Monte Carlo Studies in Statistics

In statistical inference, certain properties of the test statistic or estimator usually must be assumed to be known. In simple cases, under rigorous assumptions, we have complete knowledge of the statistic. In testing a mean of a normal distribution, for example, we use a t statistic, and we know its exact distribution. In other cases, however, we may have a perfectly reasonable test statistic but know very little about its distribution. For example, suppose that a statistic T, computed from a differenced time series, could be used to test the hypothesis that the order of differencing is sufficient to yield a series with a zero mean. If enough information about the distribution of T is known under the null hypothesis, that value may be used to construct a test that the differencing is adequate. This, in fact, was what Erastus Lyman de Forest studied in the 1870s, in one of the earliest documented Monte Carlo studies of a statistical procedure. De Forest studied ways of smoothing a time series by simulating the data using cards drawn from a box. A description of De Forest's Monte Carlo study is given in Stigler (1978). Stigler (1991) also describes other Monte Carlo simulations by nineteenth-century scientists and suggests that "simulation, in the modern sense of that term, may be the oldest of the stochastic arts."

Another early use of Monte Carlo was the sampling experiment (using biometric data recorded on pieces of cardboard) that led W. S. Gosset to the discovery of the distribution of the t-statistic and the correlation coefficient. (See Student, 1908a, 1908b. Of course, it was Ronald Fisher who later worked out the distributions.)

Major advances in Monte Carlo techniques were made during World War II and afterwards by mathematicians and scientists working on problems in atomic physics. (In fact, it was the group led by John von Neumann and S. M. Ulam who coined the term "Monte Carlo" to refer to these methods.) The use of Monte Carlo methods by statisticians gradually increased from the time of De Forest, but after the widespread availability of digital computers, the usage greatly expanded.

In the mathematical sciences, including statistics, simulation has become an important tool in the development of theory and methods. For example, if the properties of an estimator are very difficult to work out analytically, a Monte Carlo study may be conducted to estimate those properties.

Often the Monte Carlo study is an informal investigation whose main purpose is to indicate promising research directions. If a "quick and dirty" Monte Carlo study indicates that some method of inference has good properties, it may be worth the time of the research worker in developing the method and perhaps doing the difficult analysis to confirm the results of the Monte Carlo study.

In addition to quick Monte Carlo studies that are mere precursors to analytic work, Monte Carlo studies often provide a significant amount of the available knowledge of the properties of statistical techniques, especially under various alternative models. A large proportion of the articles in the statistical literature include Monte Carlo studies. In recent issues of the *Journal of the American*

Statistical Association, for example, almost half of the articles report on Monte Carlo studies that supported the research.

1 Simulation as an Experiment

A simulation study that incorporates a random component is an experiment. The principles of statistical design and analysis apply just as much to a Monte Carlo study as they do to any other scientific experiment. The Monte Carlo study should adhere to the same high standards of any scientific experimentation:

- control
- reproducibility
- efficiency
- careful and complete documentation

In simulation, *control*, among other things, relates to the fidelity of a *nonrandom* process to a *random* process. The experimental units are only simulated. Questions about the computer model must be addressed (tests of the random number generators, etc.)

Likewise, *reproducibility* is predicated on good random number generators (or else on equally bad ones!). Portability of the random number generators enhances reproducibility and in fact can allow *strict* reproducibility. Reproducible research also requires preservation and documentation of the computer programs that produced the results (see Buckheit and Donoho, 1995).

The principles of good statistical design can improve the efficiency. Use of good designs (fractional factorials, etc.) can allow efficient simultaneous exploration of several factors. Also, there are often many opportunities to reduce the variance (improve the efficiency). Hammersley and Hanscomb (1964, page 8) note,

> ... statisticians were insistent that other experimentalists should design experiments to be as little subject to unwanted error as possible, and had indeed given important and useful help to the experimentalist in this way; but in their own experiments they were singularly inefficient, nay negligent in this respect.

Many properties of statistical methods of inference are analytically intractable. Asymptotic results, which are often easy to work out, may imply excellent performance, such as consistency with a good rate of convergence, but the finite sample properties are ultimately what must be considered. Monte Carlo studies are a common tool for investigating the properties of a statistical method, as noted above. In the literature, the Monte Carlo studies are sometimes called "numerical results". Some numerical results are illustrated by just one randomly

generated dataset; others are studied by averaging over thousands of randomly generated sets.

In a Monte Carlo study there are usually several different things ("treatments" or "factors") that we want to investigate. As in other kinds of experiments, a factorial design is usually more efficient. Each factor occurs at different "levels", and the set of all levels of all factors that are used in the study constitute the "design space". The measured responses are properties of the statistical methods, such as their sample means and variances.

The factors commonly studied in Monte Carlo experiments in statistics include the following.

- statistical method (estimator, test procedure, etc.)
- sample size
- The problem for which the statistical method is being applied, that is, the "true" model, which may be different from the one for which the method was developed. Factors relating to the type of problem may be:
  - distribution of the random component in the model (normality?)
  - correlation among observations (independence?)
  - homogeneity of the observations (outliers?, mixtures?)
  - structure of associated variables (leverage?)

The factor whose effect is of primary interest is the statistical method. The other factors are generally just blocking factors. There is, however, usually an interaction between the statistical method and these other factors.

As in physical experimentation, observational units are selected for each point in the design space and measured. The measurements, or "responses" made at the same design point, are used to assess the amount of random variation, or variation that is not accounted for by the factors being studied. A comparison of the variation among observational units at the same levels of all factors with the variation among observational units at different levels is the basis for a decision as to whether there are real (or "significant") differences at the different levels of the factors. This comparison is called analysis of variance. The same basic rationale for identifying differences is used in simulation experiments.

A fundamental (and difficult) question in experimental design is how many experimental units to observe at the various design points. Because the experimental units in Monte Carlo studies are generated on the computer, they are usually rather inexpensive. The subsequent processing (the application of the factors, in the terminology of an experiment) may be very extensive, however, so there is a need to design an efficient experiment.

2 Reporting Simulation Experiments

The reporting of a simulation experiment should receive the same care and consideration that would be accorded the reporting of other scientific experiments. Hoaglin and Andrews (1975)outline the items that should be included in a report of a simulation study. In addition to a careful general description of the experiment, the report should include mention of the random number generator used, any variance-reducing methods employed, and a justification of the simulation sample size. The *Journal of the American Statistical Association* includes these reporting standards in its style guide for authors.

Closely related to the choice of the sample size is the standard deviation of the estimates that result from the study. The sample standard deviations actually achieved should be included as part of the report. Standard deviations are often reported in parentheses beside the estimates with which they are associated. A formal analysis, of course, would use the sample variance of each estimate to assess the significance of the differences observed between points in the design space; that is, a formal analysis of the simulation experiment would be a standard analysis of variance.

The most common method of reporting the results is by means of tables, but a better understanding of the results can often be conveyed by graphs.

3 An Example

One area of statistics in which Monte Carlo studies have been used extensively is robust statistics. This is because the finite sampling distributions of many robust statistics are very difficult to work out, especially for the kinds of underlying distributions for which the statistics are to be studied. A wellknown use of Monte Carlo is in the important study of robust statistics described by Andrews et al. (1972), who introduced and examined many alternative estimators of location for samples from univariate distributions. This study, which involved many Monte Carlo experiments, employed innovative methods of variance reduction and was very influential in subsequent Monte Carlo studies reported in the statistical literature.

As an example of a Monte Carlo study, we will now describe a simple experiment to assess the robustness of a statistical test in linear regression analysis. The purpose of this example is to illustrate some of the issues in designing a Monte Carlo experiment. The results of this small study are not of interest here. There are many important issues about the robustness of the procedures that we do not address in this example.

3.1 The Problem

Consider the simple linear regression model,

$$Y = \beta_0 + \beta_1 x + E,$$

where a response or "dependent variable", Y, is modeled as a linear function of a single regressor or "independent variable", x, plus a random variable, E, called the "error". Because E is a random variable, Y is also a random variable. The statistical problem is to make inferences about the unknown, constant parameters β_0 and β_1, and about distributional parameters of the random variable, E. The inferences are made based on a sample of n pairs, (y_i, x_i), with which are associated unobservable realizations of the random error, ϵ_i, and which are assumed to have the relationship

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i. \tag{1}$$

We also generally assume that the realizations of the random error are independent and are unrelated to the value of x.

For this example, let us consider just the specific problem of testing the hypothesis

$$\mathrm{H}_0\colon\ \beta_1 = 0, \tag{2}$$

versus the universal alternative. If the distribution of E is normal and we make the additional assumptions above about the sample, the optimal test for the hypothesis (using the common definitions of optimality) is based on a least squares procedure that yields the statistic

$$t = \frac{\widehat{\beta}_1 \sqrt{(n-2)\sum (x_i - \bar{x})^2}}{\sqrt{\sum r_i^2}}, \tag{3}$$

where $\bar{x}$ is the mean of the x's, $\widehat{\beta}_1$ together with $\widehat{\beta}_0$ minimizes the function

$$L_2(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2,$$

and

$$r_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i).$$

If the null hypothesis is true, t is a realization of a Student's t distribution with $n-2$ degrees of freedom. The test is performed by comparing the p-value from the Student's t distribution with a preassigned significance level, α, or by comparing the observed value of t with a critical value. The test of the hypothesis depends on the estimates of β_0 and β_1 used in the test statistic t.

Often, a dataset contains outliers, that is, observations that have a realized error that is very large in absolute value, or observations for which the model is not appropriate. In such cases, the least squares procedure may not perform so well. We can see the effect of some outliers on the least squares estimates of β_0 and β_1 in Figure 1. For well-behaved data, as in the plot on the left, the least squares estimates seem to fit the data fairly well. For data with two outlying points, as in the plot on the right in Figure 1, the least squares estimates are affected so much by the two points in the upper left part of the graph that the estimates do not provide a good fit for the bulk of the data.

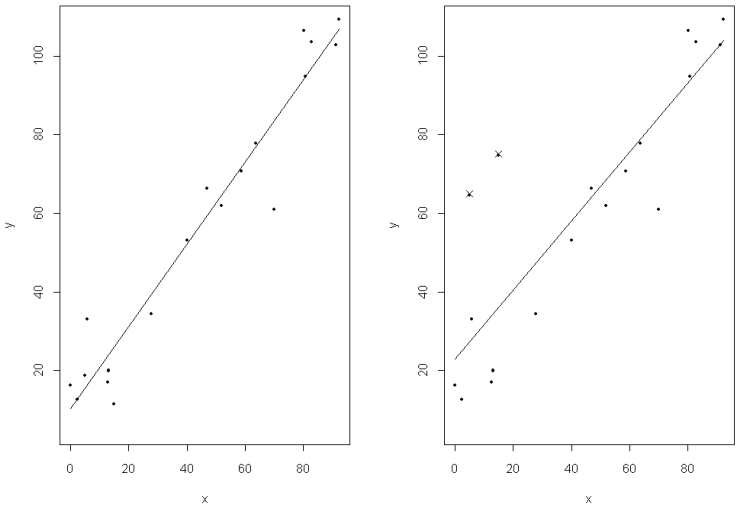

Figure 1: Least Squares Fit Using Two Datasets that are the Same Except for Two Outliers

Another method of fitting the linear regression line that is robust to outliers in E is to minimize the absolute values of the deviations. The least absolute values procedure chooses estimates of β_0 and β_1 so as to minimize the function

$$L_1(b_0, b_1) = \sum_{i=1}^{n} |y_i - b_0 - b_1 x_i|.$$

Figure 2 shows the same two data sets as before with the least squares (LS) fit and the least absolute values (LAV) fit plotted on both graphs. We see the least absolute values fit does not change because of the outliers.

Another concern in regression analysis is the unduly large influence that some individual observations exert on the aggregate statistics because the values of x in those observations lie at a large distance from the mean of all the x_i, that is, those observations whose values of the independent variables are outliers. The influence of an individual observation is called *leverage*. Figure 3 shows two datasets together with the least squares and the least absolute values fits for both. In both datasets, there is one value of x that lies far outside the range of the other values of x. All of the data in the plot on the left in Figure 3 lie relatively close to a line, and both fits are very similar. In the plot on the right, the observation with an extreme value of x also happens to have an outlying value of E. The effect on the least squares fit is marked, while the least absolute values fit is not affected as much. (Despite this example, least absolute values fits are generally not very robust to outliers at high leverage points; especially if there are multiple such outliers. There are other methods of fitting that are

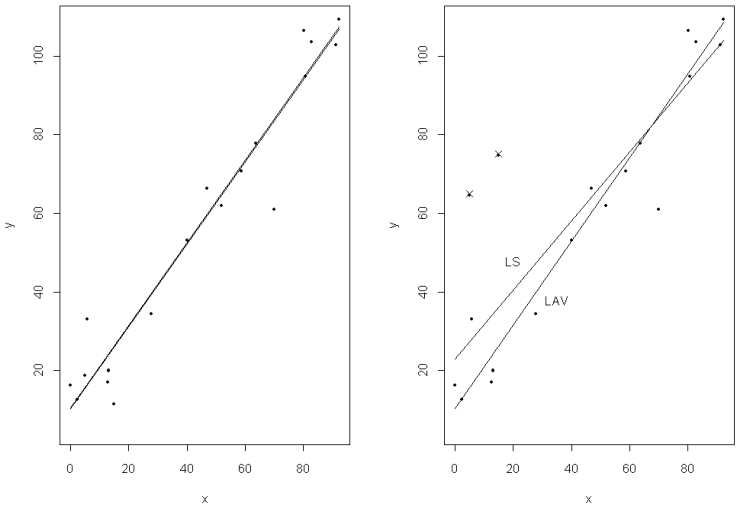

Figure 2: Least Squares Fits and Least Absolute Values Fits

more robust to outliers at high leverage points. We refer the interested reader to Rousseeuw and Leroy, 1987, for discussion of these issues.)

Now we continue with our original objective in this example; that is, to evaluate ways of testing the hypothesis (2).

A test statistic analogous to the one in equation (3), but based on the least absolute values fit, is

$$t_1 = \frac{2\tilde{\beta}_1\sqrt{\sum(x_i - \bar{x})^2}}{(e_{(k_2)} - e_{(k_1)})\sqrt{n-2}}, \tag{4}$$

where $\tilde{\beta}_1$ together with $\tilde{\beta}_0$ minimizes the function

$$L_1(b_0, b_1) = \sum_{i=1}^{n} |y_i - b_0 - b_1 x_i|,$$

$e_{(k)}$ is the k^{th} order statistic from

$$e_i = y_i - (\tilde{\beta}_0 + \tilde{\beta}_1 x_i),$$

k_1 is the integer closest to $(n-1)/2 - \sqrt{n-2}$, and k_2 is the integer closest to $(n-1)/2 + \sqrt{n-2}$. This statistic has an approximate Student's t distribution with $n-2$ degrees of freedom (see Birkes and Dodge, 1993, for example).

If the distribution of the random error is normal, inference based on minimizing the sum of the absolute values is not nearly as efficient as inference based on least squares. This alternative to least squares should therefore be

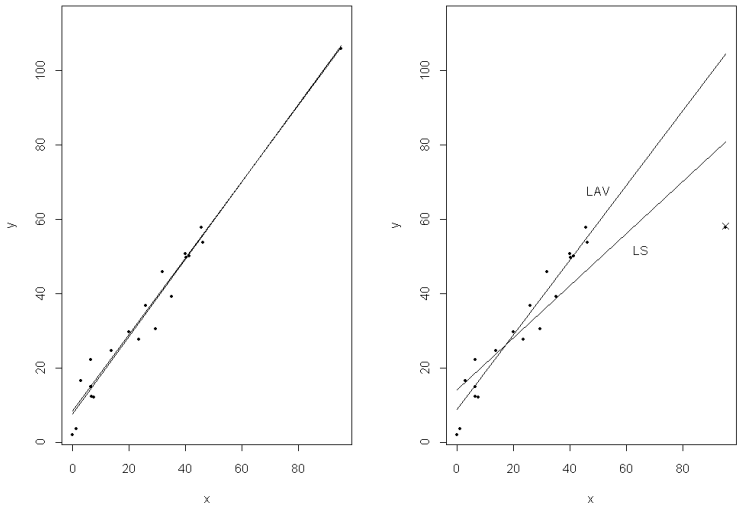

Figure 3: Least Squares and Least Absolute Values Fits

used with some discretion. Furthermore, there are other procedures that may warrant consideration. It is not our purpose here to explore these important issues in robust statistics, however.

3.2 The Design of the Experiment

At this point, we should have a clear picture of the problem: we wish to compare two ways of testing the hypothesis (2) under various scenarios. The data may have outliers, and there may be observations with large leverage. We expect that the optimal test procedure will depend on the presence of outliers, or more generally, on the distribution of the random error, and on the pattern of the values of the independent variable. The possibilities of interest for the distribution of the random error include

- the family of the distribution, that is, normal, double exponential, Cauchy, and so on
- whether the distribution is a mixture of more than one basic distribution, and if so, the proportions in the mixture
- the values of the parameters of the distribution, that is, the variance, the skewness, or any other parameters that may affect the power of the test.

In textbooks on design of experiments, a simple objective of an experiment is to perform a t test or an F test of whether different levels of response are

associated with different treatments. Our objective in the Monte Carlo experiment we are designing is to investigate and characterize the dependence of the performance of the hypothesis test on these factors. The principles of design are similar to those of other experiments, however.

It is possible that the optimal test of the hypothesis will depend on the the sample size or on the true values of the coefficients in the regression model, so some additional issues that are relevant to the performance of a statistical test of this hypothesis are the sample size and the true values of β_0 and β_1.

In the terminology of statistical models, the factors in our Monte Carlo experiment are the estimation method and the associated test, the distribution of the random error, the pattern of the independent variable, the sample size, and the true value of β_0 and β_1. The estimation method together with the associated test is the "treatment" of interest. The "effect" of interest, that is, the measured response, is the proportion of times that the null hypothesis is rejected using the two treatments.

We now can see our objective more clearly: for each setting of the distribution, pattern, and size factors, we wish to measure the power of the two tests. These factors are similar to blocking factors, except that there is likely to be an interaction between the treatment and these factors. Of course, the power depends on the nominal level of the test, α. It may be the case that the nominal level of the test affects the relative powers of the two tests.

We can think of the problem in the context of a binary response model,

$$\mathrm{E}(P_{ijklqsr}) = f(\tau_i, \delta_j, \phi_k, \nu_l, \alpha_q, \beta_{1s}), \tag{5}$$

where the parameters represent levels of the factors listed above (β_{1s} is the s^{th} level of β_1), and $P_{ijklqsr}$ is a binary variable representing whether or not the test rejects the null hypothesis on the r^{th} trial at the $(ijklqs)^{\text{th}}$ setting of the design factors. It is useful to write down a model like this to remind ourselves of the issues in designing an experiment.

At this point it is necessary to pay careful attention to our terminology. We are planning to use a statistical procedure (a Monte Carlo experiment) to evaluate a statistical procedure (a statistical test in a linear model). For the statistical procedure we will use, we have written a model (5) for the observations we will make. Those observations are indexed by r in that model. Let m be the sample size for each combination of factor settings. This is the Monte Carlo sample size. It is not to be confused with the data sample size, n, that is one of the factors in our study.

We now choose the levels of the factors in the Monte Carlo experiment.

- For the estimation method, we have decided on two methods: least squares and least absolute values. Its differential effect in the binary response model (5) is denoted by τ_i, for $i = 1, 2$.

- For the distribution of the random error, we choose three general ones:

    1. Normal $(0, 1)$

2. Normal $(0, 1)$ with $c\%$ outliers from normal $(0, d^2)$
3. Standard Cauchy

We choose different values of c and d as appropriate. For this example, let us choose $c = 5$ and 20, and $d = 2$ and 5. Thus, in the binary response model (5), $j = 1, 2, 3, 4, 5, 6$.

- For the pattern of the independent variable, we choose three different arrangements:

  1. Uniform over the range:
  2. A group of extreme outliers:
  3. Two groups of outliers:

  In the binary response model (5), $k = 1, 2, 3$. We use fixed values of the independent variable.

- For the sample size, we choose three values: 20, 200, and 2,000. In the binary response model (5), $l = 1, 2, 3$.

- For the nominal level of the test, we choose two values: 0.01 and 0.05. In the binary response model (5), $q = 1, 2$.

- The true value of β_0 is probably not relevant, so we just choose $\beta_0 = 1$. We are interested in the power of the tests at different values of β_1. We expect the power function to be symmetric about $\beta_1 = 0$, and to approach 1 as $|\beta_1|$ increases.

The estimation method is the "treatment" of interest.

Restating our objective in terms of the notation introduced above, for each of two tests, we wish to estimate the power curve,

$$\Pr(\text{reject } \mathrm{H}_0) = g(\beta_1 \mid \tau_i, \delta_j, \phi_k, \nu_l, \alpha_q),$$

for any combination $(\tau_i, \delta_j, \phi_k, \nu_l, \alpha_q)$. For either test, this curve should have the general appearance of the curve shown in Figure 4.

The minimum of the power curve should occur at $\beta_1 = 0$, and should be α. The curve should approach 1 symmetrically as $|\beta_1|$.

To estimate the curve we use a discrete set of points; and because of symmetry, all values chosen for β_1 can be nonnegative. The first question is at what point does the curve flatten out just below 1. We might arbitrarily define the region of interest to be that in which the power is less than 0.99, approximately. The abscissa of this point is the maximum β_1 of interest. This point, say β_1^*, varies depending on all of the factors in the study. We could work this out in the least squares case for uncontaminated normal errors, using the noncentral Student's t distribution, but for all other cases, it is analytically intractable. Hence, we compute some preliminary Monte Carlo estimates to determine the maximum β_1 for each factor combination in the study.

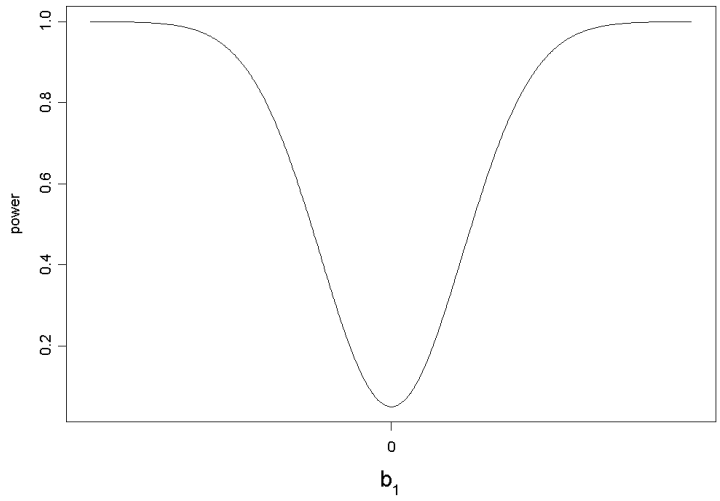

Figure 4: Power Curve for Testing $\beta_1 = 0$

To do a careful job of fitting a curve using a relatively small number of points, we would chose points where the second derivative is changing rapidly, and especially near points of inflection where the second derivative changes sign. Because the problem of determining these points for each combination of (i, j, k, l, q) is not analytically tractable (otherwise we would not be doing the study!), we may conveniently chose a set of points equally spaced between 0 and β_1^*. Let us decide on five such points for this example. It is not important that the β_1^*'s be chosen with a great deat of care. The objective is that we be able to calcuate two power curves between 0 and β_1^* that are meaningful for comparisons.

3.3 The Experiment

The observational units in the experiment are the values of the test statistics (3) and (4). The measurements are the binary variables corresponding to rejection of the hypothesis (2). At each point in the factor space, there will be m such observations. If z is the number of rejections observed, the estimate of the power is z/m, and the variance of the estimator is $\pi(1-\pi)/m$, where π is the true power at that point. (z is a realization of a binomial random variable with parameters m and π.) This leads us to a choice of the value of m. The coefficient of variation at any point is $\sqrt{(1-\pi)/(m\pi)}$, which increases as π decreases. At $\pi = 0.50$, a 5% coefficient of variation can be achieved with a sample of size 400. This yields a standard deviation of 0.025. There may be some motivation to choose a slightly larger value of m, because we can assume

the minimum of π will be approximately the minimum of α. To achieve a 5% coefficient of variation at that point (i.e., at $\beta_1 = 0$) would require a sample of size approximately 160,000. That would correspond to a standard deviation of 0.0005, which is probably much smaller than we need. A sample size of 400 would yield a standard deviation of 0.005. Although that is large in a relative sense, it may be adequate for our purposes. Because this particular point (where $\beta_1 = 0$) corresponds to the null hypothesis however, we may choose a larger sample size, say 4,000, at that special point. A reasonable choice, therefore is a Monte Carlo sample size of 4,000 at the null hypothesis, and 400 at all other points. We will, however, conduct the experiment in such a way that we can combine the results of this experiment with independent results from a subsequent experiment.

The experiment is conducted by running a computer program. The main computation in the program is to determine the values of the test statistics, and to compare them with their critical values so as to decide on the hypothesis. These computations need to be performed at each setting of the factors and for any given realization of the random sample.

We design a program that allows us to loop through the settings of the factors, and at each factor setting, to use a random sample. The result is a nest of loops. The program may be stopped and restarted, so we need to be able to control the seeds.

Recalling that the purpose of our experiment is to obtain estimates, we may now consider any appropriate methods of reducing the variance of those estimates. There is not much opportunity to apply methods of variance reduction, but at least we might consider at what points to use common realizations of the pseudorandom variables. Because the things that we want to compare most directly are the powers of the tests, we perform the tests on the same pseudorandom datasets. Also, because we are interested in the shape of the power curves we may want to use the same pseudorandom datasets at each value of β_1, that is, to use the same set of errors in the model (1). Finally, following similar reasoning, we may use the same pseudorandom datasets at each value of pattern of the independent variable. This implies that our program of nested loops has the structure shown in Figure 5.

After writing a computer program with this structure, the first thing is to test the program on a small set of problems and to determine appropriate values of β_1^*. We should compare the results with known values at a few points. (As mentioned earlier, the only points we can work out correspond to the normal case with the ordinary t statistic. One of these points, at $\beta_1 = 0$, is easily checked.) We can also check the internal consistency of the results. For example, does the power curve increase? We must be careful, of course, in applying such consistency checks, because we do not know the behavior of the tests in most cases.

Initialize a table of counts.
Fix the data sample size. (Loop over the sample sizes $n = 20$, $n = 200$, and $n = 2000$.)
Generate a set of residuals for the linear regression model (1). (This is the loop of m Monte Carlo replications.)
Fix the pattern of the independent variable. (Loop over patterns P_1, P_2, and P_3.)
Choose the distribution of the error term. (Loop over the distributions D_1, D_2, D_3, D_4, D_5, and D_6.)
For each value of β_1, generate a set of observations (the y values) for the linear regression model (1), and perform the tests using both procedures and at both levels of significance. Record results.
End distributions loop.
End patterns loop.
End Monte Carlo loop.
End sample size loop.
Perform computations of summary statistics.

Figure 5: Program Structure for the Monte Carlo Experiment

3.4 Reporting the Results

The report of this Monte Carlo study should address as completely as possible the results of interest. The relative values of the power are the main points of interest. The estimated power at $\beta_1 = 0$ is of interest. This is the actual significance level of the test, and how it compares to the nominal level α is of particular interest.

The presentation should be in a form easily assimulated by the reader. This may mean graphs similar to Figure 4, except only the nonegative half, and with the tick marks on the horizontal axis. Two graphs, for the two test procedures, should be shown on the same set of axes. It is probably counterproductive to show a graph for each factor setting. (There are 108 combinations of factor settings.)

In addition to the graphs, tables may allow presentation of a large amount of information in a compact format.

The Monte Carlo study should be described so carefully that the study could be replicated exactly. This means specifying the factor settings, the loop nesting, the software and computer used, the seed used, and the Monte Carlo sample size. There should also be at least a simple statement explaining the choice the Monte Carlo sample size.

As mentioned earlier, the statistical literature is replete with reports of

Monte Carlo studies. Some of these reports (and, likely, the studies themselves) are woefully deficient. An example of a careful Monte Carlo study and a good report of the study are given by Kleijnen (1977). He designed, performed, and reported a Monte Carlo study to investigate the robustness of a multiple ranking procedure. In addition to reporting on the study of the question at hand, another purpose of the paper was to illustrate the methods of a Monte Carlo study.