Texas A&M University Interlibrary Loan

RAPID:UUS

ILLiad TN: 3159960

Ms

Borrower: RAPID:UUS

9/13/2016 2:06 PM
(Please update within 24 hours)

Lending String:

Call #: HB74~~.5 .T427 2006~~ .8 .T4 1998

Patron:

Location: stk

Journal Title: Teaching economics

Scan
Charge
Maxcost:

Volume: **Issue:**
Month/Year: 1998 **Pages:** 141-159

Article Author: Peter Kennedy

Shipping Address:
NEW: Main Library
Main Library

Article Title: Using Monte Carlo studies for teaching econometrics

Fax:
Ariel:
Email:
Odyssey: 129.123.124.234

Imprint:

ILL Number: -10962894

Note:

CHAPTER 7

USING MONTE CARLO STUDIES FOR TEACHING ECONOMETRICS

Peter Kennedy

What do we want our students to know upon completion of an undergraduate econometrics course? Most instructors would agree that we want them to be able to analyze data with an econometrics software package, and we want them to have some understanding of the theory of econometrics. I argue that what currently is done fails miserably in regard to the second of these goals and propose an alternative approach.

Instructors do not realize that students do not understand the basic logic of statistics as reflected in the concept of a sampling distribution and consequently are inherently uncomfortable with econometrics. Because instructors do not realize this deficiency, they press on with technical material, which students deal with by becoming adept at mathematical manipulation without understanding what is actually going on. At the end of the course, students continue to be uncomfortable with econometrics. This can be remedied if instructors begin by emphasizing the concept of a sampling distribution and thereafter force students to generate explanations of how Monte Carlo studies would be structured to examine the econometric issues dealt with at each stage of the course. Examples are given in Appendix 7.A. (By Monte Carlo study I mean a specific variant of a generic technique known as resampling; to smooth exposition, discussion of resampling and its related literature is presented in Appendix 7.B.)

I. WHAT'S WRONG

Upon completion of introductory statistics courses, the vast majority of students do not understand the basic logic of classical statistics as captured in the concept of repeated samples and a sampling distribution. They know how to do mechanical things such as compute a sample variance, run a regression, and test an hypothesis, but they do not have a feel for the "big picture." They have learned a bunch of techniques, but to them they are just that, a bunch of techniques, and they know they can pass the course by remembering how these techniques work. They view statistics as a branch of mathematics because it uses mathematical formulas, so they look at statistics through a mathematical lens. What they are missing is the statistical lens through which to view this world, allowing this world to make sense. The concept of a sampling distribution is this statistical lens. My own experience discovering this lens was a revelation, akin to the experience I had when I put on my first pair of eyeglasses - suddenly everything was sharp and clear.

Those of us who teach statistics should be aware of this phenomenon. We frequently encounter students with A grades in their introductory statistics courses who clearly have no understanding of statistics beyond an ability to mechanically apply standard statistical procedures. We often see graduate students with an impressive ability to derive statistical formulas but a remarkable inability to explain how to evaluate those formulas via a Monte Carlo study. Occasionally technically competent students admit that although they can do the mathematics, they are not really comfortable with statistics. These shortcomings of students may be a result of the way we teach this material.

How has this state of affairs come about? It is not because the introductory statistics books ignore sampling distributions - they all have plenty of good material on this concept and give it appropriate emphasis. And it is not because instructors ignore this dimension of the introductory textbooks - all instructors will swear that they teach this concept thoroughly. But students are utility maximizers. They know that their exams are going to require an ability to interpret regressions, calculate t statistics, and perform hypothesis tests, not an ability to explain the concept of a sampling distribution. They can not imagine, and instructors and textbooks do not provide, examples of exam questions that in any meaningful way probe student understanding of this concept.

As a result, most students, even good students, go on from introductory statistics to econometrics without an adequate understanding of the fundamentals of statistics. The better instructors realize that students have mostly forgotten their introductory statistics material, or never learned it properly in the first place, and proceed on the basis that this introductory material needs to be reviewed. Unfortunately, this review suffers from the same problem noted earlier - there is little motivation for students to buckle down and properly learn the concept of a sampling distribution and the role it plays in econometrics because instructors' expositions of this concept, however clear, are not accompanied by appropriate example exam questions. Such questions provide motivation and, more importantly, force students to work out answers. Although we lecturers like to think otherwise, brilliant expositions seldom cause students fully to understand – such understanding comes through working out problems based on the concept to be learned. An old adage bears repeating: I hear and I learn; I see and I realize; I do and I understand.

Econometrics textbooks do not help much. They are in too big a hurry to produce the theorems, proofs, and formulas that define theoretical econometrics. Some review introductory statistics but do so without much emphasis on sampling distributions and seldom provide meaningful questions that force student understanding of the sampling distribution concept. Two things are needed. First, instructors need to produce for students a good exposition of what is a sampling distribution and what role it plays in econometrics. Second, a means must be found to construct problems that cause students to learn this concept properly. In the next section, I offer my own exposition of sampling distributions and suggest that the needed problems can be constructed by asking students to explain how to conduct Monte Carlo studies.

II. SAMPLING DISTRIBUTIONS: AN EXPOSITION

In my experience, students fail to grasp the meaning and import of a sampling distribution if it is described in general terms. A successful exposition must be couched in terms of a specific example such as the following.

Suppose that you have 45 observations on variables x and y. You know that y has been produced from x by using the formula $y = \beta x + \varepsilon$ where ε is an error with mean zero and variance σ^2. Note that there is no intercept in this linear relationship, so there is only one unknown parameter, β, the slope of x. This means that y_1, the first observation on y, was created by multiplying x_1, the first observation on x, by the unknown number β and then adding ε_1, the first error, obtained by randomly grabbing an error from a bowl of errors with mean zero and variance σ^2. The other 44 observations on y were created in similar fashion.

You are interested in estimating the unknown parameter β. Professor A suggests using the formula $\beta^* = \Sigma y/\Sigma x$ and professor B suggests using the formula $\beta^{**} = \Sigma xy/\Sigma x^2$, where the subscripts have been omitted for convenience. Both of these suggestions just involve putting the data into an algebraic formula to produce a single number that is proposed as the estimate of β. Suppose this is done using your data, producing $\beta^* = 2.34$ and $\beta^{**} = 2.73$. Which of these two estimates should you choose?

Take a closer look at these two formulas to see if a rationale can be produced for choosing one in preference to the other. Since $y = \beta x + \varepsilon$ you can substitute this into professor A's formula to get

$$\beta^* = \Sigma\,(\beta x + \varepsilon)/\Sigma x = \beta + \Sigma\varepsilon/\Sigma x$$

and you can substitute it into professor B's formula to get

$$\beta^{**} = \Sigma\,(\beta x + \varepsilon)/\Sigma x^2 = \beta + \Sigma\varepsilon/\Sigma x^2.$$

From this it is apparent that both these formulas are such that they are equal to β plus an expression that involves the unknown ε values. Because the ε values are positive about half the time and negative about half the time, it looks as though in both cases this expression involving the errors is probably going to be fairly close to zero, suggesting that both of these formulas are reasonable – they each appear to be creating a suitable estimate of β.

Consider professor A's estimate $\beta^* = \beta + \Sigma\varepsilon/\Sigma x = 2.34$. How close $\beta^* = 2.34$ is to β clearly depends on the particular set of 45 error terms that were drawn out of the bowl of errors to produce the y observations. If in obtaining the data, mostly large positive errors were drawn, β^* would substantially overestimate β. If mostly large negative errors were drawn, β^* would substantially underestimate β. If a more typical set of errors were drawn, β^* would produce an estimate fairly close to β. The point here is that the set of 45 unknown error terms drawn determines the estimate produced by the β^* formula and so there is no way of knowing how close the particular estimate 2.34 is to the true β value.

Now suppose for a moment that you know the true value of β. Visualize grabbing 45 error terms, calculating y using β and the x values, computing β^*, and recording the result. Mentally do this 2,000 times, each time grabbing 45 new error terms. This produces 2,000 hypothetical β^* values, each produced by mimicking exactly the process thought to be generating the actual y data. These 2,000 β^* values can be used to construct a histogram of possible β^* values. This histogram should show that very high values of β^* will be relatively rare because such a β^* value would require an unusual draw of 45 error terms. Similarly, very low values of β^* will also be rare. Values of β^* closer to the true β value will be more common because they would result from more typical draws of 45 errors. This histogram estimates a distribution of possible β^* values, providing a picture of the relative probabilities of obtaining different β^* values. This distribution is called the sampling distribution of β^*. Professor B's formula β^{**} will also have a sampling distribution, but because it is a different formula its sampling distribution should look slightly different. (I exposit the algebra above for the two estimators side-by-side on the overhead/blackboard, finishing with a sampling distribution picture for each, both drawn as normal-looking distributions centered over the true value of β but with the variance of β^* noticeably larger than the variance of β^{**}.)

The logic of all this delivers the following punch line: **Using β^* to produce an estimate of β can be conceptualized as the econometrician shutting his or her**

eyes and obtaining an estimate of β by reaching blindly into the sampling distribution of β^* to obtain a single number.

An immediate implication of this is that our choice between β^* and β^{**} comes down to the following: **Would I prefer to produce my estimate of β by reaching blindly into the sampling distribution of β^* or by reaching blindly into the sampling distribution of β^{**}?**

At this stage, the instructor can confess that most of econometricians' fancy algebraic derivations are just efforts to investigate what the sampling distributions of suggested estimators or test statistics look like and that much of the essence of econometrics can be summarized by addressing four basic questions related to the sampling distribution concept:

1. What would make one sampling distribution better than another in this regard? This motivates a discussion of bias, variance, and mean square error.

2. How does one find out what a sampling distribution looks like? This underlines the importance of knowing how to calculate expected values and variances, either algebraically, perhaps relying on asymptotics, or via Monte Carlo procedures.

3. What is a Monte Carlo study? This permits exposition of Monte Carlo studies as using the computer to perform the mental experiment of generating thousands of β^* values for cases in which the algebra is "too difficult."

4. Do sampling distributions play any other role in econometrics? This allows hypothesis testing to be exposited in terms of the sampling distributions of test statistics.

III. WHAT DO THESE SAMPLING DISTRIBUTIONS LOOK LIKE?

It is useful to pursue this example to illustrate these basic questions associated with the sampling distribution concept. The example can be exploited/extended in several ways.

1. Begin by explaining that an estimator is unbiased if its sampling distribution is centered over the true value of the parameter being estimated. Note that as we will see later, these two estimators are unbiased so their sampling distributions are both drawn (as I suggested earlier) centered over β. Now ask students which of these two sampling distributions they would prefer to shut their eyes and draw from to produce their estimate of β. They should quickly agree that they would prefer the sampling distribution of β^{**}, allowing you to announce the minimum variance (efficiency) criterion.

2. Next, illustrate the possibility of a biased estimator. Do this by changing the data-generating process so that there is now a nonzero intercept: $y = \alpha + \beta x + \varepsilon$. The algebra on β^{**} now produces

$$\beta^{**} = \alpha \sum x / \sum x^2 + \beta + \sum x\varepsilon / \sum x^2$$

so that it appears as though β^{**} will be biased by an amount $\alpha\Sigma x/\Sigma x^2$. An important point to stress here is that the sampling distribution properties of an estimator depend crucially on the data-generating process – an estimator with an attractive sampling distribution for one data-generating process may have a very unattractive sampling distribution for another data-generating process.

3. Econometricians work hard to find ways of modifying popular estimators so that they will have attractive sampling distributions in new data-generating situations. In this case consider the estimator

$$\beta^{***} = \sum(x-\bar{x})(y-\bar{y}) / \sum(x-\bar{x})^2 = \beta + \sum(x-\bar{x})(\varepsilon-\bar{\varepsilon}) / \sum(x-\bar{x})^2$$

which removes the bias. If we were to use bias as our criterion of choice, we would choose β^{***} in preference to β^{**}.

4. Next, draw the sampling distributions of β^{**} and of β^{***} for this case of a nonzero intercept. Draw the sampling distribution of β^{***} as centered over β and the sampling distribution of β^{**} as biased slightly but with a much smaller variance than the variance of β^{***}. At this point, just state that it turns out that β^{**} has a smaller variance, a result you will explore later. Now if you have drawn your diagram judiciously, the students should quickly agree that despite β^{**} being biased they would prefer to obtain their estimate of β by reaching blindly into the sampling distribution of β^{**}, rather than the sampling distribution of β^{***}. Conclude that maybe unbiasedness is not as important as we thought.

5. Now redraw the sampling distribution of β^{**} so that its variance is just a bit smaller than that of β^{***}, so that it is not clear which of β^{**} and β^{***} is preferred. Note that what is needed is some formal way of trading off bias and variance. Introduce mean square error, MSE, as a tradeoff taking the form MSE = bias^2 + variance, a result equivalent to minimizing the expected square of the difference between the estimate and what it is estimating. You might wish to show this algebraically at this point, if that is your style.

6. Now that you have introduced the three main criteria defining an attractive sampling distribution for an estimator, move on to how in simple cases such as this example we can use algebra to deduce these properties. (I will continue the example using β^{**} and β^{***}, but you may wish to return to comparing β^{*} and

β^{**}.) You can show formally that β^{**} is biased and β^{***} unbiased. Then you can derive

$$\mathrm{Var}(\beta^{**}) = \sigma^2/\sum x^2 \text{ and } \mathrm{Var}(\beta^{***}) = \sigma^2/\sum(x-\bar{x})^2$$

which is clearly bigger. Then you can note that β^{***} will have a smaller MSE than

$$\beta^{**} \text{ if } \left[\alpha\sum x/\sum x^2\right]^2 + \sigma^2/\sum x^2 \text{ is greater than } \sigma^2/\sum(x-\bar{x})^2.$$

This is an example of a major problem associated with using MSE as a criterion: Which estimator has the smaller MSE depends on the unknown true values of the parameters, in this case α and σ^2.

7. By now, students should be receptive to a means of finding the characteristics of the sampling distribution without having to do algebra. Announce that Monte Carlo studies are designed to do just that, by using a computer to simulate the repeated draws of error terms that affect the estimates produced by competing formulas estimating β. Show explicitly how a Monte Carlo study would be performed in the context of this example:

a. Choose 45 values for x, either by using the x values from an actual empirical study or by using the computer to generate 45 values in some way.
b. Choose values for the unknown parameters, say $\alpha = 1$, $\beta = 2$ and $\sigma^2 = 4$.
c. Get the computer to draw 45 error terms (ε) randomly from a distribution with mean zero and variance 4.
d. Calculate 45 y values as $1 + 2x + \varepsilon$.
e. Calculate β^{**} and β^{***} and save them.
f. Return to step c and repeat this procedure until you have, say, 2,000 β^{**} and β^{***} values.
g. Use the 2,000 β^{**} values and 2,000 β^{***} values to draw a histogram showing the features of the sampling distributions of β^{**} and β^{***}. Then announce that this is seldom done because usually the sampling distribution is characterized by only three measures, its mean, its variance, and its MSE. The mean is estimated by the average of the 2,000 estimates, the variance is estimated by the sample variance of the 2,000 estimates, and the MSE is estimated by subtracting the true value of β, in this case 2, from each estimate, squaring these numbers and then averaging these 2,000 squared numbers.

IV. EXPLAINING MONTE CARLO STUDIES: DO'S AND DON'TS

A careful exposition of sampling distributions and a clear discussion of their connection to econometricians' focus on deriving expected values, variances, and test statistics should be a welcome addition to the beginning of an econometrics course, but it is only that – a beginning. Some way must be found to hammer home this lesson so that the perspective it provides for later study is not forgotten. I suggest the use of explain-how-to-do-a-Monte-Carlo-study problems. By forcing students to structure a Monte Carlo study to examine an econometric question, they are forced to spell out very clearly their understanding of this question. Some advice in this regard follows, based on many years of experience teaching using this approach.

- Do not ask students actually to do a Monte Carlo study. Although students are supposed to be computer literate, they are not able to quickly learn how to program. Be content with having them write down a detailed step-by-step set of instructions for a hypothetical research assistant.
- Do prepare for the question: Won't our choice of parameter values affect the results of the Monte Carlo study? There are three possible answers. First, the kinds of problems addressed by undergraduates are such that the results about variances and about test statistics usually do not depend on the true parameter values. Admit that if bias exists its magnitude is affected by the true parameter values and so in these situations MSE also is affected. Second, if there is reason to believe that the results are affected, then the study should be repeated for a variety of parameter values to see in what way the results are affected. Third, for a particular study, we can choose the true parameter values equal to the parameter estimates from the data at hand, so that the results of the Monte Carlo study are specific to parameter values that are likely close to the true parameter values for that study.
- Do make students describe a Monte Carlo study for every major topic in the course. The detail required in the expected answers can diminish markedly over time because certain things (such as formulas for estimating bias, variance, and mean square error) become unquestionably known by all, so such tasks should not be time-consuming. As noted later, they could be no more than exercises designed to ensure that the students understand how the data-generating process operates.
- Do evaluate students' instructions carefully. Small mistakes can reflect major misunderstandings. Be particularly on guard for vagueness in these instructions; this is students' favorite way of attempting to conceal lack of understanding. For example, they may not specify their parameter values because they are not sure if they should use values satisfying the null hypothesis, or they may omit describing how to get the first error needed to generate a sequence of first-order autocorrelated errors. They may not know how to do it or whether it is important.
- Do spend lots of time on this. Most students have a remarkably difficult time grasping the logic of what is going on. At first their attention is monopolized by the mechanical steps of a Monte Carlo study, at the expense of

understanding the overall logic of what the particular study they are explaining is trying to accomplish. Consequently the first Monte Carlo problems should be very easy in terms of the econometric question they address because students will be preoccupied with questions such as: What is the computer capable of doing? How should fixed-in-repeated-samples observations be selected? How do we decide what parameter values to use? What sample size should be used? What variance of the error should be chosen? How can an error with this variance be created? What formulas should be used to estimate the mean and variance of the sampling distribution? Exactly which steps need to be repeated? How many repeated samples should be taken?

- Do reverse the problem – provide students with an outline of a Monte Carlo study and ask them a question about what results would be expected, as illustrated in the first three examples in Appendix 7.A. This approach makes students think about relevant concepts as it provides examples of the set of instructions students should later be able to produce on their own. The next stage is to ask them to describe how to conduct a study to address a specific issue in the context of a specific problem. It is important that the question spell out many of the details of the study, such as is done in examples E, F, G and I in Appendix 7.A. (The problems gathered in example H of Appendix 7.A illustrate a variety of issues that can form the basis of a Monte Carlo study problem but as written are unsuitable for students because most students require a very specific context.)
- Do ask students how to simulate the data-generating process. The fact that the properties of an estimator's or a test statistic's sampling distribution depend on the process generating the data is not quickly appreciated. If you ask students to explain how to generate data to perform a Monte Carlo study for every major econometric topic you cover in the course, you will be unpleasantly surprised by the problems they have, but they will benefit immensely by understanding the role of the data-generating mechanism. If they cannot explain how to generate raw data for a heteroskedastic error model, for a two-equation simultaneous equation system, or for a qualitative choice model, can you be confident that they understand what any of the equations you have been putting on the blackboard really mean?
- Do call on a student to suggest the first step, then go around the room asking each student in turn to supply the next step, or correct an incorrect step suggested by an earlier student. In-class development of a step-by-step set of Monte Carlo instructions can be an effective means of generating classroom participation.
- Do consider introducing students to bootstrapping as an advanced topic, after traditional Monte Carlo studies have been mastered. (Bootstrapping is a special type of Monte Carlo study that is becoming quite common in econometrics.) Further comment on this is provided in Appendix 7.B.

- Do allow better students who wish to perform a Monte Carlo study the opportunity, perhaps as a term paper. The popular econometrics packages have programming capabilities. Be cautious if using a statistical package designed to perform resampling experiments, such as Resampling Stats. Although these packages are an extremely useful means of showing students the logic of repeated samples, they mix a variety of resampling techniques without drawing students' attention to the fact that some of these different techniques do not correspond to the rationale that underlies econometric analysis. For example, for some problems, they employ Monte Carlo, for others, they use bootstrapping, and for still others they exploit a randomization methodology. Furthermore, they do not provide packaged econometric techniques, so investigation of the properties of econometric methods is awkward. See Appendix 7.B for further discussion.

V. OPPORTUNITY COST

As should be evident from the discussion above, my proposal requires that instructors spend considerable time on sampling distributions and Monte Carlo studies, something they do not now do. What should be given up to accommodate this new feature?

This is a matter for each individual instructor to decide. I believe that the benefit of my innovation is so large that no instructor could possibly claim that the least-valued subset of their course has more benefit. Indeed, my innovation should enhance student understanding of all dimensions of a course, creating benefits beyond those attached solely to learning about sampling distributions and Monte Carlo studies. Advanced estimation or testing techniques do not mean much to students if they do not understand the fundamental principles that lie behind them; such advanced material is quickly forgotten once the course is completed.

In my own courses, I have given up most mathematical derivations. Do we really want our students at the end of their course to be able to do technical things like derive the OLS estimator and prove that it is BLUE, show that the expected value of s^2 is σ^2, or show mathematically that MSE is the sum of bias squared and variance? I think not. These things in any event have little meaning if the concept of a sampling distribution is not thoroughly understood.

VI. CONCLUSION

By the end of my econometrics course, I want my students to understand the basic principles of classical statistics well enough to allow them to sit down with a textbook and figure out the rationale of any new statistical procedure that they are likely to encounter in their postundergraduate lives. In keeping with this, all my exams are open-book. Is my approach easier for the student? In my experience, the answer to this is yes and no. For those who quickly learn the Monte Carlo view of the statistical world, the entire econometrics course becomes easier because these students understand what is going on. But for all students this approach is intellectually very challenging because it forces them to think hard to understand an abstract concept. Simon and Bruce (1991, 29) make the same point in the context of their resampling approach to learning basic statistics, claiming that it "requires only hard, clear thinking. You cannot beg off by saying 'I have no brain for math!'" Some students do not rise to this challenge; they are more comfortable learning by rote a bunch of techniques and mathematical proofs and so find my approach more difficult.

I like to think that my approach works well for those who obtain jobs in which statistical analysis plays some role, but unfortunately such students tend to disappear permanently from one's life. Feedback from students who have gone on to graduate studies has been illuminating, however. One such student confessed that all he really needed to know about econometrics he learned in my course. Another told me that during the first half of his graduate econometrics course he was far behind the rest of the class because he had to learn from scratch a lot of formulas and algebraic derivations that other students had encountered in their more-traditional undergraduate econometrics courses. But he found that after the first month or so he had become comfortable with the math and suddenly leapfrogged far ahead of the rest of the class because, he said, he understood what was going on but they understood only the mathematics. Can an instructor ask for a better endorsement?

Appendix 7.A: Monte Carlo Problem Examples

This appendix provides some examples of how questions based on Monte Carlo studies can be structured. Additional examples can be found in Kennedy (1998). Examples A, B, and C reverse the sequence of the usual problem. Rather than asking students to describe a Monte Carlo study, these questions provide descriptions of Monte Carlo studies and then ask students to explain what results are expected. This makes them suitable as introductory Monte Carlo problems because they ease students into this unfamiliar territory by providing illustrations of how Monte Carlo studies are structured. Example D is similar in that it provides the results of a Monte Carlo study and asks students to work with them. Examples E, F and G illustrate typical Monte Carlo problems, in which the question provides considerable detail regarding the setup of the study. Example H suggests a variety of topics that can be investigated by Monte Carlo studies, but offers the student no specific context. I recommend that before being given to students these examples be converted into problems with specific context, such as is illustrated in example I. Some examples of typical mistakes students make when structuring Monte Carlo studies are given after the examples.

A. Suppose you have programmed a computer to do the following:

1. Draw randomly 25 values from a standard normal distribution.
2. Multiply each of these values by 3 and add 2.
3. Take their average and call it A1.
4. Repeat this procedure to obtain 500 averages A1 through A500.
5. Compute the average of these 500 A values. Call it Abar.
6. Compute the variance of these 500 A values. Call it Avar.
   a. What number do you think Abar should be close to? Explain your logic.
   b. What number do you think Avar should be close to? Explain your logic.

B. Suppose you have programmed a computer to do the following:

1. Draw randomly a hundred values from a standard normal distribution.
2. Multiply each of these values by 5 and add 1.
3. Average the resulting hundred values.
4. Call the average A1 and save it.
5. Repeat the procedure above to produce 2,000 averages A1 through A2,000.
6. Order these 2,000 values from the smallest to the largest.
   a. What is your best guess of the 1,900th ordered value? Explain your logic.
   b. What is your best guess of how many of these values are negative? Explain your logic.

C. Suppose you have programmed a computer to do the following:

1. Draw 20 x values from a distribution uniform between 2 and 8.
2. Draw 20 z values from a normal distribution with mean 12 and variance 2.
3. Draw 20 ε values from a standard normal distribution.
4. Create 20 y values using the formula $y = 2 + 3x + 4z + 5\varepsilon$.
5. Regress y on x and z, obtaining the estimate b_z of the coefficient of z and the estimate seb_z of its standard error.
6. Subtract 4 from b_z, divide this by seb_z and call it $w1$.
7. Repeat the process described above beginning with step 3 until 5,000 w values have been created, $w1$ through $w5,000$.
8. Order the 5,000 w values from smallest to largest.
9. Average the 4,750th and 4,751st values.

   What is your best guess of this average? Explain your reasoning.

D. Suppose you have conducted a Monte Carlo study to investigate, for sample size 25, the bias of an estimator β^* of the slope coefficient in the relationship $y = 2 + 3x + \varepsilon$ where you drew 400 repeated samples of errors (ε) from a normal distribution with mean zero and variance 9.0. Your study estimates the bias of β^* as 0.04 and the variance of β^* as .01. You are not sure if 0.04 is small enough to be considered zero.
Test the null hypothesis that it is insignificantly different from zero.

E. Consider the case of a sample of size 20 of a variable x distributed normally with mean 8 and variance 4. The usual estimator of the variance of x has a divisor of $N-1$. Two alternative estimators use divisors of N and $N+1$.
Explain in detail how to conduct a Monte Carlo study to investigate the relative merits of these three formulas.

F. Suppose the classical linear regression model applies to the money demand function $m = \alpha + \beta y + \delta r + \varepsilon$ and you have 25 observations on income y and on the nominal interest rate r which in your data are positively correlated. You wish to compare the ordinary least squares β estimates including, versus omitting, the relevant explanatory variable r.

1. Explain in detail how to do this with a Monte Carlo study.
2. What results do you expect to get? Why?

G. Suppose the classical normal linear regression model applies to the money demand function $m = \alpha + \beta y + \delta r + \varepsilon$ and you have 25 observations on income y and on the nominal interest rate r which in your data are negatively correlated. You regress m on y (erroneously omitting r) and use a t test to test the null hypothesis that $\beta = 1$ at the $\alpha = .05$ significance level.

1. Explain in detail how to conduct a Monte Carlo study to find the type I error of this t test.
2. What results do you expect to get? Why?

H. Explain in detail how you would undertake a Monte Carlo study to examine:

1. the validity of the textbook formula for the variance of the sample mean statistic;
2. the payoff to incorporating extraneous information in an OLS regression;
3. how inference using OLS is biased when the errors are heteroskedastic;
4. the relative merits of OLS and estimated generalized least squares when the errors are first-order autocorrelated;
5. how OLS is biased in the presence of measurement errors;

6. the relative merits of linear and logit regression in a qualitative choice model;
7. the risk function of a pretest estimator; and
8. the power curve of a Lagrange multiplier test versus a Wald test or a likelihood ratio test.

I. Explain how to undertake a Monte Carlo study to examine the relative merits of ordinary least squares and two-stage least squares in the simultaneous equation system

$$\text{D: } Q = \alpha_0 + \alpha_1 P + \alpha_2 Y + \alpha_3 A + \varepsilon_d$$
$$\text{S: } Q = \beta_0 + \beta_1 P + \varepsilon_s$$

where Q and P are the endogenous variables quantity and price, and Y and A are the exogenous variables income and advertising.

Typical Student Errors

There are several generic errors that students tend to make when describing Monte Carlo studies for Questions D, E, F, G and I above.

Question D

Bias is estimated using the average of the 400 β^* values, so the variance of the bias estimate should be calculated using the formula for the variance of the sample mean statistic, in this case estimated as 0.01/400. Many students just use 0.01.

Question E

1. The results here are sensitive to sample size, with the bias of the formulas dividing by N and by $N+1$ disappearing as the sample becomes larger. Students should note that their choice of sample size is important here but many do not.
2. Some students become confused by the fact that this question is looking at the bias and variance of estimators of variance and are careless in the way in which they word their exposition.
3. Looking at only bias and variance misses the fact that dividing by $N+1$ in this example happens to produce the smallest MSE; sometimes students neglect to look at MSE.
4. Students have a tendency to estimate MSE by summing their estimate of variance and the square of their estimate of bias. This is not quite as good as estimating MSE directly – by finding the average of the squared difference between the parameter estimate and the true parameter value.

Question F

1. Some students use two data-generating mechanisms and so create two sets of data, one with a nonzero coefficient on the interest rate and then another, inappropriate set with a zero coefficient on the interest rate.
2. Some students do the Monte Carlo study for one estimator separately, then repeat it for the other, drawing new error terms. Although this is not incorrect, it does not exploit the efficiency to be gained from comparing the two estimators using the same errors.

Question G

1. Because the Monte Carlo study is to find the type I error of a test, the null hypothesis, that $\beta = 1$, must be true in the data. Many students neglect to ensure that this is the case by using $\beta = 1$ when generating data.
2. Some students think that the researcher failing to include r as a regressor means that the data-generating process has a zero coefficient on r and so they erroneously generate data with the coefficient on r set equal to zero.

Question I

1. A common mistake is failure to note that only the supply equation is identified so that the Monte Carlo study must be confined to investigation of the estimation of the supply equation's parameters.
2. A second common error is not to generate the P and Q values such that they have been generated as the equilibrium values of the simultaneous equations. To do this students need to employ the reduced form expressions, including the reduced error expressions, when generating the P and Q values.

Appendix 7.B: Alternative Forms of Resampling

Throughout this chapter I have argued for the use of descriptions of Monte Carlo studies to enhance teaching of econometrics. The purpose of this appendix is to discuss how this suggestion fits into the related literature on resampling. As defined by Simon and Bruce (1991, 28) resampling is "Using the entire set of data you have in hand, or using the given data-generating mechanism (such as a die) that is a model of the process you wish to understand, to produce new samples of simulated data, and to examine the results of those samples." They claim (and cite confirming empirical evidence) that by teaching students in terms of resampling rather than in terms of mathematical formulas, better understanding of statistical concepts and a higher success rate in finding numerical answers are achieved. Boomsma and Molenaar (1991) and Albert and Berliner (1994) are interesting critiques.

Before discussing how my Monte Carlo suggestion fits into this literature it is necessary to describe the three fundamentally-different ways of conducting resampling experiments to find sampling distributions. I exposit them below in the context of an ordinary least squares (OLS) regression of y on x, sample size N.

Monte Carlo

Sampling distributions of statistics of interest are calculated by drawing errors from a known distribution, using them in conjunction with a known data-generating mechanism to create new data, and recalculating the statistics. Its major drawback is that any conclusions it reaches may be sensitive to the error distribution assumed. The essence of a Monte Carlo study should be clear from earlier examples, so I will not review the procedure here. What will be of relevance is how it differs from the two alternative procedures described below.

Bootstrapping

Bootstrapping is a special case of Monte Carlo, with two main variants. The first variant begins by calculating the OLS estimates and their associated N residuals. Then a Monte Carlo study is conducted using the OLS estimates as the true parameter values and using the residuals as the pool of errors from which one draws (with replacement) errors when constructing y values. In effect, the error distribution is chosen as discrete, with equal probability on each of the N residual values. The big advantage of the bootstrap is that it does not rely on a questionable assumption regarding the distribution of the true errors, allowing the distribution of the residuals to serve as a proxy for this unknown distribution. This variant of bootstrapping is squarely in the tradition of Monte Carlo studies; it does exactly what a Monte Carlo study does but changes the nature of the bowl of errors used in constructing the repeated sample y values. The big disadvantage of this method is that it must be the case that each of these residuals is equally likely to be attached to each observation, something that may not be true. An example would be a case of heteroskedasticity when large errors are more likely to be attached to observations having true errors with larger variances.

To deal with problems such as this heteroskedasticity, a second variant of the bootstrap is employed. In this approach, each value of the dependent variable along with its associated independent variable values is placed in a vector to form a set of N such vectors. A new sample of size N is created by drawing with replacement from this set of N vectors and the OLS estimator is calculated. By repeating this process several times, a sampling distribution can be calculated. The big advantage of this approach is that it circumvents the problem of all errors not being equally likely to be attached to each observation – in these repeated samples each error stays with its associated dependent and independent variable values.

This variant of the bootstrap has two big disadvantages. First, no known true parameter value is associated with this repeated sampling process so that although the variance of the estimator in repeated samples can be calculated, its bias cannot. Second, this procedure changes the meaning of the sampling distribution. Earlier the sampling distribution reflected how the drawing of error terms affects our estimates, given the values of the explanatory variables drawn when we obtained our sample. This second bootstrap procedure produces a sampling distribution that must be interpreted differently. It reflects how our estimates are affected by the drawing of both the values of the explanatory variables and their associated error terms, assuming the values of the explanatory variables drawn when we drew our sample are typical of explanatory variable values that could have been obtained. The difference between these two types of sampling distributions can be of consequence. Suppose, for example, we are interested in examining the determinants of public expenditure in cities of population greater than 20,000 in the state of Washington. The city of Seattle is very different from all other Washington cities. It would not be possible for more than one city with Seattle's characteristics to appear in the sample, but exactly that could happen in this second bootstrap procedure, producing a misleading sampling distribution.

Bootstrapping is becoming very popular in applied econometric work. Jeong and Maddala (1993) provide a good survey for econometricians.

Randomization Tests

A third resampling method, found in the context of hypothesis testing, consists of shuffling the existing data in some suitable fashion, calculating a test statistic and repeating this procedure to build up an estimate of the sampling distribution of this test statistic. For example, suppose the null hypothesis is that the slope of x is zero. If the slope of x is zero, then it should make no difference which x value is paired with y when running a regression. By shuffling the x values several times and rerunning the regression to produce several t statistics, one can estimate the sampling distribution of the t statistic and see if the t statistic produced by the actual ordering of the data is unusual. In this case, the meaning attached to the sampling distribution is different yet again. It tells us how the t statistic differs whenever the x data are shuffled.

Randomization testing does not appear much in econometrics; for discussion of its conceptual difference from traditional econometric thinking, and analysis of its use in econometrics, see Kennedy (1995) and Kennedy and Cade (1996).

Why Choose Monte Carlo?

There are several major differences between Monte Carlo and the other two resampling methods.

1. Bootstrapping and randomization are designed to produce information specifically tailored to the sample at hand, whereas Monte Carlo produces information regarding an assumed-known data-generation mechanism.
2. Because of point 1 above, bootstrapping and randomization can be used as techniques to undertake hypothesis testing for a given sample, whereas Monte Carlo can only be used to describe the character of a statistical procedure if that sample had been generated by a known data-generating mechanism.

3. Monte Carlo encompasses bootstrapping and randomization in the sense that if we wanted to know the character of a bootstrapping or randomization procedure in a known setting we would investigate via a Monte Carlo study.
4. The meaning of a sampling distribution is not necessarily the same in these three types of resampling methods.
5. There are some situations, such as the heteroskedasticity case for Washington state discussed earlier, in which neither bootstrapping nor randomization can be employed.
6. In many situations it is quite difficult to deduce how to perform a bootstrap or a randomization test, but in all cases a Monte Carlo procedure is straightforward.

To some extent these differences speak for themselves. My choice of Monte Carlo exercises to ensure that students learn the resampling concept was not intended to denigrate bootstrapping or randomization - they clearly have special advantages (and disadvantages). Monte Carlo studies are used in econometrics mainly to investigate the properties of an estimator or test in the context of a specific data-generating mechanism. Because much of textbook econometric theory is concerned with just this issue, Monte Carlo studies are of particular value in teaching econometric theory. Bootstrapping and randomization are better viewed as inference techniques and introduced to students as such.

My suggestion that students be asked to describe how to perform a Monte Carlo study to investigate an econometric issue should be interpreted as a means of enhancing student understanding of the resampling concept and the rationale behind the use of specific econometric techniques, rather than introducing them to a resampling technique (such as bootstrapping or randomization) which itself is a means of conducting inference. Simon and Bruce, in their statistical package Resampling Stats, are interested in showing how bootstrapping or randomization can be used to circumvent traditional statistical formulas in undertaking statistical inference. The strength of this approach is that, as the title of Simon and Bruce (1991) implies, it allows students more easily to produce answers to questions of "everyday statistical work" such as what is the probability that three of four children will be girls, or what is the probability that a basketball player who averages 47 percent success in shooting will miss eight of his next 10 shots? When we move into the world of the econometrician and ask about things like the properties of two-stage least squares or Lagrange multiplier tests, this approach becomes problematic. For example, one must begin with the mathematical formula in question, exactly what Simon and Bruce wish to avoid.

But what I suggest does share one major theme with Simon and Bruce. They state (1991, 29):

> *The first step in using probability and statistics is to translate the scientific question into a statistical question. ..... Though this step is difficult, it*

> *involves no mathematics. Rather, this step requires only hard, clear thinking. You cannot beg off by saying "I have no brain for math!" The need is for a brain that will do clear thinking, rather than a brain for mathematics......But resampling pushes you to do this thinking explicitly.*

What I propose is a means of pushing students to do clear, hard thinking about econometrics.

REFERENCES

Albert, J. and M. Berliner. 1994. Review of the resampling method of teaching statistics. *American Statistician* 48 (2): 129-31.

Boomsma, A. and I. W. Molenaar. 1991. Resampling with more care. *Chance* 4 (4): 25-9.

Jeong, J. and G. S. Maddala. 1993. A perspective on application of bootstrap methods in econometrics. In G. S. Maddala, C. R. Rao and H. D. Vinod (eds), *Handbook of statistics, Vol. 11*, pp. 573-610. Amsterdam: North Holland.

Kennedy, P. E. 1995. Randomization tests in econometrics. *Journal of Business and Economic Statistics* 13 (1): 85-94.

_____. 1998. *A guide to econometrics*, 4th ed. Cambridge, Mass.: MIT Press.

Kennedy, P. E. and B. S. Cade. 1996. Randomization tests for multiple regression. *Communications in Statistics - Computation and Simulation* 25 (4): 923-36.

Simon, J. L. and P. C. Bruce. 1991. Resampling: A tool for everyday statistical work. *Chance* 4 (1): 22-32.