Universität Bern Kurt Schmidheiny/Manuel Wälti
Doing Economics with the Computer Summer Term 2001

Session 9: Monte Carlo Experiments

1 The Topic

Doing statistics means often estimating parameters, such as the mean of a population, the coefficients in a linear regression or the autocorrelation of a time series, given a sample of real world data. Besides the point estimate itself, we would like to know how close our estimate is to the true value. In other words we would like to know its “quality” or “precision”. The properties of an estimator can be described by various aspects of the distribution of the estimator (the so-called *sampling distribution*), such as the mean and the variance of an estimator. The variance of an estimator can then be used to perform tests against hypothesis.

In some cases it is possible to calculate the sampling distribution from the statistical model. But sometimes, especially for finite (small) samples, this is either not possible or very difficult. In these cases Monte Carlo experiments are an intuitive way to obtain information about the sampling distribution and hence about the “quality” of the estimator.

2 The Method

The term “Monte Carlo” refers to procedures in which quantities of interest are approximated by generating many random realisations of some stochastic process and averaging them in some way. In statistics, the quantities of interest are the distributions of estimators and test statistics, the size of a test

statistic under the null hypothesis, or the power of a test statistic under some specified alternative hypothesis (see Davidson and MacKinnon 1993, 731). In economic theory Monte Carlo techniques are used to explore the quantitative properties of models with stochastic elements, for example the correlation between variables in real business cycle models.

How can we use Monte Carlo techniques to find the sampling distribution of an estimator? In the real world, we usually observe just one sample of a certain size T, that will give us just one estimate. The Monte Carlo experiment is a lab situation, where we replicate the real world study many (N) times. Every time, we draw a different sample of size T from the original population. Thus, we can calculate the estimate many times and any estimate will be a bit different. The empirical distribution of these many estimates approximates the true of the estimator.

A Monte Carlo experiment involves the following steps:

(1) Draw a (pseudo) random sample of size T for the stochastic elements of the stochastic model from their respective probability distribution functions
(2) Assume values for the exogenous parts of the model or draw them from their respective distribution function
(3) Calculate the endogenous parts of the statistical model
(4) Calculate the value (e.g. the estimator) you are interested in
(5) Replicate step 1 to 4 N times
(6) Examine the empirical distribution of the N values

Let's explain the above elements in an example: the bivariate ordinary least squares model

$$y_t = \alpha + \beta x_t + \varepsilon_t \qquad \text{with } \varepsilon_t \sim N(0, \sigma^2) \; .$$

The stochastic element in the model is ε_t, the exogenous part is x_t .Assuming values for the true parameters α and β, we can calculate the endogenous variable y_t. The values of interest are the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$.

In the core of Monte Carlo Experiments is the *random number generator*. A random number generator produces a sequence of numbers, that are draws from a specific identically and independently distributed random variable. In practice, this is a mathematical algorithm, that produces a sequence of numbers that look as if they were drawn from this specific identically and independently distributed random variable.[1]

There is an important drawback of Monte Carlo experiments: We must completely specify the Statistical Model (Data Generating Process DGP). This implies, that we must assume the deterministic parts of the model, the form and the exact parameters of the distribution of the stochastic (error) term and the distribution of exogenous variables. This is a great loss of generality as the results of the experiment depends on these assumptions.

3 The Software

Matlab has a good built-in random number generator. We will use the functions `rand` and `randn`, which are included in the basic Matlab

[1] These numbers are called *pseudo* random numbers, because they are not really random. In fact, the algorithm describes the purely deterministic relationship between the numbers. But with a good generator, they are indistinguishable from sequences of genuinely random numbers and pass usual statistical tests of independence. Judd (1998) provides a thorough treatment of different pseudo-random number generators.

installation. The `rand` function generates random numbers whose elements are uniformly distributed in the interval (0,1).

```
Y = rand(m,n)
```

returns an m-by-n matrix of random entries. `rand`, by itself, returns a scalar whose value changes each time it's referenced. We can reset the random number generator with

```
rand('state',0).
```

The `randn` function generates random numbers whose elements are normally distributed with mean 0 and variance 1.

```
Y = randn(m,n)
```

returns an m-by-n matrix of random entries. `randn`, by itself, returns a scalar whose value changes each time it's referenced. Again,

```
randn('state',0).
```

resets the generator.

Matlab's statistics toolbox contains random number generators for a variety of distribution functions.

4 References

Davidson, Russell and James G. MacKinnon (1993), *Estimation and Inference in Econometrics*, Oxford: Oxford University Press, chapter 21.

Greene, William H. (1997), *Econometric Analysis*, Upper Saddle River: Prentice Hall, chapter 5.3.3.

Judd, Kenneth L. (1998), *Numerical Methods in Economics*, Cambridge: MIT Press, chapter 8.

5 Today's Task

Exercise 1: Explore the random number generator

a) Draw 100 random number from a standard uniform distribution.
b) Plot the random numbers. Do you think they are independent?
c) Plot the histogram of the random numbers. Does it meet your expectations?
d) Calculate mean and standard deviation. Is this what you expected?
e) Draw another random sample of size 100. Is it different from the previous one? Reset the random number generator and draw again a random sample.
f) Do a) to d) the same with the standard normal and the Cauchy distribution (the appendix describes how to draw from this distribution).
g) Can you think of a way to draw random numbers from a normal distribution with mean 2 and variance 9 or a uniform distribution with limits 5 and 20.

Exercise 2: A Monte Carlo study of the OLS coefficients

We will now explore the properties of the coefficient estimates in the linear regression model as described in part 2. The value of interest are the two estimators:

$$\hat{\beta} = \frac{\sum_{t=1}^{T}(y_t - \bar{y})(x_t - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}$$

and

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} .$$

Assume that the true constant is zero and the true slope coefficient is one. Further assume that the exogenous variable x is drawn form a uniform distribution between -5 and +5 and that the error term is normally distributed with zero mean and variance set to unity.

a) Write an m-file that explores the properties of $\hat{\alpha}$ and $\hat{\beta}$ in a Monte Carlo experiment. Follow the 6 steps described in section 2. Implement the $N = 500$ replications in a for loop. Each random sample has size $T = 30$. This m-file will produce a 1 by N vector of both $\hat{\alpha}$ and $\hat{\beta}$.
b) Plot a histogram of both $\hat{\alpha}$ and $\hat{\beta}$. Compare this empirical sampling distribution with what you know from statistical theory. Contented?
c) Try sample sizes of 10, 100, 1000 and look at the sampling distribution.
d) Run similar experiments when the error term follows the uniform or the Cauchy distribution. Report your findings.

Exercise 3: A Monte Carlo study of the z-Test of an OLS coefficient

We will now look at the properties of a test against the null hypothesis that the true slope coefficient is one. The test statistic for this test is

$$t = \frac{\hat{\beta} - \beta_0}{\sqrt{V\hat{\beta}}} = \frac{\hat{\beta} - \beta_0}{\sqrt{\hat{\sigma}/S_{xx}}}$$

with

$$S_{xx} = \sum_{t=1}^{T}(x_t - \bar{x})^2 \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{T-2}\sum_{t=1}^{T}\hat{\varepsilon}_t^2 = \frac{1}{T-2}\sum_{t=1}^{T}(y_t - \hat{\alpha} - \hat{\beta}x_t)^2 .$$

a) Run a Monte Carlo experiment, where the value of interest is the above t-statistic.
b) Show the empirical distribution of the t-values. What is the theoretical distribution?

c) How many times was the test rejected at the 5% significance level (you may use the asymptotic rejection level of 1.96). Was that to be expected?

d) Try sample sizes of 10, 100, 1000 and look at the sampling distribution.

e) Run similar experiments when the error term follows the uniform or the Cauchy distribution. Report your findings.

Exercise 4: A Monte Carlo study of spurious regression

We will now explore the properties of the beta coefficient in the above linear regression model, when variables *x* and *y* follow two independent random walk processes with drift:

$$x_t = a + x_{t-1} + u_t \quad u_t \sim WN(0, \sigma_u^2)$$
$$y_t = b + y_{t-1} + v_t \quad v_t \sim WN(0, \sigma_v^2)$$

where the error terms *u* and *v* are independently distributed.

a) Derive the expected value and other properties of $\hat{\beta}$ in a Monte Carlo experiment. Take $a = b = 0.5$, $\sigma_u^2 = 1, \sigma_v^2 = 1$ and $x_1 = 0$, $y_2 = 0$.

b) Simulate the rejection rate of the Null hypothesis that the slope coefficient is 0. Again, take $a = b = 0.5$, $\sigma_u^2 = 1, \sigma_v^2 = 1$ and $x_1 = 0$, $y_2 = 0$. Simulate the rejection rate for different sample sizes and draw a graph.

c) Do b) with $a = b = 0$. This replicates the famous study by Granger and Newbold, 1974, Spurious Regression in Econometrics, Journal of Econometrics 2, 111-120. Show that there is a mistake in the report of the original Granger/Newbold simulation.

6 Appendix: Some standard probability distributions

The most important distribution is the *standard normal distribution*, N(0,1). Its probability density function is described as follows:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

Its expected value (mean) is $\mu = 0$ and its variance $\sigma^2 = 1$.

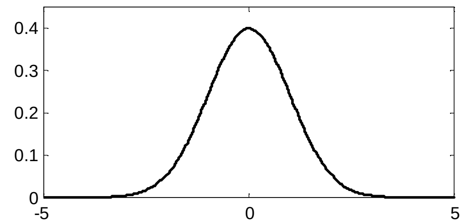

Figure 1: Standard normal probability density function

Another very common distribution is the *standard uniform distribution*. Its probability density function is described as follows:

$$f(x) = \begin{cases} 1 & \text{for } 0 \le x < 1 \\ 0 & \text{otherwise} \end{cases}.$$

Its expected value (mean) is $\mu = 0.5$ and its variance $\sigma^2 = 1/12$.

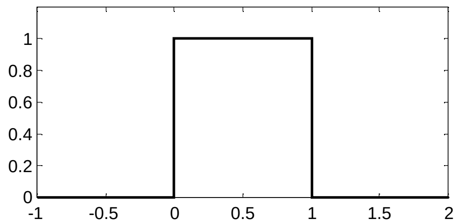

Figure 2: Standard uniform probability density function

A very peculiar is the *Cauchy distribution*. Its probability density function is described as follows:

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

The strange thing about the Cauchy distribution is that it has no moments, i.e. neither mean nor variance. The Cauchy distribution is equal to a Student distribution with one degree of freedom. It can be generated by dividing two independently drawn standard normally distributed variables.

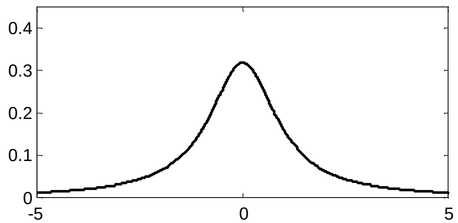

Figure 3: Cauchy probability density function