Statistical Education via Simulation.

Jorge Luis Romeu
State College at Cortland, Cortland NY 13045

Keywords: Statistical education, simulation, data analysis.

Abstract

Discrete event simulation has nurtured from statistical analysis for many years. The converse is not true. However, recent advances in computer technology and software development has made it possible to have PC's running specialized simulation languages, readily available. This paper discusses how discrete event simulation, developped via specialized simulation languages (e.g. GPSS) can become a useful class resource to teach and motivate statistics students. In addition, simulation helps to present, more effectively, interdisciplinary case studies, to increase group learning and to relieve students and instructors from statistical drudgery. Examples of such GPSS simulation are developed. [1]

1 Introduction.

For some time now, statisticians have been trying to change the way statistics is taught in college. For example, the 1992 ASA Winter Conference, in Louisville, KY, was dedicated to Teaching. The Statistical Education Section and some hard copy and electronic journals are dedicated to improving the teaching of statistics (e.g. the February 1995 issue of The American Statistician). Many statisticians agree that it is time to change the way statistics is taught, especially to non majors, applied statistics students and practicing professionals.

For example, Dr. Hogg describes the workshops in statistical education (Hogg, 1985, 1991) and discusses the need for clearly specifying the objectives that statistics service courses should pursue, according to the audience and the level of the students to which it is intended for.

Prof. Bisgaard describes his engineering statistics course (Bisgaard, 1991) where he uses practical experimentation in the classroom to motivate the students and to introduce statistical concepts. Bisgaard states that such live experiments constitute the core of his entire course, since he uses them to introduce basic concepts in experimental design, data collection problems, parameter estimation and statistical comparisons of the results.

With respect to teaching applied statisticians, Kettenring argues that industry needs problem solvers who are able team players and can communicate effectively their findings (Kettenring, 1995). He states that in industry, problems are interdisciplinary (as in most real life environments). Therefore, bridges should be built between disciplines. Bickel also argues in favor of an interdisciplinary approach to teaching statistics (Bickel, 1995) and lists some topics and types of talents that statisticians should have.

In the same vein, Ross expresses the requirements to succeed as a government statistician (Ross, 1995). Again the demand for *team players* arises due to the interdisciplinary problems encountered. The *entreprenurial* element, that is the drive to search within one's own organization, for potential statistical problems that would not otherwise be proposed to the government statistician by the (non statistician) client, is another positive characteristic. Bailar states that Academe fails to involve students because we teach heavily abstract Statistics topics, unrelated to the student taking a service course (Bailar,

[1] To Prof. J. P. Vilaplana, of the U. of the Basque Country; statistician, mentor and friend.

1995). He argues that there must be cuts in the present curriculum; that priorities must be established and that specific topics must be selected. Also that the range of Statistics electives should be widened even at the expense of currently required material, perhaps far too theoretical.

Finally, Lehoczky discusses ways of modernizing the doctoral curriculum (Lehoczky, 1995). He describes the Carnegie Mellon Ph.D. program (masters as well as upper level courses) and argues that it is evenly distributed between theory and the practice of Statistics. In addition, this program trains doctoral students in teaching, in written and oral communication and in cross disciplinary activity.

All of the above shows that there is a serious and longstanding concern among the statistical community about the teaching of our subject. And that a serious quest for new methods of teaching it, in a more practical and interdisciplinary manner, is justified.

This paper presents one such method: teaching Statistics through simulation modelling and output analysis. With this approach, we belive applied and service students can be better reached and motivated. And that some students' perception that Statistics is a useless, dull course requirement can be changed to that of a useful and interesting topic, worth dedicating time to. We must remember that statistics competes for the students' attention and time among several other subjects and extra curricular activities.

In addition, simulation allows the application of several new educational approaches such as *workshops and cooperative learning.* Traditional lecturing lends poorly to developing these new techniques. Finally, via simulation we can *teach more statistical methods to undergraduate science and engineering students.* Many undergraduate curriculums today teach (at most) two statistics courses. And most undergraduates never pursue a masters degree, even when they deal with data on a daily basis, in research, development and production environments.

The rest of this paper expands on these problems. Section 2 discusses several uses of simulation in the teaching of Statistics and the advantages of using simulation packages. Section 3 deals with several types of teaching approaches using simulation and discusses their advantages and disadvantages. Section 4 briefly presents Monte Carlo methods as a teaching tool. Finally, in Section 5, we summarize our work.

2 Simulation in Teaching.

There are several ways of teaching Statistics: from the traditional lecturing approach to that of developing in-class experiments, as in Bisgaard (1991). The first, essentially uses book examples to illustrate the theory, often remote from the students' main interests and lacking in data collection and manipulation. The second, the experimental approach, is complicated in nature and even risky (e.g. the ladder experiment may be prone to accidents). Or at least, time consuming (inside and out of the classroom) so that only a very limited number of real life experiments can be carried out during the semester.

Simulation modelling and analysis, on the other hand, is an intermediate solution, halfway between the above two. Simulation still retains the flavor of the uncertainty in the experimental outcomes, as occurs in real-life. This uncertainty is created by the simulation model (for it represents a real situation). On the other hand simulation does not require that the student spends hours in or outside the classroom, gathering data. For, the student can do this automatically during the simulation run. Students are no longer subjected to the inherent physical risks of an experimentation process. Nor they incur in the time and money expenses that some experiments necesssarily bring about.

Throughout this paper, we will refer to simulation as the art of modelling in a computer, via a specialized language (e.g. GPSS), a real system whose operation through time we study by running the computer model. This definition differs from that of (static, Monte Carlo) simulation, sometimes used when teaching elementary service courses. There, a set of data following a pre-specified distribution (say the normal), is generated via some statistical package (e.g. Minitab). Then one performs some simple statistical procedures on them (say, a test of hypothesis).

Such elementary uses of static (Monte

Carlo) simulation are adequate for introductory courses, but will not be discussed further in this paper. We, in turn, will concentrate in the pedagogical uses of discrete event simulation models in intermediate and advanced service and/or applied statistics courses.

In the past, simulation was seldom used in teaching Statistics. At least two good reasons can be argued for this. One is that simulation models were written in an (HOL) programming language (e.g. FORTRAN). These models were time consuming and difficult to implement. The second reason is that specialized simulation languages (e.g. GPSS) were only available in main frame computers making them (i) expensive and (ii) constrained to a small group of specialized users.

With the advent of the PC and the subsequent transfer to it, of many of these specialized simulation languages, the above two disadvantages have all but disappeared. First, it is now possible to write a complicated and challenging simulation model with relative ease. And they are still easier to run by the students. Then, there are several new simulation books that include an inexpensive student version (e.g. Karian and Dudewicz (1991), Thesen and Travis (1992), Schreiber (1991)). Such versions are, for all practical classroom purposes, quite adequate.

The approach presented in this paper was developed from several years of teaching applied Statistics, then simulation and finally Statistics via simulation. And more recently, from our experiences in teaching short courses in the uses of simulation in the intermediate and advanced statistics courses, to Statistics Faculty of several Mexican and Spanish universities. And from helping them implement these techniques in their classrooms. In general, Statistics Faculty have been very positive to it, and so have their students.

3 Simulation Approaches.

We propose here three different approaches to teaching Statistics and/or stochastic Operations Research (OR) courses via simulation with a specialized (e.g. GPSS) language. And these depend on the level of expertise the instructor has with the language. They are: (i) simulation as an independent, statistically oriented course; (ii) simulation as a companion course lab and (iii) simulation as integral part of the statistics course.

As an independent course (as it is most often taught today) simulation is an excellent complementary course for statistics and OR students. In them, students have an opportunity to apply, in an interdisciplinary environment (models may differ widely) multiple data analyses techniques.

Majors in industrial engineering, OR, systems analysis, computer science and business form the bulk of this student body. However, we should encourage (even require) applied or service course statistics students with at least two semesters (including regression and ANOVA) to take simulation as an applications course.

Requirements to teach simulation include, in addition to statistics, knowledge of a specialized simulation package (e.g. GPSS) and of the simulation techniques themselves. For, simulation is an art that includes programming, systems modelling and statistical analysis. Hence, its great potential for students in the statistical data analysis area.

As a Lab course, companion to the statistics course, simulation is excellent and the requirements for the instructor are much smaller. One does need a minimum level of understanding of the simulation activity and language. But an introductory level of programming in languages such as GPSS, is not difficult to acquire. And the examples in the introductory simulation books are easily understood and can be quickly modified to provide useful applications of statistical methods.

The Lab would follow the theory class, providing an inexpensive alternative to physical experimentation. One would explain the physical meaning, needs and objectives of the simulation model in question. Then, the student would be provided with the simulation program and an instruction sheet on how to run it. And of course, an individual seed. This is the key issue.

With individual seeds we guarantee individual and even possibly conflicting results. And such capability provides many advantages. First, students can now work in *teams* and cooperate. Different seeds generate different random samples. With this, students can engage in **cooperative learning** while re-

maining individually accountable for their results.

In addition, if we have a class with enough students (say 20 or more) and we test at, say level $\alpha = 0.05$, it is likely that one or more of the students obtain, by chance, contradictory results. We can make excellent use of such contradictory results to elaborate on the practical implications of testing, on the size of the test and on the sample size.

We can also control the model variables and the model itself. In real life, one seldom knows the exact model, nor all the model assumptions hold. One usually is not even sure which is the best approximation to the *real* model. In simulation, since we are *building* the model, it is totally under our control.

This capability also allows us to *violate* model assumptions and have the students assess its impact in the experimental results. This way we show (i) practical importance of checking model assumptions and also (ii) which assumptions are more important and which can be relaxed and to what extent.

Another advantage of the realistic character of the examples presented, via simulation, is the introduction of statistics students into potentially inter-disciplinary problems and team work. Finally, if a report and a presentation are required from each student or team of students (if using cooperative learning techniques) the communication skills, oral and written, are also developed.

Labs can take place once a week, for two or three hours. Students first must understand the simulation model and the learn how to operate it. Finally, the objectives of the statistical analyses are explained and the task is rehearsed. The students, then and on their own, design and develop the computer experiments and obtain the results.

The third possibility is that of using simulation as an intrinsic part of the Statistics course. This approach has yet smaller requirements on the instructor. One only needs to operate the simulation programs and understand how to change the control values of the variables under analysis. One instructor can act as a *focal* point. This instructor develops and passes on to the other instructors the required simulation programs for their use. The focal point does need an intermediate level of programming in such simulation languages. However, the other instructors only need to have a basic knowledge of the programming language.

Under this setting, the course instructor would use, as examples, the simulations in class. And out of class, as homework and/or project materials. An example of a mid-complexity problem, consisting of the modelling and optimization of the operation of a network of small dams, is shown in Table 1. This problem yields several discrete variables (i.e. number of dams, maximum and minimum dam capacity, replenishment policies). It also provides several continuos variables (e.g. pumping rates and costs, rain distribution) and responses (e.g. Total/partial operating costs, number of water transfers from dam to dam, stock-out probabilities, days to stock-out) One can easily form cooperative learning groups (teams) and still have students working independently, if different seeds are assigned to different students. Also, different sets of independent variables and different responses can be assigned to different members of a team. This approach insures *individual accountability* and preserves group learning.

Many statistical topics can be reviewed via simulation. For example, in Figure 1, one can have different numbers of dams, with different (max/min) water capacities (replenishing levels). One can also have different costs associated with inter-dam water transfers. And one can have different rain distribution schedules (dry, rainy season). All this provides multiple independent variables.

One can then design as many replications (or batches) as one needs and use (i) ANOVA or ANCOVA, (ii) multiple regression, (iii) goodness-of-fit techniques, (iv) residual analysis, (v) methods of variable selection and (vi) response surface methodology, just to mention a few. And one can optimize the operation of this system of dams, under different conditions.

In addition, if the simulation batches or replications are made short enough, one will violate the assumptions of normality and independence of the observations. One can use this to discuss the importance of residual checking and analysis and of data transformations. Discrete variables can be used

in ANOVAs and continous ones in regression. Combinations of discrete and continous variables permit the implementation of ANCOVAs. There is much flexibility and realistic flavor in these different combinations, yielding many interesting models and statistical procedures.

In addition, since one is able to computer generate and store in files these statistical results, regardless of sample size, students can easily use statistical packages in the course. Under such setting, students' numerical results and data will be (i) meaningful for them (not retrieved form a package library) without (ii) having to collect and enter them by hand, as one would if they had been obtained from a physical experiment.

4 Monte Carlo Simulations.

Monte Carlo is also a very useful tool for teaching Statistics. Hence, we would not like to conclude without dedicating a few paragraphs to its many potential educational uses.

Assume one gives the students of an intermediate or advanced Statistics course a final proyect to compare the power of two or more hypotheses tests. For example, students are requested to compare several multivariate normality goodness-of-fit tests (Romeu, 1994a). And one gives them the assignment to do so under a subset of non normal alternatives, say under kurtic or skewed alternatives.

Students can then analyze several important concepts. First, the impact of the dimensionality of the problem. Then, the importance of defining the specific type of alternative one is covering against. Also, the effect of the sample sizes and the inter-correlation between the variable components.

Then, students can get into different methods of random variate generation and compare them (by testing them for fit). This also helps uncover the nature and relevance of the concepts of skewness and kurtosis. And of how to measure and interpret these concepts in a data set or in the shape of a density function. Many Statistics students have a good handle of the theory but are unable to use theoretical results in practical situations. But in real-life data analysis, one needs to see through the data, during the EDA phase of a study.

Finally, Monte Carlo simulation gives students another opportunity to use statistical packages and another excuse to brouse through statistical papers, becoming more familiar with statistical journals.

5 Conclusions.

During many years, simulation modelling has nurtured from Statistics. Simulation output analysis is another good example of applied Statistics. On the other hand, statisticians have seldom used simulation in their teaching. For, writing simulations in a HOL programming language was complicated and because simulation packages were difficult to get and operate.

With the advent of the PC, these two problems have been minimized. Today one has ready access to good simulation packages such as GPSS, easy to learn and to use by students and Faculty.

Simulation, in turn, provides an intermediate approach to pure lecturing or to physical experimentation. Computer or simulation modelling is easier and faster to implement and to learn and more efficient to control in the classroom. It provides more flexibility for the Statistics instructor, by allowing a small simulation model to serve as a teaching tool for several statistical methods, with the adequate change of a few parameters and instructions.

This author has experimented with this approach, with excellent results, in several institutions in the U.S. and abroad (Romeu, 1986 and 1994b). This author has also conducted short training workshops in teaching statistics with simulation in GPSS, during his recent Fulbright Lecturing Award, in several universities in Mexico and also in Spain. Student evaluations and responses to this approach have consistently been very positive.

References

[1] Bailar, John C.; "A Larger Perspective"; The American Statistician. Vol 49, Number 1. February 1995.

[2] Bickel, Peter J.; "What Academia Needs"; The American Statistician. Vol. 49, Number 1. February 1995.

[3] Bisgaard, S.; "Teaching Statistics to Engineers"; The American Statistician. Vol 49, Number 1. February 1991.

[4] Karian, V and E. Dudewicz; "Modern Statistical, System, and GPSS Simulation: The First Course"; Freeman. 1991.

[5] Hogg, R., et al.; "Statistical Education for Engineers: An Initial Task Force Report"; The American Statistician. Vol. 39. 1985.

[6] Hogg, R.; "Statistical Education: Improvements are Badly Needed"; The American Statistician. Vol 45, Number 4. 1991.

[7] Kettenring, Jon R.; "What Industry Needs"; The American Statistician. Vol. 49, Number 1. February 1995."

[8] Lehoczky, John; "Modernizing Statistics Ph.D. Programs"; The American Statistician. Vol. 49, Number 1. 1995.

[9] Romeu, J. L.; "Teaching Engineering Statistics With Simulation: A Classroom Experience"; The Statistician (RSS Series D); Vol 35. 1986.

[10] Romeu, J. L.; "Validation of Multivariate Monte Carlo Studies"; Proceedings of the IMSIBAC-4. Edited by M. L. Puri and J. P. Vilaplana. VSP Int. Science Publ. (Netherlands). 1994a.

[11] Romeu, J. L.; "Simulacion, Pedagogia y Estadistica". Actas del Primer Seminario de la Ensenanza de las Matematicas. ITAM. Mexico. 1994b.

[12] Ross, P. ; "What Government Needs"; The American Statistician. Vol. 49, Number 1. February 1995.

[13] Schreiber, T.; "An Introduction To Simulation Using GPSS/H". Wiley. 1991.

[14] Snee, R. et al.; "Preparing Statisticians for Careers in Industry"; The American Statistician. Vol 34. 1980.

[15] Thesen A. and L. Travis; "Simulation for Decision Making"; West. 1992.

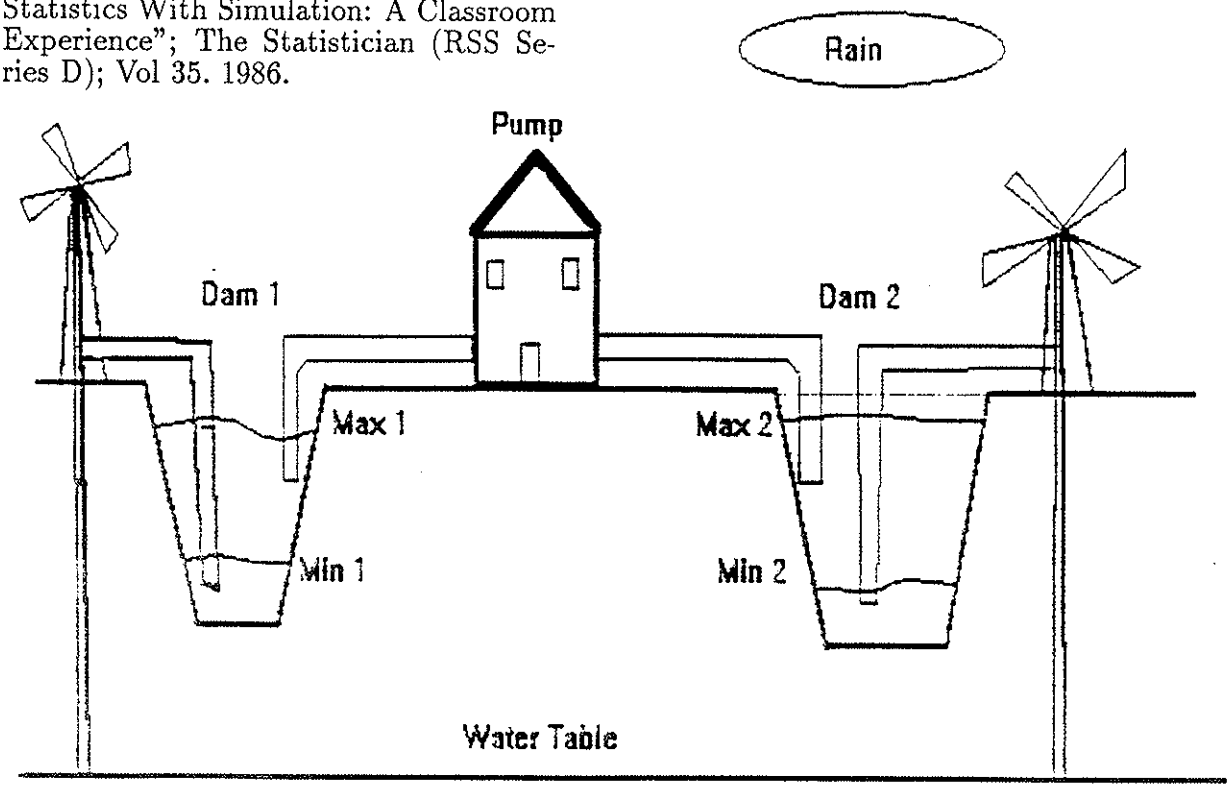

Figure 1: Sketch of Dam System