 1% of packs fall below the advertised 9 Kg weight.

What mean weight should the machine be set at, in order to guarantee the firm meets this regulation ?



Pr (X < 9 Kg) = 0.01

Pr { $(X-\mu)/\sigma$ < 9Kg) = 0.01

The value of Z at which the area to the left is 1% or .01 is 0.5-0.01=0.49 giving Z as 2.33.

-2.33=$(9-\mu)/0.06$ Note the -ve sign of -2.33

-0.1398=$9-\mu$

μ = 9.1398 Kg.

Part B.

If this firm could use an alternative packing machine which operates to a SD of 0.03Kg. They produce 2.500.000 of these packs each month. The cost C in TL of producing a pack weighing g Kilograms is given by the expression;

C = 6.5 + 0.5g

Determine the monthly savings which could be achieved by changing to the alternative machine?

In the above solution replace sd = σ = 0.03 instead of 0.06

This will give a value of mean μ = 9.0699 Kg

So for packs that weigh 9.1398 on average the cost is

C= 6.5 + 0.5 x 9.1398 = 11.0699 TL

For packages having an average weight of 9.0699 the cost is

C= 6.5 + 0.5 x 9.0699 =11.03495 TL

The cost saving would thus be:

11.06990-11.03495 = 0.03495 for each pack

For 2.5 M packs this would be 2.500.000 x 0.03495 = 873.75

Things to remember

Pr[E∪F]=Pr[E]+Pr[F]-Pr[E∩F]

Combinations & Permutations

For equally likely outcome experiments we need set theory and counting (combination & permutation).

$P(n,r) = \frac{n!}{(n-r)!} =$ Arrangements = the number of ways of placing n distinct objects into r distinct places.

$C(n,r) = \frac{n!}{r!(n-r)!} =$ Selections = the number of ways of selecting r distinct objects from n distinct objects.

Stirling's Approximation

$n! \cong \sqrt{2\pi\, n} n^n e^{-n}$ (can be used when calculating Binomial probabilities)

① Sampling with replacement: Choosing r objects, one after the other, with replacement from a population of n distinct objects. Number of ways of doing this is n^r.

② Sampling without replacement: Choosing r objects, one after the other, without replacement from a population of n distinct objects. Number of ways of doing this is

$P(n,r) = \frac{n!}{(n-r)!}$.

③Simultaneous selection: Number of ways of selecting r objects simultaneously from a population of n distinct objects is

$$C(n,r)=\frac{n!}{r!(n-r)!}.$$

Problems of coincidences

Poincare counting principle: Let S be a sample space, E_1, E_2,…,E_n be events in S.

Then

$$\Pr[E_1 \cup E_2 \cup \ldots \cup E_n]$$

$$=\sum_{i=1}^{n}\Pr[E_i]-\sum_{1\le i<j\le n}\Pr[E_i \cap E_j]+\sum_{1\le i<j<k\le n}\Pr[E_i \cap E_j \cap E_k]-\ldots \mp \Pr[E_1 \cap E_2 \cap \ldots \cap E_n]$$

Axiom of countable additivity:

$$\Pr\left[\bigcup_{k=1}^{\infty}E_k\right]=\sum_{k=1}^{\infty}\Pr[E_k] \text{ in case of } \forall i,j \quad E_i \cap E_j = \varnothing.$$

First Borel-Cantelli

Let E_1, E_2,…, E_n,…be events with probabilities

$$P_n \equiv \Pr[E_n]. \text{ If } \sum_{n=1}^{\infty}P_n<\infty$$

Then the probability of occurring only finitely many of the events E_1,E_2,…,E_n,…is 1 (i.e. the probability of occurring infinitely many of the events E_1,E_2,…,E_n,…is 0).

$$0 \le \Pr[E] \le 1$$

$$\Pr[S]=1$$

$$\Pr[S-E]=1-\Pr[E]$$

Further Result:

$$\Pr\left[\bigcup_{k=1}^{\infty}E_k\right]\le\sum_{k=1}^{\infty}\Pr[E_k]$$

Conditional Probabilities

$$\Pr[E|F] \equiv \text{the probability of occuring } E \text{ in the case } F \text{ occurs}$$
$$(\text{the probability of } E \text{ given } F)$$
$$= \frac{\Pr[E \cap F]}{\Pr[F]}$$

$$= \frac{\mathrm{n}(E \cap F)/\mathrm{n}(S)}{\mathrm{n}(F)/\mathrm{n}(S)}$$

$$= \frac{\mathrm{n}(E \cap F)}{\mathrm{n}(F)}$$

Statistical Independence

E and F are independent of each other ≡

$$\Pr[E|F] = \Pr[E] \Leftrightarrow \Pr[E \cap F] = \Pr[E]\Pr[F] \Leftrightarrow \Pr[F|E] = \Pr[F]$$

The sequence of events $E_1, E_2, \ldots, E_n$ are independent of each other ≡

$$\forall k \quad \Pr[E_{i_1} \cap E_{i_2} \cap \ldots \cap E_{i_k}] = \Pr[E_{i_1}] \ldots \Pr[E_{i_k}]$$

Bayes Rule

To calculate inverse probabilities) Let E_i's be mutually exclusive events.

Then

$$\Pr[E_k|F] = \frac{\Pr[F|E_k]\Pr[E_k]}{\Pr[F]} = \frac{\Pr[F|E_k]\Pr[E_k]}{\sum_{i=1}^{n} \Pr[F|E_i]\Pr[E_i]}$$

Second Borel-Cantelli

Let $E_1, E_2, \ldots, E_n, \ldots$ be independent events with probabilities

$P_n \equiv \Pr[E_n]$. If $\sum_{n=1}^{\infty} P_n = \infty$

Then the probability of occurring infinitely many of the events $E_1, E_2, \ldots, E_n, \ldots$ is 1.

Probable, Possible, Improbable Events

Event E ≡ the monkey will be able to type the complete works of Shakespeare

Event E' ≡ A coin falls middle (not head and not tail)

The events E and E' are possible but improbable. Probability of an improbable event is, by convention, 0. So $\Pr[E], \Pr[E'] = 0$. But we will not be interested in improbable events. That is to say, all the events in the sequel will be assumed to be probable, i.e. hereafter

$\forall$Eevent $\Pr[E] > 0$ (1)

Mutually exclusive (Pairwise disjoint) events & Independent events

When two (or more) events cannot occur at the same time we say they are mutually exclusive (because they exclude each other!).

A: A coin falls heads

B: A coin falls tails

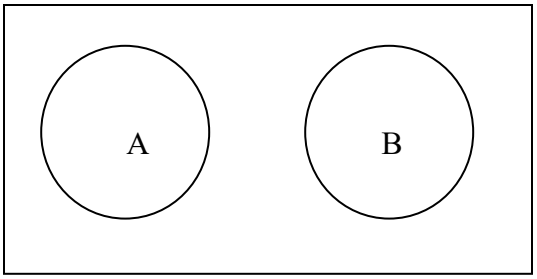
A and B are mutually exclusive because a coin cannot fall both heads and tails. Only one of them can happen.

When two events, A and B, are mutually exclusive there are no outcomes in A∩B and

so Pr[A∩B] = 0.

A and B are mutually exclusive ⇔ Pr[A∩B] = 0

When two events are mutually exclusive their two loops on a Venn diagram do not intersect as shown below:



Pr[E∪F]=Pr[E]+Pr[F]− Pr[E∩F]

Pr[E∪F]=Pr[E]+Pr[F]-0 (mutually exclusive)

Pr[E∪F]=Pr[E]+Pr[F] (mutually exclusive)

Pr[E OR F]=Pr[E]+Pr[F] (mutually exclusive)

Two events are independent when the occurrence/non-occurrence of one does not change the probability of the occurrence/non-occurrence of the other.

$\Pr[A|B] \neq \Pr[A]$ (when B occur the probability of A change from its original probability $\Pr[A]$)

$\Leftrightarrow \Pr[B|A] \neq \Pr[B] \Leftrightarrow A$ and B are dependent

$\Pr[A|B] = \Pr[A] \quad \Leftrightarrow \quad \Pr[B|A] = \Pr[B] \quad \Leftrightarrow \quad A$ and B are independent

A: A student plays table tennis

B: A student is good at mathematics.

There is no reason why a student who plays table tennis should be good or bad at mathematics. The above events A and B are independent.

A: An individual has a high IQ

B: An individual is accepted for a university place

Someone with a high IQ is more likely to be accepted for a university place than someone with a low IQ, so the events are dependent.

Confusion:

Independent events vs. mutually exclusive events

Recall that two events are mutually exclusive if they cannot both happen. For instance when a single dice is thrown the events A an even number is observed and B a 5 is observed, are mutually exclusive. On the other hand, the events C an even number is observed and D the number is a 4 are not mutually exclusive. When two events are mutually exclusive the probability that they both happen is zero, that is

Pr[A AND B] = Pr[A$\cap$B] = 0.

Two events are independent when the occurrence or non-occurrence of one event does not affect the probability of the other event. That is, when two events A and B are independent,

$\Pr[A|B] = \Pr[A]$ and

$\Pr[B|A] = \Pr[B]$

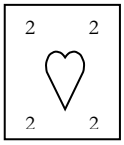
For instance, when a card is selected at random from a pack of cards the event A card is a heart and the event B card is a 2 are independent, because the probability of a two is 1/13, whether A happens or not.

Mutually exclusive $\Rightarrow$ dependent

When A and B are mutually exclusive, we know that if A occurs, B cannot occur, so $\Pr[B|A]=0$ whereas $\Pr[B]\neq 0$, the reader has already warned in (1) that we are interested in only probable events (that's why $\Pr[B]\neq 0$).

$\therefore$ Two mutually exclusive events are not independent.

$$\textit{independent} \Rightarrow \underbrace{\text{events intersect}}_{\textit{not mutually exclusive}} \text{ (contrapositive):}$$

In the above example the intersection is just the card 

Random Variables

The outcomes of a chance situation are often numbers. For instance, the number observed when throwing a dice or the temperature in Istanbul at midday tomorrow. When a variable can take different values according to chance it is called a random variable (RV). In other words, a random variable is an assignment of numerical values to all possible events of an experiment. The results of tossing a coin and the nationality of a student chosen from a class are not random variables because their outcomes are not numbers.

Discrete & Continuous RVs

Discrete RV(X):

X: Events$\rightarrow$Discrete set of values. e.g. {1,2,…,n}, natural numbers N, integers Z, $\{1,\frac{1}{2},\frac{1}{3},.....,\frac{1}{n}\}$ or $\{e,\sqrt{\pi},3,8\}$ (there are 'countably' many elements in the range)

Continuous RV(X):

X: Events$\rightarrow$Real numbers R (there are 'uncountably' many elements in the range)

The result of throwing a dice is a discrete random variable (X ≡ the result of throwing a dice) because it is only take some of the values (namely 1,2,3,4,5,6) between 1 and 6 inclusive. We cannot have a dice throw of 5.2 or 3.5, for instance!

The time taken for an athlete to run 1500 meters (X ≡ the time taken for an athlete to run 1500 meters) is a continuous random variable because it is any time within the possible range.

Probability density function (pdf)& Cumulative distribution function(Φ):

As it will be explained later, probabilities can be calculated from pdf. Given a random variable X, the probability density function pdf(X) (or P(X)) and cumulative distribution function Φ are defined as

for discrete random variable:

$$\Pr[s \le X \le t] \equiv \sum_{s \le X \le t} P(X)$$

(s and t don't need to be a random variable; s and t are just some, arbitrary but fixed, real numbers)

$$\Phi(x) \equiv \sum_{X_k \le x} P(X_k) = \sum_{-\infty \le X_k \le x} P(X_k) \qquad (= \Pr[\ -\infty \le X_k \le x\ ]) \qquad (-\infty, x \in \Re^*)$$

Notice that we put subscript to indicate discreteness and also observe that X_k is the value of the random variable X,

for continuous random variable:

$$\Pr[s \le X \le t] \equiv \int_s^t P(X)dX$$

(remember that integral is a continuous summation),

$$\Phi(x) \equiv \int_{-\infty}^{x} P(X)dX \quad (= \Pr[-\infty \le X \le x])$$

Remarks:

1) $\Phi(-\infty) = 0$

2) $\Phi(x) \ge 0$

3) $\Phi(x) \xrightarrow[x \mapsto \infty]{} 1$

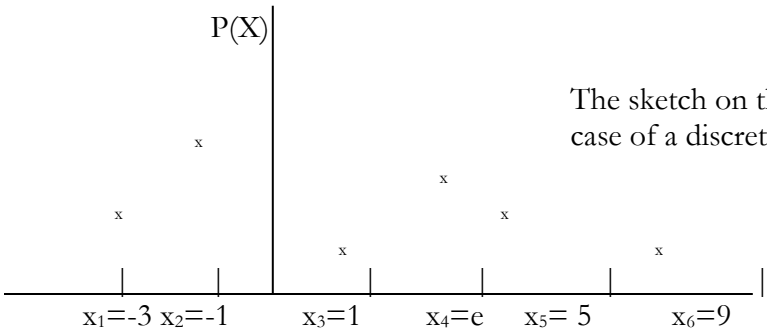
4). By Leibnitz's rule

5). $\dfrac{d\Phi}{dx} = P(X) \ge 0$ (at each point where $\overbrace{P(X) \text{ is continuous}}^{\text{applicability condition of Leibnitz}}$)

6) $\Phi(x)$ is an increasing function (an immediate consequence of 4th remark)

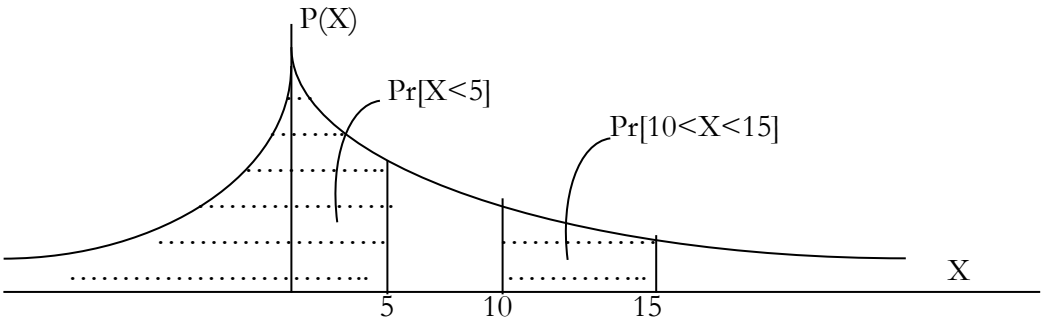
7). $\Phi(x \ge a) = 1 - \Phi(x \le a)$ } AREA EXPLANATION!

8). $\Phi(a \le x \le b) = \Phi(x \le b) - \Phi(x \le a)$ } AREA EXPLANATION!



The sketch on the left is for the case of a discrete random variable.

$$\Pr[s \le X \le t] \equiv \int_s^t P(X)dX$$



The pdf is chosen so that the area underneath the curve between any two values on the X-axis is the probability that X falls between these two values. The above sketch is for the case of a continuous random variable.

Expected value (mean) μ & Variance σ^2 of a random variable:

Discrete case:

$$P(X = X_k) \ge 0 \text{ (here } X_k \text{ is the value of the random variable)} \qquad \sum P(X = X_k) = 1$$

$$\mu \equiv \text{expected value of } X \text{ (or } E[X]) \equiv \sum_{k=1}^{\infty} X_k P(X = X_k) \qquad (-\infty \le X_k \le \infty).$$

$$E[X^2] = E[X^2 = X_k{}^2] = \sum_{k=1}^{\infty} X_k{}^2 P(X^2 = X_k{}^2) = \sum_{k=1}^{\infty} X_k{}^2 P(X = X_k \text{ OR } X = -X_k)$$

$$= \sum_{k=1}^{\infty} X_k{}^2 P(X = X_k) \text{ by assuming that we deal with only positive valued RVs.}$$

$$E\left[(X - E[X])^2\right] = E\left[X^2 - 2XE[X] + E^2[X]\right] = E[X^2] - 2\mu\mu + \mu^2$$

$$= E[X^2] - \mu^2 = E[X^2] - (E[X])^2$$

$$\sigma^2 \equiv \text{variance of } X \text{ (or } \text{Var}[X]) \equiv E[X^2] - (E[X])^2 = E\left[(X - E[X])^2\right]$$

$$\sigma_X \equiv \sqrt{\text{Var}[X]}$$

Continuous case:

$$\forall x \in \mathbb{R} \quad P(X = x) \geq 0 \text{ (If } \exists a \in \Re, X \neq a, \text{ then } P(X = a) \equiv 0 \text{ by convention)} .$$

$$\Pr[a \leq X \leq b] = \int_a^b P(X)dX \qquad \int_{-\infty}^{\infty} P(X)dX = 1$$

$$\mu \equiv \text{expected value of } X \text{ (or } E[X]) \equiv \int_{-\infty}^{\infty} XP(X)dX$$

$$E[X^2] = \int_{-\infty}^{\infty} X^2 P(X)dX$$

(due to the same assumption as above)

$$\sigma^2 \equiv \text{variance of } X \text{ (or } \text{Var}[X]) \equiv E[X^2] - (E[X])^2 = E\left[(X - E[X])^2\right]$$

$$\sigma_X \equiv \sqrt{\text{Var}[X]}$$

Binomial Distribution (BD)

Binomial situations:

A binomial situation is a random situation which has the following form.

1) There are n identical "trials". The word "trial" here just means "happening" or "repetition".

2) Each "trial" has two possible outcomes. We will generally call these success and failure, even if the outcome we have labelled "success" is not particularly desirable.

3) At each trial the probability of a success is the same. We will call this p, so at each trial

$$\Pr[success] = p$$

4) The result of each trial is independent of all the others. (That is, the success or failure of any trial does not affect the probabilities of the success or failure of the other trials.)

A coin is thrown.

$X_n \equiv$ number of heads in n trials $\equiv$ number of successes in n trials[Binomial RV], a discrete RV.

(A head is assumed to be a success).

$$P(X_n = i) = \binom{n}{i} p^i q^{n-i}$$ (probability density function for Binomial RV)

$$\Phi(\infty) = \sum_{X_n \le \infty} P(X_n) = \sum_{X_n \le n} P(X_n) = P(X_n = 0) + P(X_n = 1) + P(X_n = 2) + \cdots + P(X_n = n)$$

$$(\text{obviously } P(X_n = n+1) = P(X_n = n+2) = \ldots = 0)$$

$$= \binom{n}{0} p^0 q^n + \binom{n}{1} p^1 q^{n-1} + \ldots + \binom{n}{n} p^n q^0$$

$$= \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} = (p+q)^n = 1^n = 1 \text{ (as expected)}$$

Mean(expected value) and variance of binomial RV (p and q are arbitrary):

First let's find μ:

$$(p+q)^n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}$$

$$n(p+q)^{n-1} = \sum_{k=1}^{n} \binom{n}{k} k p^{k-1} q^{n-k} \text{ (by applying } \frac{\partial}{\partial p}\text{) index increases!!!}$$

$$pn(p+q)^{n-1} = \sum_{k=1}^{n} \binom{n}{k} k p^k q^{n-k} \text{ (by multiplying by } p\text{)}$$

$$pn(\underbrace{p+q}_{1})^{n-1} = \sum_{k=1}^{n} \binom{n}{k} k p^k q^{n-k}$$

$$pn = \sum_{k=1}^{n} \binom{n}{k} k p^k q^{n-k}$$

$$pn = \sum_{k=0}^{n} \binom{n}{k} k p^k q^{n-k} \text{ (starting from } k = 0 \text{ doesn't make any difference because } \binom{n}{0} 0 p^0 q^n = 0.\text{)}$$

$$pn = \sum_{k=0}^{n} k P(X_n = k) \quad (X_k = k \ (k = 0,1,2,\ldots,n,n+1,\ldots) \text{ are the values of the random variable } X_n)$$

$$pn = \underbrace{\sum_{k=0}^{\infty} k P(X_n = k)}_{\mu(X_n)} \quad (\text{notice that} \quad \forall k \ge n+1 \quad P(X_n = k) = 0)$$

The variance of a binomial random variable is $\sigma^2 = npq$.

Calculating Probabilities (Example):

10 Customers walk into a shop. The probability that an individual customer buying something is 0.3.

- ❖ What is the probability that exactly 2 customers buy?
- ❖ What is the probability that exactly 6 customers buy?
- ❖ What is the probability that no customers buy?

Solution: For this example, n=10 and p=0.3

a. Using the formula

$$P(2)=\binom{10}{2}p^2q^8=\frac{10!}{2!\,8!}0.3^2 0.7^8=0.233474$$

There is a 23.3% chance that exactly 2 customers buy.

b.
$$P(6)=\binom{10}{6}p^6q^4=\frac{10!}{6!\,4!}0.3^6 0.7^4=0.036757$$

There is a 3.7% chance that exactly 6 customers will buy.

c.
$$P(0)=\binom{10}{0}p^0q^{10}=\frac{10!}{0!\,10!}0.3^0 0.7^{10}=0.028248$$

There is a 2.8% chance that no customers buy.

Poisson Distribution(PD)

$$P(X)=\frac{e^{-\mu}\mu^X}{X!}\quad(\text{Generally written as } P(X)$$

$$=\quad\overbrace{\frac{e^{-a}a^X}{X!}}^{a\text{ stands for average}}\quad\text{so that}\quad\overbrace{P(X=k)=\frac{e^{-a}a^k}{k!}}^{\text{pdf for Poisson RV}})$$

PD is used, in particular, in Queuing theory to model the number of people who join a queue during a particular time interval:

There are two broad sets of circumstances for which a Poisson distribution is appropriate:

(1) To model the number of occurrences of a particular rare event in a time interval when, on average, the events occur μ times in the time interval and the event occur at the same average rate throughout the time interval events happen individually and independently

(2) as an approximation to the binomial distribution.

The Poisson approximation to the Binomial Distribution:

Recall that the conditions for a binomial situation are

① There are n identical "trials"

② Each trial has two possible outcomes, success or failure.

③ At each trial Pr[success]=p

④ The result of each trial is independent of all the others.

The number of successes has a binomial probability density function. However, suppose further that

⑤ The chance of success, p, is very small and may be unknown.

⑥ The number of trials, n, is very large and again may be unknown.

⑦ The average number of successes, $\mu=np$ (because we are in a specific situation of Binom), is known and is not large, say $\mu \leq 7$.

$X_n \equiv$ number of successes in n trials [Poisson RV]

When conditions ⑤,⑥,⑦ hold as well as ①,②,③ and ④, the number of successes still has a binomial probability distribution function but the Poisson distribution function provides a very good approximation. This is useful because when n is large binomial probabilities can be very tedious to calculate (will be seen in the following example) and they are not published in tables.

Mean(expected value) and variance of Poisson RV (p and q are arbitrary):

$$ae^a = a\frac{d}{da}(e^a) = a\frac{d}{da}(\sum_{k=0}^{\infty}\frac{a^k}{k!}) = a\sum_{k=1}^{\infty}k\frac{a^{k-1}}{k!} = \sum_{k=1}^{\infty}k\frac{a^k}{k!} = \sum_{k=0}^{\infty}k\frac{a^k}{k!}\ (**)$$

$$\mu(X) = \sum_{k=0}^{\infty}kP(X=k) = \sum_{k=0}^{\infty}ke^{-a}\frac{a^k}{k!} = e^{-a}\sum_{k=0}^{\infty}k\frac{a^k}{k!} \underset{(**)}{=\!=\!=} e^{-a}(ae^a) = a$$

$$\sigma^2(X) = \mu(X^2) - (\mu(X))^2 = \mu(X^2) - a^2$$

$$\mu(X^2) = \sum_{k=0}^{\infty}k^2P(X=k) = \sum_{k=0}^{\infty}k^2e^{-a}\frac{a^k}{k!} = e^{-a}\sum_{k=0}^{\infty}k^2\frac{a^k}{k!} = e^{-a}\underbrace{(a+a^2)e^a}_{\text{to be shown below}} = a + a^2\ (***)$$

$$a\frac{d}{da}(ae^a) \overset{?}{=\!=\!=} \sum_{k=0}^{\infty}k^2\frac{a^k}{k!}$$

$$\frac{d}{da}(ae^a) \overset{?}{=\!=\!=} \sum_{k=0}^{\infty}k^2\frac{a^{k-1}}{k!} \text{ (dividing by } a)$$

$$ae^a \overset{?}{=\!=\!=} \sum_{k=0}^{\infty}k^2\frac{a^k}{k.k!} \text{ (upon integration)}$$

$$ae^a \overset{?}{=\!=\!=} \sum_{k=0}^{\infty}k\frac{a^k}{k!} \text{ (OKEY we have already shown this in (**))}$$

$$a\frac{d}{da}(ae^a) = a(e^a + ae^a) = ae^a + a^2e^a = (a+a^2)e^a$$

$$\sigma^2(X) = \mu(X^2) - (\mu(X))^2 = \mu(X^2) - a^2 = a + a^2 - a^2 = a$$

Example - Calculating probabilities

The probability that a particular automobile part is defective is known to be 0.001. 3000 parts are required in the assembly of a car.

a. Use the binomial probability distribution to calculate the probability that there are no defectives. Now calculate the approximate probability using the Poisson distribution.

b. Calculate the probability of exactly 5 defectives. Try using both the binomial distribution (binomial probability density function) and the Poisson approximation.

Solution:

The distribution of the number of defectives is binomial, with n=3000 and p=0.001, so the mean is μ=np=3000×0.001=3. As n is large, and p is small so that $\mu \le 7$ we can approximate by a Poisson distribution.

a. Using the binomial distribution, the probability of 0 defectives is

$$P(0) = \binom{3000}{0} 0.001^0 0.999^{3000} = 0.999^{3000} = 0.0497124 \qquad 4.97\%.$$

Using the Poisson approximation, $\mu=3$ and so

$$P(0) = \frac{e^{-3}3^0}{0!} = 0.049787 \quad 4.98\%.$$

Notice that the results are very similar.

b. We require P(5). The binomial probability is

$$P(5) = \binom{3000}{5} 0.001^5\, 0.999^{2995} = 0.1008356 \; .$$

Using the Poisson approximation is much more straightforward and gives

$$P(5) = \frac{e^{-3}3^5}{5!} = 0.1008188.$$

Normal (Gaussian) Distribution (ND)

(standard normal $\mu=0$ $\sigma=1$)

When number of trials n is very large and $p = q = \frac{1}{2}$ then

$$\binom{n}{k} p^k q^{n-k} \xrightarrow[n \mapsto \infty]{} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(X-\mu)^2}{2\sigma^2}} \qquad (\sigma \equiv \sqrt{npq},\ \mu \equiv np)$$

$$P(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(X-\mu)^2}{2\sigma^2}} \qquad (\sigma \equiv \sqrt{npq},\ \mu \equiv np)$$ (probability density function for normal variable).

Mean (expected value) and variance of normal RV ($p=q = \frac{1}{2}$):

Continuity of mixed partial derivatives, double integrals and Polar Coordinates give:

$$\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}} \; .$$

$$\Phi(\infty)=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx \underset{y\equiv x-\mu}{=\!=\!=\!=}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-y^2}{2\sigma^2}}dy=\frac{1}{\sqrt{2\pi}\sigma}\sqrt{\frac{\pi}{(1/2\sigma^2)}}=1\,(\text{as expected})$$

$$\mu\equiv\int_{-\infty}^{\infty}XP(X)dX=\int_{-\infty}^{\infty}X\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(X-\mu)^2}{2\sigma^2}}dX=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}Xe^{\frac{-(X-\mu)^2}{2\sigma^2}}dX$$

$$(y\equiv X-\mu)\qquad=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}(y+\mu)e^{\frac{-y^2}{2\sigma^2}}dy$$

$$=\frac{1}{\sqrt{2\pi}\sigma}\left[\underbrace{\int_{-\infty}^{\infty}ye^{\frac{-y^2}{2\sigma^2}}dy}_{=0\text{ since odd fnc.}}+\mu\int_{-\infty}^{\infty}e^{\frac{-y^2}{2\sigma^2}}dy\right]$$

$$=\frac{1}{\sqrt{2\pi}\sigma}\mu\sqrt{\frac{\pi}{(1/2\sigma^2)}}$$

$$=\mu\quad(=np)$$

$$\text{Var}[X]=E[X^2]-(E[X])^2=\underbrace{\int_{-\infty}^{\infty}X^2\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(X-\mu)^2}{2\sigma^2}}dX}_{\sigma^2+\mu^2}-\mu^2=\sigma^2\quad(=npq)$$

Standardization procedure:

$X \sim N(\mu,\sigma^2)$. Any normal random variable X, which has mean μ and variance σ^2 can be standardized as follows:

Take the RV X, and

(i) subtract its mean, μ and then

(ii) divide by its standard deviation, σ

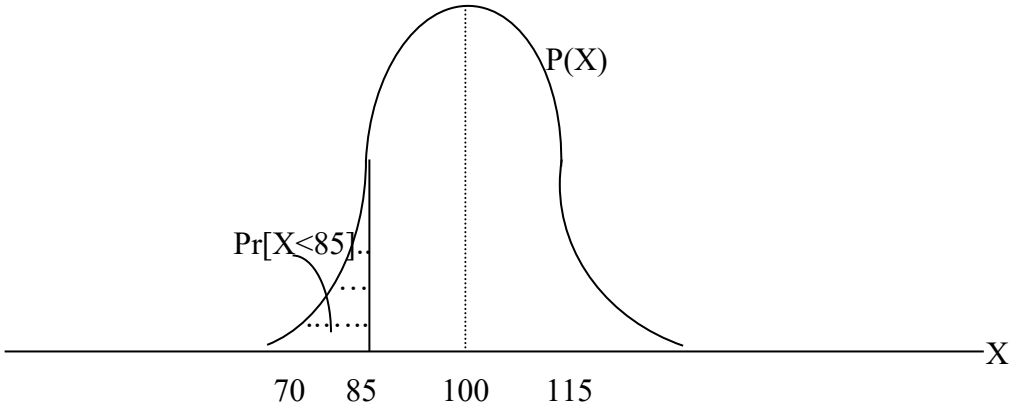
We will call the result, Z, so

$$Z\equiv\frac{X-\mu}{\sigma}.\qquad Z\sim N(0,1)$$

We are going to use this procedure to calculate non-standard normal probabilities.

Example - Calculating probabilities (

Consider the probability that an individual's IQ score is less than 85, P(X<85). The corresponding area under the $N(\mu,\sigma^2)=N(100,15^2)$ pdf is shown below:



We cannot use normal tables directly because these give N(0,1) probabilities. Instead, we will convert the statement X<85 into an equivalent statement which involve the standardized score,

$Z \equiv \frac{X-100}{15}$ because we know it has a standard normal distribution. Since X<85,

$$Z \equiv \frac{X-100}{15} < \frac{85-100}{15} = -1$$

.

We have X<85 ⇔ Z<−1 and Pr[X<85]=Pr[Z<−1].

Pr[Z<−1]is just a standard normal probability and so we can look it up in the following table

Table: Cumulative Standard normal Probabilities P(Z<a)

a	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1537	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611

So Pr[Z<−1]=Pr[X<85]=0.1537.This table is according to juxtaposition, not summation: Pr[−1.08]=Pr[−1 juxt 08]=0.1401.

Summary

x (random variable)	μ	variance=σ^2
Binom	np	npq
Poisson	a≡np	a≡np
Normal	μ≡np	σ^2≡npq

Bibliography

1. Ziarati, R. Information Management, Dogus University publication, 2003
2. Eden, A. Probability lecture notes (not published), 1997.
3. Gülmez, E. Basic Data Analysis for Experiments in the Physical Sciences, Boğaziçi University Publications, Istanbul, 1997.
4. Fleming, M. The Essence of Statistics for Business, Prentice Hall, New York, 1996.
5. Mcleod, E. Management Information Systems, Prentice Hall, New Jersey, 1995.
6. Morris, C., Quantitative approaches in Business Studies, Pitman, London, 1996.