A Course in

STATISTICAL ANALYSIS AND SIMULATION

JOHN H. BEAMER
Clarkson College of Technology
Potsdam, New York 13676

MOST CHEMICAL ENGINEERING curricula devote several courses to the topic of model building. In general these courses demonstrate how fundamental laws and relationships in fluid mechanics, mass and heat transfer and thermodynamics can be used to develop mathematical models which describe the complex system being studied. This approach works well when the available theory and evidence is adequate to develop a unique model. However, there are many areas of chemical engineering where such theory often does not exist, such as in kinetics, where one may be faced with a wide variety of possible reaction mechanisms. This is an area where statistics could be useful in selecting the most appropriate model. Unfortunately, statistics is often neglected in chemical engineers' backgrounds and as a result, they are weak in this area of modeling.

In the extreme cases where little or no theory exists, statistics can be used to develop empirical models. Typically, this model may be just that linear equation or polynomial which best fits a given set of data. The use of statistics in the social sciences for such purposes is well established. Recently, there have been a number of attempts at developing models of societal systems, such as in the description of the population and economic growth of a region. Necessarily these models use empirical relationships as there is very little theoretical basis for developing models of such systems. Such studies are often interdisciplinary and frequently a chemical engineer may find himself involved. Certainly this is true for the increasing number of studies concerning the environment or population which now are being undertaken. Therefore, it seems that statistical theory appropriate for building models of various types should be studied by the chemical engineer.

There are several ways in which statistics should play an important role in a chemical engineer's modeling procedures such as in parameter estimation, design of experiments and model discrimination. Where a unique theoretical model exists, the parameters must often be estimated from experimental data. This may be done by the least squares method or by maximizing the likelihood function. The likelihood function is related to the probability distribution for the random variable being measured. Maximizing this function with respect to the parameters in the model determines the values of the parameters that are most likely to produce the observed values of the random variable.

In the problem of deciding between several models, statistical theory can be used to design the placement of experiments so that the maximum amount of discriminating power is given by each experiment. The data may be analyzed to determine which model best fits the data. The analysis of variance, residual analysis and related techniques can be used to further determine the adequacy of the selected model and to discriminate between alternative models.

The analysis of variance allows one to decide if the proposed model adequately describes the experimental data. It also permits one to decide which terms are insignificant and can be eliminated. Residual analysis provides a further check on the adequacy of the model. By plotting the residuals (observed minus predicted values) versus time, time trends in the data can be identified. Similarly, plotting the residuals against other independent variables will indicate if the fundational relationship of each variable in the model is correct. Confidence limits on the model can also be established. A good deal of work in chemical engineering has been done in this area in the problem of modeling reation kinetics. Recent surveys include Bard and Lapidus [1] and Kittrell [2]. Here statistical analysis is used to determine the best model from several possible candidates by first estimating the best parameters for each model and then utilizing the analysis of variance and residual analysis to eliminate inferior models. Once the best model has been determined, the mechanism of the reaction is inferred.

Where no theoretical model exists, the ap-

John Beamer received his B.S. degree from the California Institute of Technology; his M.S. and Ph.D. degrees from Stanford University. He has been employed as Manager, Environmental Systems Division, Urban Systems Research and Engineering, 1969-1970. His areas of interest include optimization and systems modeling and he has taught courses in these subjects since coming to Clarkson in 1970. Current research is in developing optimization theory and applications and in modeling societal systems.

proach most frequently used is to determine the simplest linear model which adequately describes the data. One approach is to propose a model including all variables which may be important. Linear regression is applied to estimate the coefficients for each variable. By using the analysis of variance it can be determined which variables are insignificant in the model and can be eliminated and also to determine when a model has been built which adequately describes the system.

An example of this approach is the societal modeling problem of determining in-migration to a region. There is no theoretical basis for determining what the functional relationship for inmigration should be. Empirically, it has been shown that in-migration can be estimated as a linear function of unemployment and population in that region [3]. Other variables which have lesser descriptive effect are wage rates in the region and the distance of the region from major population sources. Out migration, on the other hand, is mainly a function of the age distribution of the population under consideration and such factors as unemployment rates are relatively unimportant. People in their twenties are the most likely to migrate and they are equally likely to migrate whether they are living in a prosperous region or a depressed one. However, once having made the decision to move, they will most likely choose to move to a prosperous area.

Statistical theory appropriate for building models of various types should be studied by the chemical engineer.

THUS FAR, ONLY deterministic models have been discussed. However, simulation models involving uncertainty (Monte Carlo models) are also very useful. They have been used for problems as diverse as the design of nuclear reactors and the optimization of the location of fire stations. These models require that the probability distributions of the random events involved be known. Using a random number generator, values for the random variable are then generated with the same frequency as the observed values. For example, suppose the percolation of a fluid in a porous media is being studied. A probability distribution for the size of channels is determined. If the channel is above some critical size, the fluid will flow through that channel. A random number is generated to determine whether or not fluid will flow through a given channel in the model. From this and an assumption on how channels are connected, one could determine the degree of penetration of the fluid. Again, statistical theory supplies the tools necessary to carry out such a simulation.

A course for seniors and graduate students has been introduced this year at Clarkson, which introduces basic statistical theory and shows how it is useful in developing the various types of models discussed. The majority of the applications discussed are chemical engineering examples, but other problems such as those noted earlier, are also presented. For example, in the area of empirical modeling, a societal modeling problem is discussed; while a traffic problem, described below, illustrates the use of Monte Carlo methods.

COURSE DESCRIPTION

THE FIRST PART of the course is an introduction to probability theory. Probability density functions are described and the normal, χ^2, t and F probability distributions are introduced. Ensemble averages such as the mean, variance and correlation coefficient are described and the ways of estimating these from sample statistics are given.

Next, certain statistical techniques are discussed, such as parameter estimation techniques, interval estimation and hypothesis testing.

There are several ways in which statistics should play an important role in modeling procedures such as in parameter estimation, design of experiments, and model descrimination.

These methods are then applied to model building. The simplest non-trivial model is a linear model with one independent variable. Parameter estimation by least squares or maximum likelihood techniques are applied to this problem, and confidence intervals are derived. F and t tests, described in the first part of the course, are applied to sample problems to determine if a significant relationship has been obtained. These one-dimensional problems are calculated by hand so that the student develops an understanding of the theory involved and gets a feeling for the problem.

The extension to linear models with several independent variables is straightforward. Homework problems are done on the computer using available linear regression programs. The stress here is on being able to use the tools available and understand what one can or can not do with them. This course adheres to the philosophy that the computer is a useful and necessary tool with which the engineer should be familiar. While analyzing multidimensional linear models, the student becomes familiar with techniques to discriminate between models and to determine whether to add or delete variables to a proposed model. This is done by using analysis of variance and stepwise regression.

The next step is to develop parameter estimation techniques for non-linear models. This is done through least squares minimization, which again requires the student to use existing computer software.

One result of interest which the students verified is that linear and non-linear analysis of a given problem can differ substantially. For example, for the catalytic oxidation of ammonia, a proposed model is:

$$r_{N_2} = \frac{k\ [NH_3]\ [O_2]}{(1+b_1[NH_3]+b_2[O_2]+b_3[H_2O])}$$

This can be linearized as follows:

$$\frac{[NH_3]\ \ [O_2]}{r_{N_2}} = \frac{1}{k} + \frac{b_1}{k}[NH_3] + \frac{b_2}{k}\ [O_2] + \frac{b_3}{k}\ [H_2O]$$

The parameters which appear on the right could then be estimated from a set of experimental data using linear regression analysis. However, this can lead to erroneous results because the effect of the experimental error is altered in the transformation. Regression techniques give unbiased estimates for parameters if the error is normally distributed with zero mean. If the error between observed values for the reaction rate and those predicted by the experimental model has a normal distribution and zero mean, this distribution will be altered by the linearizing transformation. Thus, the frequently used method of linearization in order to use linear regression analysis can result in erroneous results and non-linear estimation should be used in these cases. In non-linear estimation, parameter values are determined by minimizing the sum of the squares of the difference between the observed values of the dependent variable and the value given by the non-linear model under consideration. This requires the use of non-linear optimization techniques. For the example described, the estimates of the parameters by non-linear and by linear analysis differed by more than an order of magnitude.

At this point, the problem of designing experiments to gain the maximum amount of information is described. Often one has to decide between several competing models, and unless experiments are designed effectively, they may be of little value. An example is given of determining the best kinetic model of a chemical reaction from several proposed models by concentrating experiments in the region where the value of the dependent variable of the proposed models differs the most.

THE FINAL PART OF the course goes into a slightly different area, that of simulation modeling. The earlier parts of the course were concerned only with deterministic models. However, the probability and statistical theory that has already been introduced provides the necessary basis for Monte Carlo simulation. This is a tool that has found widespread application in operations research but applications to chemical engineering problems have been very limited.

The basics of simulation techniques are described. One essential is a uniform random number generator which can then be used to generate any required probability distribution, so that events will occur, in the simulation, in accord with the desired probability distribution. In a simulation, one must keep a record of events as they occur and their effect on the system. There are two ways of following the course of events in a simulation, which could be classified as time-oriented and event-oriented. In a time-oriented simulation, the clock is updated one time unit and a check is made to see what events have occurred. In an event-oriented simulation, the clock is updated to the time at which the next event occurs. Each is preferable for certain types of problems.

In applying a simulation to a given problem, the statistics of that problem must be examined. For example, the problem of examining the length of queues at a traffic light was analyzed. It must be determined from experimental data what probability distribution is applicable to describe the arrival of cars, and then the parameter for this distribution must be estimated. A X^2 - test is performed to determine if the theoretical curve adequately fits the data.

Once the probability distribution for the random events involved has been determined, a random number generator can be used to simulate the occurrence of events. Most of the rest of the program consists of maintaining records of desired information.

As an example of Monte Carlo programming a local traffic problem was analyzed and simulated. First data were collected. It was found that at Postdam's busiest intersection the probability distribution for the inter-arrival time between cars is given by:

$$p(t) = \frac{1}{4.7} \exp \left(- \frac{(t-2.5)}{4.7} \right)$$

for $t \geq 2.5$.
The arrival of cars can be simulated by the relationship,

$$t = 2.5 - (4.7)\ 1n\ Y$$

where Y is a random number between 0 and 1

Data on operation of the traffic light were also gathered. The length of time between signal changes is not constant, but is affected by pedestrian traffic demands, which also could be approximated by an exponential distribution. A computer program was written and information such as the average waiting time per car was collected. Then alternative traffic control options could be considered to determine how they would affect the average waiting time.

The application of Monte Carlo methods to chemical engineering has been limited but some work related to it has been done. In numerical analysis the technique may be used to approximate solutions to various problems such as evaluating complex integrals or solving differential equations. Physical problems also have been solved by this technique [5]. Problems involving collisions between molecules have been analyzed by this method, such as the design of nuclear reactor shielding and statistical mechanical problems. As previously mentioned, problems involving percolation of liquids in porous media have been studied by this approach. The traffic problem analyzed in class demonstrated the methodology and accents the wide range of applicability of the subject. It is hoped that looking at problems of a diverse nature will help chemical engineers approach their problems with a fresh viewpoint and stress the wide range of applicability of the methods.

Simulation models involving uncertainty have been used for problems as diverse as design of nuclear reactors and optimization of fire station location.

This course is designed to complement those courses which emphasize model building from a theoretical basis by demonstrating the ways in which the methods of statistics can be used to aid model building. It provides the student with background that is useful for continued study in the diverse areas in which statistical concepts are employed. These would include for example, statistical mechanics, quality control and kinetic theory. The text used for the course is Himmelblau's "Process Analysis by Statistical Methods" [4], which covers the basic statistical theory. Examples have been taken from a wide range of sources. Hammersley and Handscomb [5] was used as the primary source of material on Monte Carlo methods.

REFERENCES

1. Bard, Y. and L. Lapidus, "Kinetics Analysis by Digital Parameter Estimation," *Catalysis Reviews, 2*, 1, 67-112 (1968).
2. Kittrell, James, "Mathematical Modeling of Chemical Reactions," *Advances in Chemical Engineering, 8*, 97-183 (1970).
3. Hamilton, H. R., et al., "A Dynamic Model of the Economy of the Susquehanna River Basin," Battelle Memorial Institute Research Report (1966).
4. Himmelblau, D. M. "Process Analysis by Statistical Methods," John Wiley & Sons, Inc., New York, (1970).
5. Hammersley, J. M., and D. C. Handscomb, "Monte Carlo Methods," John Wiley & Sons, Inc., New York, (1965).