CHAPTER 2

CONCEPT AND METHODOLOGY FOR STATISTICAL EXPERT SYSTEM FOR INFORMATION MANAGEMENT

2.1 Introduction

The present chapter provides an understanding of the basic concepts and approaches to developing statistical expert system, in order that the system developed is comprehensive. The previous, introductory chapter has provided a focus on information manager and information management, and, expert systems components and characteristics. The introductory chapter has also reviewed recent literature to indicate to the research directions in expert systems and expert system development. Advice-giving and pattern-recognising programs have been discussed. The rudiments of a statistical knowledge base and statistical inference engine have been discussed in order to work towards a Statistical Expert System as an Information Manager. Architectures and interfacing languages for the statistical expert system have also been discussed in as detailed a fashion as necessary. It is here again that the problem of study and the scope of the study have been discussed.

A comparison of statistical expert systems in use is the starting point for discussion in the present chapter, where the advice-giving and pattern-recognising systems are dealt with as they are in use in various institutions. The discussion then turns to the requirements of a typical statistical expert system, emphasising statistical problems requiring solutions. The steps in solutions and factors influencing them are outlined here. A comparison of statistical procedures offered by statistical packages such as SPSS, SYSTAT, STATISTICA, and WinSTAR is then provided to give an
appreciation of the range of problems that could be resolved using the statistical packages. This is then followed by a model of Statistical Expert System, with four main components of knowledge base, inference engine, and user interface and knowledge acquisition phase. Further down, the design of the requirements for the design of the model is indicated, even as the structure of the statistical expert system being conceived is given and explained.

Organising and representing knowledge of statistics in the expert system and the problems addressed form the next step. The description then turns to the programming language Turbo Prolog and the rule-based techniques and logics of the language used in the development of the system. Considering the need for the statistical expert system, the statistical inference chain for inferring statistical methods is elaborated with a model. Working memory and compression as means of storing and saving memory space are highlighted before the design and development defect reduction strategies are outlined.

2.2 Statistical Expert Systems in Use

Advice-giving programs have coded rules for decision-making and have the goals of solving or pointing to solutions to a select set of problems (for example: diagnosing a disease, or determining an appropriate statistical analysis). These involve statistical reasoning if they give statistical advice. The SESYS designed and reported in the thesis is one such expert system which gives advice on statistical techniques for given types of data and also helps with the analysis of data either by interfacing the users with an appropriate statistical package or by computing results using the user-given data. Currently however advice-giving programs are written in two modes. The information managers use their rules to solve the problems but turn to a human expert only if they lack rules to continue. Most such programs also
include explanation systems that allow the human expert to query the programs' reasoning (Pregibon and Gale, 1982; Gottinger, 1995; Buchanan and Shortliffe, 1984). Such programs are often called consultant systems as well, good examples of such being the MYCIN (Buchanan and Shortliffe, 1984), which provides medical diagnosis and prescribes drugs for bacterial infections and REX (Pregibon and Gale, 1982), for simple linear regression.

Pattern-finding programs often are such that they search through a database for interesting facts (for example: side effects of a drug to be administered for a given disease; or rules for reaching certain conclusions). These search a database to discover relationships; use statistical techniques such as correlations and discriminants as well as knowledge about their domain. For example, the RX program (Blum, 1982) searches a medical database for possible causal relationships; where causation is defined by lagged correlation. This program then uses medical knowledge to rule out common causes and clearly spurious correlations. Finally, both statistical and medical knowledge are used to test correlations, after controlling for other associated variables. The SESYS we are designing uses statistical knowledge and also knowledge about the structure of plausible rules. Typically, it has language-handling facility like all expert systems.

Current systems which access statistical packages such as RX (Blum, 1982) and REX (Pregibon and Gale, 1982) handle the tedious task of passing information back and forth from the packages to the system, by allowing some rules to create calls to the associated package and then scanning the output for relevant values (See Table 2.1 for a sample of statistical expert systems designed for advice-giving and pattern-recognising purposes).

Chambers (1981) was the expert who suggested the idea of building a statistical consultant expert system. This led to considerable debate in the
statistical community about the program as a consultant. One of the criticisms about the consultant was that the untrained users are given a powerful tool in their hands. But the consultant system is certainly a good system for those who use statistical packages without consulting a statistical expert. The argument against the consultant is that it lacks real world knowledge and natural language recognition. This means that it cannot deal with vague questions. Inasmuch as this system clarifies the premises, actions, and goals of the client, it is a useful system.

It is however useful to think that experts in other fields are uncomfortable about using consulting systems. It leads us to the fact that:

- The computer does not understand how complex the data set really is (in terms of XSAMPLE); or
- It cannot understand how sick the patient really is (in terms of MYCIN).

Yet, in blind assessments, MYCIN compared well with human experts and that XSAMPLE worked well for those seeking advice. There are however two basic differences between a consultant system for statistics and that for other fields. In a field such as chemistry or medicine, the consultant system has knowledge about the problem domain. In the consultant system for statistics however there is knowledge about the statistics domain, not about the problem domain. Secondly, medical consultant systems are designed for use by doctors and not by patients. On the other hand, statistical consultant systems are primarily designed for use by the untrained users and managers and not expert statisticians or experienced managers. Given the current use of statistical packages by clients with almost no knowledge of statistics, the developers of consultant systems must realistically expect that these systems will be used by clients with little statistical experience. So it may be necessary to design expert
systems, which could give consultation as well as computing or provide access to statistical packages and do analysis for the users. Such a system could then be very useful for novices and experts in statistics alike.

The field of expert systems (ExpSys), and the community designing and using such systems, has grown tremendously, worldwide. As the field has matured, ExpSys are being developed and deployed in a myriad of applications and fields (Liebowitz, 1998). An excellent review of existing expert systems applications is by Waterman (1986), who catalogues over 150 systems, in areas such as medicine, and military science. The quality of operational ExpSys is outstanding. Application of ExpSys range from medicine and military science, blast furnace control in Japan to strategic management support in Germany, to scheduling crew in Portuguese railways to management of sheep production in Australia, and to hurricane damage assessment in the Caribbean to modeling a black teenager or subjects of teenage pregnancy, drug, and alcohol abuse. Only a fewer Expert Systems (ExpSys) have been developed in the area of information management. As the area of ExpSys is expanding rapidly, both the Government and Industry the world over are beginning to invest in commercial expert systems and, it is expected that, within a few years a very large number of companies may be involved in Expert Systems Research and Development (ESRD) in India. Companies that ignore this high technology advances will certainly find themselves at a competitive disadvantage in the near future.

The education and training of middle and top-level management in Information Technology (IT) is less than adequate (Black and Trippi, 1990). Today, for example, many administrators at all levels have not received an education and training necessary to overcome the fear of computer technology. The communications between statisticians and engineers or physical scientists range from outstanding to awful (Hoadley and
Kettenring, 1990), but overall there are strong indications of serious problems that need resolution.

Table 2.1: Statistical Expert Systems Designed for Advice-Giving and Pattern Recognising Purposes

<table>
<thead>
<tr>
<th>Category</th>
<th>Name of System</th>
<th>Year of Design</th>
<th>Stage</th>
<th>Authors/ Designers</th>
<th>University/ Institutions</th>
<th>Purpose/ What is done?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advice Giving</td>
<td>REX</td>
<td>1982</td>
<td>Prototype</td>
<td>Gale W.A and Pregibon D.</td>
<td>Bell Lab.</td>
<td>Simple linear regression</td>
</tr>
<tr>
<td></td>
<td>Automated</td>
<td>1982</td>
<td>Prototype</td>
<td>O'Keefe, R.</td>
<td>Brunei University</td>
<td>Advice giving - insufficient in communicating with a statistically naive client... [because] you have to know too much to use it</td>
</tr>
<tr>
<td></td>
<td>Statistical</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Advice giving:</td>
</tr>
<tr>
<td></td>
<td>analyst (ASA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ASC was suggested but not implemented</td>
</tr>
<tr>
<td></td>
<td>STATPATH</td>
<td>1983</td>
<td>Prototype</td>
<td>Portier and Lai</td>
<td>Not available</td>
<td>Only advice consultant, does entire analysis</td>
</tr>
<tr>
<td></td>
<td>ASC</td>
<td>1984</td>
<td>Hypothetical Example</td>
<td>Hahn</td>
<td>Not available</td>
<td></td>
</tr>
<tr>
<td></td>
<td>XSAMPLE</td>
<td>1984</td>
<td>Complete System</td>
<td>Van Melle W., Scott A.C., Bennett J.S., Peuris M.</td>
<td>Stanford University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>KBSSP</td>
<td>1993</td>
<td>Prototype</td>
<td>Mani T and NRV Prabhu</td>
<td>University of Madras</td>
<td>Advice on appropriate methods</td>
</tr>
<tr>
<td></td>
<td>EASSA</td>
<td>1993</td>
<td></td>
<td>Zhang, Anung</td>
<td>Carleton University, Canada</td>
<td>Advisory System, Access GIS for mapping</td>
</tr>
<tr>
<td></td>
<td>GCA</td>
<td>1995</td>
<td>Hybrid methods</td>
<td>Lenard, Mary Jane</td>
<td>Kent state University</td>
<td>Evaluates firms at risk of receiving auditors report</td>
</tr>
<tr>
<td></td>
<td>SESA</td>
<td>2000</td>
<td>Prototype</td>
<td>Mani T</td>
<td>University of Madras</td>
<td>Access Statistical packages</td>
</tr>
<tr>
<td></td>
<td>SGESC &amp; LS</td>
<td>1992</td>
<td>Prototype</td>
<td>Madizon, Elinor M</td>
<td>Florida Institute of Technology, USA</td>
<td>Elementary Statistics Interactive capability Immediate feedback</td>
</tr>
<tr>
<td></td>
<td>ESFDHP</td>
<td>1992</td>
<td>Prototype</td>
<td>Al-Jabri, Ibrahim Muhammad</td>
<td>Illinois Institute of Technology</td>
<td>Access statistical packages</td>
</tr>
<tr>
<td>Pattern-Recognisa</td>
<td>RX</td>
<td>1982</td>
<td></td>
<td>Blum</td>
<td>Not available</td>
<td>Selection of methods</td>
</tr>
</tbody>
</table>
Research continues at greater pace and the computer industry is working towards integrating the existing abilities of data processing with the new knowledge processing.

The Governments, industry, planners, policy makers, researchers and scientists use statistics in making, besides many others, policy decisions, monitoring their results and in trying to understand economic and social relationships.

The administrators, who have to take responsibility for any of the decisions, based upon their recommendations, should understand the nature of their reasoning. There are situations where many managers and administrators (from now on, the two categories are also representative of other top-cadre professionals and so will be referred to as only administrators and managers, hereafter) are looking for answers to the question: which statistical technique should I use for my data? They should know the statistical tools for analysis, estimation, hypothesis testing, examination of strength of relationships and multivariate analyses. The guidelines of when to use these techniques involve a number of criteria such as the type and level of measurement, sample size, and whether the observations are independent, or not.
Statistical software is available for various applications and they vary widely in quality. Unfortunately, softwares are widely misused by non-statisticians. Users and managers of statistics or information need to be versatile and resourceful in tackling statistical problems. It is important to note the different stages of a statistical investigation, from problem formulation through the presentation of results (Charfield, 1989). Statistics is concerned with collecting, processing and interpreting data in the best possible way. Any statistical problem needs the following for better solution:

a) The formulation of a real problem in statistical terms;

b) The efficient way of data/information collection;

c) Processing of data/information effectively; and

d) Interpretation and reporting of results.

To execute the four steps above successfully, they should identify the following factors accordingly:

1) the nature of problem;

2) the number of variables involved in the problem;

3) the type of variables;

4) the variable constraints;

5) the variable strengths and limitations;

6) the assumptions made on variables;

7) the total number of observations;

8) the concept(s) involved in the problem;

9) the objective(s) of the problem;

10) the structure of the data; and

11) the more important features of data.

Some researchers have addressed the general issue of intelligent statistical software and ExpSys (Mellichamp and Park, 1989). However,
these efforts have been more from a theoretical design point of view and do not describe actual systems, which have been developed. SESYS is an attempt to develop a system, which may be used to lead to better statistical data analysis, decision-making and forecasting for a variety of people. As ExpSys become larger and more complex and come to be used in environments away from their developers (Jacob and Froscher, 1990), maintenance of the knowledge base becomes an overwhelming problem.

As statistical packages vary widely in quality, the non-statistician user has difficulties in using them. Frequently, learning how to use a sophisticated package will require learning a branch of statistics requiring investment of time, which cannot be made. Unfortunately, packages are widely misused by non-statisticians (Charfield, 1988) and we need packages, which incline towards ExpSys where the packages will be able to say, for example, whether a given set of data is suitable or not for fitting an appropriate model.

The proliferation of PC-based statistical softwares, increase in computer processing speed, and the availability of larger and faster hard disks have made it possible for the vendors to increase the sophistication of their products and to accommodate large data sets. The PC-based packages now offer a comprehensive array of statistical procedures while giving users greater flexibility in exploring and analyzing data. This is to say that there is no need for writing modules for various statistical procedures for use by the SESYS users, but rather interface the statistical packages with the SESYS such that the packages can be accessed by the users of the SESYS. In fact, such access can be established with more than one statistical package so that the users can access the one they most prefer.

Lurie (1994) has made a review of five statistical packages for Windows, which offers a glimpse at the packages, some of which can be
considered for interfacing with the SESYS. The statistical packages most commonly used in India, such as the SPSS, STATISTICA and SYSTAT, may be accessed from the SESYS so that data can be processed by the users, once they receive advice from the SESYS for a particular or any specific set of procedures. Morgan (1998) has made another review of eight statistical packages for general use.

In general, the more procedures offered by a statistical package, the more useful it is. It is often very disconcerting to find that once the users have become comfortable using a statistical package for a particular set of analyses, it cannot perform additional analyses the users need. Therefore, they must switch to another package, taking up more disk space and incurring additional cost. Of course, no single package can possibly cover the universe of statistical procedures available. SPSS offers the most procedures of the packages. Table 2.2 summarises the statistical procedures offered by each of the packages that could be interfaced and accessed.

Some researchers have addressed the general issue of intelligent statistical software and ExpSys (Mellichamp and Park, 1989). However, these efforts have been more from a theoretical design point of view and do not describe actual systems, which have been developed. SESYS is an initial attempt to develop a statistical expert system, which may be used to lead to better statistical data analysis, decision-making and forecasting for administrators and managers. As ExpSys become larger and more complex and come to be used in environments away from their developers (Jacob and Froscher, 1990), maintenance of the knowledge base becomes an overwhelming problem. This study also covers statistical knowledge from the following statistical methods and these may be classified, for convenience of handling, as falling under analysis, decision and forecasting (Table 2.3).
Table 2.2: Comparison of Statistical Procedures Offered by Statistical Packages

<table>
<thead>
<tr>
<th>No</th>
<th>Statistical Procedure</th>
<th>SPSS</th>
<th>SYSTAT</th>
<th>STATISCA</th>
<th>WinSTAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Basic statistics</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>Analysis of variance</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>3.</td>
<td>General linear models</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>Canonical correlation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>Descriptive statistics</td>
<td>*</td>
<td>*</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>6.</td>
<td>Cluster analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>Discriminant analysis</td>
<td>*</td>
<td>*</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>8.</td>
<td>Factor analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>Principal components</td>
<td>*</td>
<td>*</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>10.</td>
<td>Logic models</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>11.</td>
<td>Multidimensional scaling</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>12.</td>
<td>Non-parametric tests</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>13.</td>
<td>Multiple regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14.</td>
<td>Stepwise regression</td>
<td>*</td>
<td>*</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>15.</td>
<td>Multiple correlation</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>16.</td>
<td>ARIMA models</td>
<td>*</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>17.</td>
<td>Frequency tables</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18.</td>
<td>Cross tabulations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19.</td>
<td>Time series analysis</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>20.</td>
<td>Spectral analysis</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>21.</td>
<td>Student t tests</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>22.</td>
<td>Chi(^2) tests</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>23.</td>
<td>F tests</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>24.</td>
<td>Kruskal-Wallis</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>25.</td>
<td>McNemar's test</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
</tbody>
</table>


The knowledge base contains the rules and facts about the statistical domain available from literature (Ashford, 1997; Caswell, 1989; Makridakis and Wheelwright, 1987; Wheelwright, 1985; Wheldon, 1981).

Expert system DEXTER was developed to assist researchers in the area of experimental design (Melichamp and Park, 1989), while REX was a prototype system, which assisted in regression analysis. The SASS was somewhat broader in scope attempting to identify an appropriate statistical
procedure, given certain research objective. The classification of statistical methods is shown in Figure 2.1.

2.3 Towards a Statistical Expert System Model

Knowledge is the essential source of problem solving power in knowledge-Based Systems (KBS) Programs. The process of constructing a KBS program involves an iterative technique of acquiring knowledge from one or more sources and representing it such that a computer can process it.

Table 2.3: Statistical Methods for Statistical Expert Systems

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Name of Statistical Methods</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Organizing Data</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>Numerical Descriptive Measures</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>Probability</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>Discrete RANDOM Variables and their Probability Distributions</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>Continuous RANDOM Variables and the Normal Distributions</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>Sampling Distributions</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>Estimation of the Mean and Proportion</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>Hypothesis tests about the Mean and Proportion</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>Estimation and Hypothesis Testing</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>Testing with Two Populations</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>Chi-Square Tests</td>
<td>7</td>
</tr>
<tr>
<td>12</td>
<td>Analysis of Variance</td>
<td>15</td>
</tr>
<tr>
<td>13</td>
<td>Simple Linear Regression</td>
<td>13</td>
</tr>
<tr>
<td>14</td>
<td>Multiple Regression</td>
<td>7</td>
</tr>
<tr>
<td>15</td>
<td>Non-Parametric Methods</td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>Analysis of Covariance</td>
<td>3</td>
</tr>
<tr>
<td>17</td>
<td>Data Management and Report Proportion</td>
<td>12</td>
</tr>
<tr>
<td>18</td>
<td>Time Series</td>
<td>13</td>
</tr>
<tr>
<td>19</td>
<td>Index Numbers</td>
<td>7</td>
</tr>
<tr>
<td>20</td>
<td>Plots: Probability (P-P) and Quantile (Q-Q) plots</td>
<td>4</td>
</tr>
</tbody>
</table>


An expert system is computer software that helps people to find a solution to a problem drawing on expert knowledge that has been embodied within (explicitly) in the form of a knowledge base. Expert software (system) has four main components:
Figure 2.1: Step-by-Step Methodology for Determining Statistical Methods and Procedures using Statistical Expert System for Information Management
1. The knowledge base contains the facts agreed by experts, the common knowledge they have acquired over years of work, and the rules of thumb that they apply to desire conclusions.

2. The inference engine is a program that simulates the deductive thought processes. Of the human expert, drawing on the knowledge base and further information entered by the expert system user.

3. The user interface helps the expert system user to enter relevant facts and to ask questions. Normally, it will use a natural 'English-like' language to conduct question-and-answer dialogue with the user.

4. The knowledge acquisition facility acquires knowledge from the human expert (or The knowledge engineer who acts as an intermediary between the human expert and the expert system) in the form of rules and facts, and encodes these to build up the knowledge base.

Why Expert Systems are not emerging?

1. The cost of developing expert systems is very high.

2. Software piracy. It is defined as the direct, unauthorised copying of a program for commercial gain - is common throughout the world. In some Asian Countries, including China, South Korea and Taiwan, more than 90 percent of all U.S. microcomputer software in use is pirated.

American software companies or the Government could even assist other nations. Interested in developing their own software industry, particularly, in developing Expert Systems.
In the development and implementation of an ExpSys application for solving problems, the two most significant problems are:

(i) the lack of qualified knowledge engineers and ExpSys designers; and

(ii) the lack of commitment on the part of top management and availability of experts from whom the necessary knowledge must be derived (Ansari and Modarress, 1990).

The basic internal structure considered here for the design of SESYS is a micro modular one. The SESYS comprises of five micro modules, namely:

1. Initialisation Module;
2. Method-Rulebase Module;
3. Attribase Module,
4. Algbase Module; and
5. Compute Module.

The integrated micro modules of SESYS are shown in Figure 2.2. The separation into micro modules reduces the complexity of the overall system and facilitates easy modification and enhancements. The initialisation micro module performs preliminary functions such as

(i) the display of the introductory panel(s) which briefly describe(s) the range of SESYS capabilities, and

(ii) the initialisation of the dynamic database as empty base.

The rulebase micro module contains the header and sub-rules to identify appropriate statistical methods the administrators and managers are looking for. The number of header rules depends on the size of the SESYS designed and the statistical area covered by the SESYS. In succeeding the goal, each sub-rule of the rulebase module invokes the attribase module.
The **attribase** micro module is mainly developed for a session of consultation. The consultation consists of a series of questions, which the system asks of the administrators and managers, in order to determine the exact nature of the statistical problem of interest and therefore the appropriate statistical procedure(s) for use (Schutzer, 1989). Once all the questions have been answered by the administrators and managers, the system response is either of displaying the statistical method it recommends along with appropriate computational procedures or of suggesting a consultation with a statistical expert, especially in cases where the problems are complex and unique. The compute micro module is included in the model to execute the identified statistical method. To execute the method, the SESYS seeks input data from the user in the format required. In general, however, the **rulebase**, **attribase**, **algbase** and compute micro modules are called into operation, sequentially.

Figure 2.3 shows the design and implementation of SESYS from a knowledge engineer point of view, including the knowledge base, inference engine and the information flow between the various modules. The knowledge base contains knowledge about the domain of statistics. The SESYS offers advice for a total of 170 statistical methods, comprising statistical analyses, decision-making, estimation, time series, index numbers, and trends/forecasting methods. Consequently, the statistical rule base module contains 170 statistical rule heads to propose 170 methods. The modular design of SESYS would enable new knowledge to be easily added or modified and when SESYS is increasingly used, its expertise will be improved, accordingly.

Once the administrator / manager is ready to begin the dialogue part of a session, one rule from rulebase module is invoked to initiate a dialogue for that particular session. A sample rule is given below:
If the proposed method is regression analysis using three-point method, then the following questions are displayed by attribase module:

1. Do you want advice on statistical decision-making?
2. Do you wish to take decision based on two variables?
3. Does one variable depend on the other?
4. Do you wish to know the relationship between the variables?
5. Do you expect advice to solve the estimation and prediction of the behavior of the variable?
6. Do you have data on the variables?
INTRODUCTORY MODULE

IDENTIFY PROBLEM FACED BY MANAGEMENT

DETERMINE INFORMATION REQUIRED TO HELP MEET THEIR STATISTICAL GOALS

SELECT APPROPRIATE MODULE (SUCH AS ANALYSIS, DECISION, FORECASTING)

STATE OBJECTIVES OF PROBLEM BY ANSWERING QUERIES

STATISTICAL KNOWLEDGE BASE

 STATISTICAL INFERENCE ENGINE

ANY MORE INFORMATION NEEDED?

EXECUTE SUGGESTED METHOD?

INPUT DATA AS SPECIFIED

COMPUTE RESULTS AND FORMULATE RECOMMENDATIONS

CLEAR DYNAMIC DATABASE, NEXT SESSION

HIGHLIGHT ALGORITHM TO SOLVE METHOD

END

FOR REENTRY?

Figure 2.3: Information Flow for Statistical Expert System
One of the most important requirements is that genuine experts exist. These are people generally acknowledged to have an extremely high level of expertise in the problem area; they are significantly better than novices at solving problems in that domain area. Without a source of extensive, powerful knowledge to draw on, the KBS development effort will fail to produce the truly skilful program.

**Rule-Based Technique.** Rule-based representation of knowledge on statistics centres on the use of IF action 1, action 2, ... action n THEN conclusion statements or IF sub-goal 1, sub-goal 2 ... sub-goal n THEN goal.

Each sub-rule of the rulebase module invokes the attribase module that initiates the query phase of the session during which the statistical problem is described.

Once the user has input answers (y/n) to all questions pertaining to the nature of the statistical problem, the answers which describe the statistical problem will then match one of the sub rules within the dynamic knowledge base resulting in a recommendation from the ExpSys.

### 2.4 The Design of the Model

Until recently, the major type of hardware used for ExpSys applications was machine-designed specifically for Artificial Intelligence (AI) applications. Now, The IBM PCs and other compatible computers have been and are used (Ansari and Modarress, 1990) for this purpose. As far as software platform is concerned, there are two ways of developing ExpSys (for example, Smith 1990; Townsend, 1988; Borland International, 1988):

i) Using an AI language like the PROLOG or LISP.

ii) Using a commercially available Expert Systems Shell
Ansari and Modarress (1990) point out that 26 per cent of the companies surveyed in the U.S reported having used AI Languages such as the PROLOG or LISP to develop specific expert system applications. The SESYS modules of the present study have been developed using Turbo Prolog, Version 2.0, Borland International, Inc. which could run on any IBM PC compatible machine.

**Expert System Structure.** The statistical expert system has three primary aspects: a user interface, an inference engine, and stored knowledge of statistics (Figure 2.4). When consulting with the Statistical Expert System, a user states a problem and then interacts with the system via its user interface. This interaction would occur both during the reasoning process and after the reasoning is concluded. The statistical expert system's inference engine is the problem / method processor software that actually carries out the reasoning needed to be solved in a statistical problem.

![Figure 2.4: Structure of a Statistical Expert System](image)

In so doing, it draws on the stored statistical reasoning expertise about the statistical problem area. In addition, it may interact with the user to find out further details about the nature of the problem being solved. When the problem is solved, the inference engine reports the solution to the user and is able to explain its line of reasoning in reaching that solution.
The stored statistical knowledge about a problem area can be represented as a rule set (as collections of predicates). The statistical rule set contains a collection of statistical rules, each of which captures some piece of text/knowledge about how to select the specific statistical method addressed by the statistical expert system. This statistical rule set we can call *Statistical Knowledge Base* (SKB).

### 2.5 The Essential Components of Statistical Expert System

**Statistical Knowledge Base.** The knowledge base of an expert system contains the factual and empirical knowledge of experts in a particular subject area. References such as books or other written materials discussing the domain can form the basis of an initial statistical knowledge base. In a book, an expert has already extracted and organised some of the domain expertise. This organised knowledge might prove useful (at least initially) in building the SESYS (Prerau, 1989).

**Statistical Inference Mechanism.** The inference mechanism of an expert system can simulate, the problem-solving strategy of a human expert.

**Explanation Component.** The explanation component explains the statistical problem solving strategy to the user.

**User Interface.** The user interface employs natural language for dialogues with the user whenever possible.

**Statistical Knowledge Acquisition Component.** The knowledge acquisition component provides support for the structuring and implementation of statistical knowledge in the statistical knowledge base.
2.6 Statistical Expert System Development Team

Statistical Expert (Statistician). Statistical experts provide knowledge of statistical methods for the statistical expert system.

Knowledge Engineer. Knowledge engineer questions the statistical experts, structure the knowledge of statistics and uses it to implement the knowledge base on the statistical techniques.

User. Users state their requirements and ideas, and above all, define the scenario in which statistical expert system will be used.

2.7 A Hierarchical Approach to Conceptual Design

The SESYS software development process comprises of eight phases that follow the problem solving method (Russ et al., 2000):

- Requirements scooping;
- Statistical domain analysis;
- Statistical application analysis;
- Architecture;
- Application design;
- SESYS program implementation;
- SESYS integration; and
- SESYS system test.

The engineering method indicates that the design problems of the SESYS be resolved by first developing very simple statistical solutions and then adding successive layers of detail. To see how this approach may be used in regard to SESYS design (Figure 2.5), there is need to consider a typical information flow for the design process, and then look for ways of stripping away layers of detail until the simplest problem of interest is obtained as a result.

To develop a conceptual design for SESYS, the following are to be done:

- Select module units needed.
Choose interconnections among these modules.
List dominant design variables
Estimate optimum working conditions.
Explain activities under each individual module.
Explain design steps to be carried out for every module.

Figure 2.5: Concurrent Design Process for SESYS
Organising Knowledge of Statistics. The term 'knowledge' means information that a computer program needs before it could behave intelligently.

This information can take the form of facts or rules (Laspeyres Index Method) like those shown below:

FACTS: 1. Prices in the current year is known; 2. Prices in the base year is known; and 3. Quantities sold in the base year is known.

RULES: Laspeyres Index IF The Prices in the current year is known AND The Prices in the base year is known AND The Quantities sold in the base year is known.

Facts and rules in SESYS are not always either true or false; sometimes there is a degree of uncertainty about the validity of a fact or the accuracy of a rule.

The knowledge base on statistics in SESYS contains statistical facts (data) and statistical rules (or other representations), which use those statistical facts as the basis for decision-making. The Statistical Inference Engine (SIE) contains two parts and they are:

i) Interpreter of knowledge on statistics; and

ii) Scheduler of knowledge on statistics.

The interpreter part of the SIE that decides how to apply the statistical knowledge. This analyses the code to decide what action to take next. The scheduler part of the SIE that decides when and in what order to apply different pieces of statistical knowledge. This organisation is shown in Figure 2.6.

Representing Knowledge of Statistics. We discuss how knowledge of statistics is structured in a program, that is, statistical knowledge
representation rule-based method/technique is selected to represent the statistical knowledge. The rule-based technique is discussed later on.

**Verification of Knowledge-Based System.** Once a knowledge base is created for use in an expert system, the verification of knowledge base is crucial in order to ensure the reliability of the knowledge-based system. Literature on verification and validation provides a variety of implementation-dependent definitions, mostly rule-based, and their associated verification methods (Marcos et al., 2000). There is a model-based approach, which consists in the verification of properties defined at the knowledge level.

![Diagram of SESYS structure](image)

**Figure 2.6: The Structure of SESYS**

This approach is also applicable to real-world knowledge-based systems. It is specially useful for systems which are not amenable to
verification using traditional implementation-dependent techniques, that is, those employing a great variety of knowledge implemented using hybrid representations (Chien, 1996; Duftschmid et al., 1998; Marcos, 1999).

Table 2.4 provides an understanding of the various generic categories of expert system applications, the problems addressed by them as well as examples of such expert systems. ASTA, which is a military software, is used to infer situational descriptions from sensor data. This is particularly in reference to the category interpretation. Likewise PROSPECTOR, a geological application, is for inferring system malfunctions from observables in the category of diagnosis. There are thus several systems of applications in use, which are indicated by a small sample of expert systems in the Table 2.4.

2.8 Prolog: The Programming Language

PROLOG (PROgramming in LOGic) is commonly used for programming Expert Systems. This language enables the user to store expert knowledge of humans on particular subjects in the form of facts and rules, and allow non-experts to ask questions, which have a bearing on that knowledge. Further, the purpose of the development of the KBSSP in PROLOG was to allow the psychologists to add facts and rules to the knowledge base, and to ask the computer to extract knowledge from that base in the form of inferences based on those facts and rules.

Development of knowledge-based systems faced a number of practical problems while trying to move this technology into routine use. Two research projects at the University of Liverpool (Bench-capon et al., 1993) have examined problems related to KBS verification and validation, and the associated issue of maintenance. The key factors in determining when it is appropriate to develop knowledge-based systems are the nature, complexity, and scope of the problem to be solved.
Table 2.4: Generic Categories of Expert System Applications.

<table>
<thead>
<tr>
<th>Category</th>
<th>Problem addressed</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpretation</td>
<td>Inferring situational descriptions from sensor data</td>
<td>ASTA (Military)</td>
</tr>
<tr>
<td>Prediction</td>
<td>Inferring likely consequences of given situation(s)</td>
<td>PTRANS (Computer)</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Inferring system malfunctions from observable(s)</td>
<td>PROSPECTOR (Geology)</td>
</tr>
<tr>
<td>Design</td>
<td>Configuring objects under constraints</td>
<td>EVRISKO (Electronics)</td>
</tr>
<tr>
<td>Planning</td>
<td>Designing actions</td>
<td>KNOBS (Military)</td>
</tr>
<tr>
<td>Monitoring</td>
<td>Comparing observations to expected outcomes</td>
<td>VM (Medicine)</td>
</tr>
<tr>
<td>Debugging</td>
<td>Prescribing remedies for malfunctions</td>
<td>TIMM (Computer)</td>
</tr>
<tr>
<td>Repair</td>
<td>Executing plans to administer prescribed remedies</td>
<td>TQMSTUNE (Chemistry)</td>
</tr>
<tr>
<td>Instruction</td>
<td>Diagnosing, debugging, and repairing patient</td>
<td>ATTENDING (Medicine)</td>
</tr>
<tr>
<td>behaviour</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>Governing overall system behaviour</td>
<td>MVS (Computer)</td>
</tr>
</tbody>
</table>


Note: Some expert systems belong to more than one category. Examples were found by the scholar from literature.
Rule-Based Technique for SESYS. Rule-based representation of knowledge on statistics centres on the use of **IF** action 1, action 2, ... action n **THEN** conclusion statements or **IF** sub-goal 1, sub-goal 2 ... sub-goal n **THEN** goal.

For example,

1. **IF**
   
   *The Prices in the current year is known* AND
   *The Prices in the base year is known* AND
   *The Quantities sold in the base year is known*

   **THEN**
   
   *Laspeyres Index.*

2. **IF**
   
   *We want to use more than one independent variable* AND
   *We want to estimate the dependent variable* AND
   *We attempt to increase the accuracy of the estimate*

   **THEN**
   
   *Multiple regression technique*.

When the current statistical problem satisfies or matches the **IF** part of a statistical rule, the action specified by the **THEN** part of the statistical rule is performed (Patterson, 2001). This action may affect the outside world, may direct program control to another module, or may instruct the system to reach a conclusion. This matching of statistical rule, that is, **IF** portion of the facts, will lead to what are called 'inference chains'. The inference chain resulting from successive execution of statistical rules 1 and 2 is shown in Figure 2.7. This inference chain indicates how the SESYS
uses the statistical rules to infer from the particular statistical method used in an analysis.

**Working Memory and Compression.** Expert Systems require a lot of working memory and compression is a heavily used technique in today's computer systems. It is used for audio, image, and video data in multimedia systems to carry out backups, to compress inverted indexes in information retrieval. Compression has two advantages, namely, it reduces costs for storage media (main memory, disk and tape), and it saves IO bandwidth (disk, tape, or network communication) which results in improved performance for IO-bound applications (Westmann et al., 2000). Expert systems and related database and knowledge base applications execute a fair amount of CPU-intensive operations. To give an example: a query could have IO costs of one minute and CPU costs of 30 seconds resulting in an overall response time of one minute if CPU and IO processing can be overlapped.

With compression, the IO costs of the same query could be easily reduced to 30 seconds. With an expensive compression system however the CPU costs can just easily be increased to become more than a minute resulting in an overall higher response time due to compression.

Today, there are several new compression algorithms. But first, how compression could be integrated into a database and knowledge base system rather than inventing new compression algorithms should be explained. Effectiveness of many (variable length) compression techniques depend on efficiently encoding and decoding length information for compressed data/knowledge items. It may be important in this connection to think of a compression that could turn an IO-bound query into a CPU-bound query. Since textbook query execution engines have been designed IO costs, they may be extended to minimise CPU costs as well, in order to take full
advantage of the database / knowledge base compression. In order that the working memory therefore is optimised, we may look at the possibility of query optimisation.

To optimise, we may make the optimiser's cost model 'compression aware'. It may be important at this point to decide whether temporary results should be compressed or not. When large and essential, it is necessary to
compress the temporary results (data/knowledge); otherwise, we could avoid compression. This also means that compression can effectively improve the performance of the database \ knowledge base systems, in addition to providing disk space savings. Studies have shown that compression improves the running times of queries by a factor of two or slightly more in many cases (Freytag and Goodman, 1989; Antoshenkov et al., 1996).

**Design and Development Defect Reduction Strategies.** There are some defect reduction strategies that should be kept in mind in the design and development of an expert system, more so a statistical expert system such as the one we are planning to do here. These strategies, briefly discussed here, seek to transform fad-based practice to an engineering based practice (Boehm and Basili, 2001). Software development, all said and done, remains a people-intensive and continually changing field. As there are possibilities for defects in human efforts of this kind, it is fitting to think of strategies that could reduce defect in the software development. These apply to the expert system we are developing as much to other system development.

1. **Finding and fixing a problem after software delivery is 100 times more expensive than finding and fixing it during the requirements and design phase.**

   This is a major driver in focusing industrial software practice on thorough requirements analysis and design. Good architectural practices have significantly reduced cost escalation.

2. **Current software projects spend 40 to 50 per cent of their efforts on avoidable rework. Twenty per cent of the defects lead to 80 per cent of avoidable rework.**
Rework consists of effort spent fixing software difficulties that could have been discovered earlier and fixed less expensively or avoided altogether. But some works always consist of unavoidable rework. Better user-interactive systems result from emergent processes. Hastily specified requirements and nominal case design and development are the two sources of rework. These often cause major architecture, design and code breakage as well. Building a defect tracking system is beneficial in the long run for system designers.

3. *About half the modules are defect-free while 80 per cent of the defects come from 20 per cent of the modules.*

Experience has shown that between 60 and 90 per cent of the defects arise from 20 per cent of the modules, with a median of about 80 per cent. Finding error-prone modules in a particular environment can prove worthwhile.

4. *About 90 per cent of the downtime comes from 10 per cent of the defects.*

Some defects disproportionately affect a system's downtime and reliability. For example, IBM software products revealed that about 0.3 per cent of the defects accounted for about 90 per cent of the downtime. Therefore, risk-based testing is clearly cost effective.

5. *Peer reviews catch most of the design and software defects. Perspective based reviews catch 35 per cent more defects than non-directed reviews.*

Peer review is one sure way of finding and fixing defects. It therefore provides an effective technique that catches between 30 to 93 per cent of the defects. Number and type of peer
reviews, the size and complexity of the system are important aspects of the defect-reduction strategies.

The union of a several perspectives into a single inspection offers broad yet focused coverage of the system being reviewed. An organisation's defects history may be useful in the generation of perspectives for reviews.

6. *Disciplined personal practices can reduce defect introduction rates by up to 75 per cent.*

There are several disciplined personal processes in practice. For example, use of Cleanroom software development process has shown 25 to 75 per cent reductions in failure rates during testing.

7. *High dependability software products cost more than those of the low-dependability software products.*

Low dependability software costs more than high dependability software. It is because their maintenance is what costs more. High dependability expert system is the ultimate goal in many of the efforts at developing expert systems.

8. *Half the programs in use contain non-trivial defects.*

Worldwide, there are literally millions of programs in use. The creators of softwares and programs face the challenge of providing their tools with the equivalent of seat belts and air bags, along with safe - driving aids and rules of the road. This means further cost to the developers but less cost for the users, in the long run.

2.9 Summary

Having conceived the concept of the SESYS and developed a methodology for writing and implementing it, the discussion in this chapter
has turned to the proposed model for SESYS. In the early pages of the chapter, a comparison of the statistical expert systems in use the world over has been made, through a review and appraisal of the existing literature. This provides the logistics of proposing a model for the SESYS, through discussions on requirements of a typical SESYS, comparison of statistical procedures offered by statistical packages in wider use in various disciplines, and statistical methods for statistical expert system. The model for statistical expert system has then been proposed, with four main components, namely, the knowledge base, the inference engine, the user interface, and knowledge acquisition facility. The basic structure of the SESYS comprising of the five micro-modules – initialisation, rulebase, attribase, algbase and compute modules – has been then erected such that information flow for statistical expert system, and the hierarchical approach to conceptual design, including organising knowledge of statistics, representing knowledge of statistics, and verification of knowledge-based system have been thoroughly modelled. Finally, Turbo Prolog, the programming language, development defect reduction strategies have been discussed so as to make the proposed model complete in all respects. In the next chapter, the logic is extended into the proposed model for SESYS.