# COMPUTER GENERATION OF STATISTICAL DATA

**Robin H. Lock**
**St. Lawrence University**
**Canton, NY**

The demand for data in applied statistics courses has increased dramatically in recent years as the growth in computer technology has enabled students to perform more sophisticated analyses on larger and more complicated data sets. This has increased the burden on instructors and textbook authors to supply interesting data to illustrate desired concepts and allow students to practice techniques. We will describe some ways the computer itself can be used to help satisfy the demand for data.

Specifically, we will discuss three programs which generate individual data sets with varying amounts of student interaction and control over the experimental design. STOCKS creates data according to various Box-Jenkins models for use in time series analysis. SAMPLER generates samples from common probability distributions which the students are then asked to identify. GOLFBALLS simulates golf shots where the user can control the characteristics of the ball and ability level of the golfer.

In each case the data are automatically stored in a file suitable for input into most statistical packages. This eliminates the burden of entering large amounts of data by hand and allows students to concentrate more on the actual analysis. Also, since data sets are computer generated, each student receives a unique set of values, although the underlying model might be the same for an entire class.

A key advantage to using the computer to simulate data is that the "right" answer is known to the instructor and can eventually be revealed to the students. This allows the instructor to control the difficulty of the assignment or to generate data for classroom use which illustrate particular concepts. Furthermore, students seem to really enjoy the challenge of trying to discover the underlying model or structure of the data.

## STOCKS

This program creates data which is assumed to represent a time series of weekly closing prices for a fictitious stock (although the data can be fictionally attributed to many other sources). Using the notation of Box and Jenkins (1976), the underlying model is of the form:

$$Z_t = \emptyset_1 Z_{t-1} + \emptyset_2 Z_{t-2} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}$$

where, $a_t$ = a random error term, $a_t \sim N(0, \sigma^2)$

$Z_t$ = the differenced series $Z_t = \nabla^d Y_t$

$Y_t$ = the actual stock price series

$\emptyset_1, \emptyset_2, \theta_1, \theta_2$ = autoregressive and moving average parameters

Thus the prices follow an ARIMA (p,d,q) model with each of the parameters p, d, and q being either 0, 1, or 2. The student's task is to identify an appropriate model, estimate coefficients, and forecast the future behavior of the stock.

The students do not actually run the STOCKS program. Instead it may be used by the instructor to generate sets of data which are then distributed to the individual students or stored is some sort of public access file. The program may also be used to provide examples for classroom discussion or exams which exhibit characteristic features of various time series models.

For each data set, the instructor specifies:

- the number of values to be generated in the series
- values for autoregressive and/or moving average terms (a zero indicates no such term)
- the number of differences (d)
- the variance for the random error term
- a seed for the random number generator

The STOCKS program automatically checks that the requested model is stationary and invertible. The $Y_t$'s are calculated by generating an i.i.d. sequence of (pseudo)- random errors, calculating the $Z_t$'s, and accounting for the differencing and rounding to get $Y_t$'s. The mean, minimum, and maximum of the resulting series are then displayed. At this point one may choose to add a constant to the entire series (negative stock prices are difficult to explain!). Finally, the entire series is displayed and, if it looks satisfactory, may be stored in a file.

The amount of noise in the system can be adjusted by changing the variance of the random errors. This affects the degree to which the sample autocorrelation and partial autocorrelation plots agree with theoretical models. It is even possible to use exactly the same sequence of random errors to generate several data sets according to different ARIMA models. The eventual time series can appear completely different, although the residuals will be very similar when an appropriate model is fit to each.

## SAMPLER

We will give a brief description of SAMPLER here. A more detailed discussion can be found in Lock (1986). The basic idea for the program is quite simple. Each student is assigned a project code number which uniquely determines a set of five distributions which they are to identify. Using the computer, they can request samples from any of their five distributions, although they are limited to a maximum of 500 data points in all. Thus they generally start by requesting a fairly small sample from each unknown distribution, making some preliminary guesses, and then taking further samples from the more difficult cases.

Although the students do not know their particular distributions, they may be given a list of potential candidates such as: normal, binomial, discrete or continuous uniform, exponential, or Poisson. To avoid a brute force process of elimination approach we have also included the sum of pairs of dice rolls in the program, but only identified it on the list as a "mystery" distribution. It is interesting to observe what little feel most students have for how random a sample might be. Typically, if they don't get a perfect bell-shaped curve, they will go to great lengths to create a mystery distribution which describes every glitch in the sample data.

The skills needed to successfully complete the SAMPLER exercise cover a fairly broad range of statistical concepts and techniques. Naturally, a thorough knowledge of the theoretical distributions and their properties is required. Graphical abilities are used in analyzing histograms and other plots for preliminary identification. A variety of estimation techniques are needed to determine the parameters for each unknown distribution and check that they agree with theoretical calculations. Formal hypothesis tests, such as the chi-square or Kolmogorov-Smirnov, can be employed to check goodness-of-fit. Students often get a good feel for the relative power (or lack of it) when, for example, a small sample might simultaneously "pass" a chi-square test for normal, uniform, and even binomial distributions!

## GOLFBALLS

This is the most involved of our simulation programs. For each "shot", the response variables are distance (DIST) and left/right deviation (DEV). The experimenter may control the ability level of the golfer (Pro, Amateur, Duffer), color of the ball (White, Orange), compression (70-110 p.s.i.), and type of construction (One-piece, Inner core). Contrary to our previous examples, the underlying model for GOLFBALLS is the same for all students, although its structure is rich enough to allow each member of a class to investigate a different aspect of the data. Typical project ideas might include:

- Find the type of ball which maximizes DIST for Pros.

- Do Orange balls travel farther than White?

- How are DIST and DEV related?

- Find a model to predict DEV for Duffers.

Obviously the sophistication of the project can range from a simple confidence interval or hypothesis test to a complex analysis of variance or regression model.

The sampling scheme in GOLFBALLS is designed so that the computer may select the characteristics of each of the balls randomly, or the student may specify a precise experimental design which controls the levels of any or all of the independent factors. As with STOCKS and SAMPLER, the simulated results, including coded values for independent variables, are displayed on the screen and stored in a file.

GOLFBALLS also contains a banking facility which limits the number of shots which can be requested. Each student is assigned a project number and given a budget from which they "pay" their sampling costs: $3 for each pro shot, $2 for Amateurs, and $1 for Duffers. A specially coded bank file keeps track of each student's account and is adjusted whenever sample values are produced. This makes students keenly aware of how sampling costs can affect the design of an experiment, something which they rarely encounter in the classroom but invariably experience later on.

The analysis of the GOLFBALLS data has some interesting features. For example, DIST exhibits a great deal of variability whenever more than one type of golfer is included, so students quickly learn to model each ability level separately. DEV is coded so that a negative value denotes a shot to the left (a hook) and positive numbers go to the right (a slice). This causes a surprising amount of difficulty for students since textbook data seems to rarely include negative numbers. Often a student happily obtains an average DEV which is close to zero before realizing that they really are interested in the average $|DEV|$. Another problem arises in analyzing data for Duffers since, occasionally, they miss the ball completely! How should those zeros be handled and do they miss Orange balls more often than White?

## Summary

These programs and related assignments are designed to be very flexible and have been used in a variety of courses including introductory and second level applied statistics, time series analysis, statistical computing, and advanced courses in regression and experimental design. Students report that the simulation projects are a valuable part of these courses, giving them a better feel for how statistical investigations really work, and that they are interesting, challenging, and fun to work on. A student, who graduated several years ago, visited campus recently and could still enthusiastically recall details of the GOLFBALLS data.

The programs themselves were written in BASIC for our IBM mainframe and have been converted to run in Microsoft BASIC on the IBM-PC. An attempt has been made to keep the programming as simple as possible to

facilitate transfer to other systems and modifications by other instructors. Copies of any of the programs are available by request. Softcopy can be obtained by sending an appropriate diskette.

## References

Box, G.E.P., & Jenkins, G.M. (1976). Time series analysis. San Francisco: Holden-Day.

Lock, R. (1986). SAMPLER: A Computer Simulation in Statistics, Collegiate Microcomputer, IV(1).