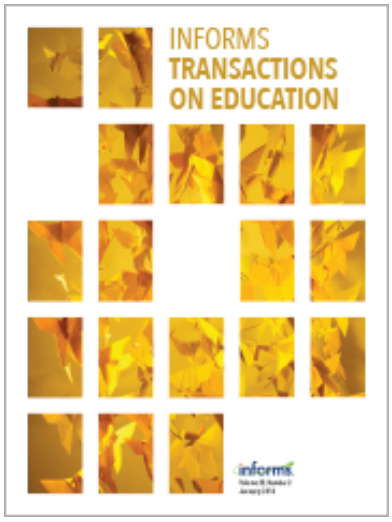


INFORMS Transactions on Education

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Monte Carlo Spreadsheet Simulation Using Resampling

Thin-Yin Leong,

To cite this article:
Thin-Yin Leong, (2007) Monte Carlo Spreadsheet Simulation Using Resampling. INFORMS Transactions on Education 7(3):188-200. https://doi.org/10.1287/ited.7.3.188

Vol. 7, No. 3, May 2007, pp. 188–200
ISSN 1532-0545 | 07 | 0703 | 0188
INFORMS Transactions on Education
DOI 10.1287/ited.7.3.188

Monte Carlo Spreadsheet Simulation Using Resampling

Thin-Yin Leong
School of Information Systems, Singapore Management University,
80 Stamford Road, Singapore S178902, tyleong@smu.edu.sg

The ubiquitous spreadsheet can be used to model situations with random values, in what is commonly referred to as Monte Carlo simulation. For simple cases, adding random functions such as Excel™'s RAND) is enough. In general business models, complex inverse distribution functions, in combination with RAND, are needed to generate the right random values. But first the modeler must determine the appropriate best-fit distribution to use. This can be a daunting process for undergraduates and typical executives. So for expediency, simulation add-ins (with additional learning time and possible costs) may be employed. The use of add-ins, however, makes the modeling less transparent. A more direct alternative is to resample the raw data, which in many cases are not sufficient in sample size to establish statistical goodness of fit. This paper reviews the limitations of current spreadsheet resampling methods and proposes new simple yet effective formulations that better accommodate classroom and practical real-world application.

1. Introduction

Monte Carlo simulations have been used to evaluate business situations where there are uncertainties and randomness. For each random variable, its values typically are assumed to vary according to common probability distributions; one is selected that best represents the variable's behavior. In spreadsheets, these would simply be cells having formulas that automatically generate random values. Simple cases, like those typically found in academic exercises, can use random functions such as Excel™'s RAND to generate the needed uniformly distributed random values. In general, complex inverse distribution functions, in combination with RAND, are required.

Modeling random behavior in a spreadsheet, as opposed to writing a computer subroutine using a programming language, is fairly straightforward and intuitive and at the same time powerful and flexible. However, there still are some difficulties. First, to be technically rigorous, field data have to be collected, candidate distribution functions test-fitted to this sample, and, for the selected function, the associated population parameters accurately estimated. Often the amount of data that can be collected is insufficient for these tasks to be properly done. Moreover, the procedures involved are challenging for typical undergraduate students and business executives, who have to do the modeling and analysis work, and to their managers, who have to understand and accept the models and analysis results.

For expediency, Excel simulation add-ins including CrystalBall™ (Decisioneering 2006), @Risk™ (Palisade Corporation 2006), XLSim™ (AnalyCorp 2006), Insight™ (Savage 2002), Simtools™ (Myerson 2006), and Resampling Stats™ (Resampling Stats 2006) have been used. These are comprehensive software packages, use of which usually means additional learning time for the students. Unless the software is applied extensively by a business, our experience suggests that it is really neither cost nor time effective. The packages have features to help modelers select the needed distribution and generate the random values. Though the technical burden is much reduced for the modeler, the process is less transparent, and the responsibility for technical accuracy is subtly relegated to the add-in software. For example, a good modeler should not blindly adopt a distribution just because the add-in software found it to have the best fit to the data sample. So although add-ins are valuable in simplifying the work, learning is compromised.

With some ingenuity, native spreadsheet features can provide better learning opportunities for students. They can appreciate and learn about the behavior of random variables, statistical thinking, and business modeling. This is consistent with industry feedback and, as elaborated in Leong and Cheong (2006), quantitative foundation and management science courses,

even with the aid of user-friendly software like Excel, can at best teach students to be better consumers of analysis (Powell 1997). Management education can be made more relevant to the market place by training students to be active modelers, end users building models to address immediate challenges and concerns.

For many years, we have been teaching a course on spreadsheet business modeling; it is mostly taken by second-year undergraduates from business, accounting, economics, information systems, and social sciences majors. The classes are conducted in an interactive manner, that is, seminar style with about 45 students per class. In this course, Monte Carlo simulation is taught at an introductory level over 2 weeks (i.e., 6 hours of class time). Our experience here is that add-ins, other than the Analysis Toolpak and Solver (both in the MS Office™ standard package), are really not needed for the short class module on Monte Carlo simulation.

Instead we employ resampling. Resampling eliminates the need to fit distributions to the sample data and the ensuing tests for goodness of fit, for which sufficient data must be present to achieve the needed level of statistical confidence. In contrast, resampling as a method requires less data, as it just uses whatever is available. Thus, it allows us to dive directly into business modeling without spending class time on statistical data fitting and distribution function inversion. These more complex topics are best covered in full-blown statistics and simulation courses, rather than in a business modeling course. Students would also find it easier to collect data for their team projects and future ad hoc modeling work. However, so far, spreadsheet resampling approaches have not been studied much. We therefore seek to review and evaluate past attempts in using spreadsheets to resample data and propose simpler yet effective formulations that better accommodate classroom and practical real-world application.

The rest of the paper is organized as follows. A brief review of the literature in resampling and how it relates to Monte Carlo simulation follows this introductory section. Next we elaborate on the current spreadsheet resampling formulations available. Then we propose new Excel formulations, first for discrete data with small ranges. This is followed by resampling methods with interpolations, for continuous data and discrete data with large ranges. A short discussion is included on bias errors arising from the use of frequency tables in resampling. The paper culminates in a section on resampling of multivariate dependent data, and finally, the paper ends with some concluding comments.

2. Literature Review

The term *bootstrapping* is often used in the literature interchangeably with *resampling*. Strictly, in bootstrapping, data are resampled from an original data sample to draw statistical inference about the population or its sampling distribution. That is, it is used for analyzing the data. This work was theoretically established as a general statistical method by Efron (1979) and Efron and Tibshirani (1993). Resampling as a method, however, is more correctly associated with Monte Carlo simulation, an idea derived from gamblers testing their chances by using repeated simulations to check their probability of winning. Simulations can be done by either sampling from a theoretical distribution or resampling from a data sample. So resampling is used here to generate data, as inputs to a business model.

Willemain (1994) first proposed that bootstrapping be made more accessible to teachers and students by implementing it in a simple spreadsheet. He demonstrated how the LOOKUP function may be used to randomly sample univariate raw data. Woodroof (2000) further elaborated with a more complete multivariate illustration on how Lotus 1-2-3™ (a spreadsheet program) can be applied to make bootstrapping easier to understand and implement.

More recently, Yu (2003) added that many resampling tools are available in standard applications, such as SAS™ and SyStat™, and also refuted the misconception that resampling is not founded on strong theoretical support. He did this by first providing a historical survey of the various resampling methods; he then reviewed popular arguments for and against the use of resampling. In many situations, as when the data population is ill defined and the sample size is small, exploring the empirical distribution by resampling should be favored over unjustified use of theoretical distributions.

Duckworth and Stephenson (2003) deliberated on how others have attempted to use resampling as a tool to teach courses in introductory statistics. They proposed that simulation (i.e., random sampling from a known population), resampling (i.e., further sampling from a given sample of the population), and the use of computing (to do either) are necessary statistical tools for health care professionals. In this paper, we are also interested in resampling, but closer to its origins in Monte Carlo simulations and for a broader purpose: its application in general business modeling.

As improvements over those proposed by other researchers, our resampling formulas are also updated to use Excel, a more commonly used spreadsheet program nowadays. One of our formulas has been applied (Leong 2007) to simulate parametric multi-server queues. For the discussion in this paper, we

assume that resampling is done with replacement and will cover both univariate and multivariate cases. A short note on the interesting resampling without replacement case is provided in Appendix 2. Variables in the multivariate case will be allowed to be dependent. If the variables are independent, we can easily use the univariate model multiple times.

3. Current Spreadsheet Resampling Approaches

To aid discussion, we will use an example (see Figure 1) to explain the resampling formulas. In this example, we seek to resample historical annual returns to simulate for next year the return rates of treasury bills, treasury bonds, and common stocks. Before we present our formulas, we first collate current available approaches and rewrite them as Excel formulas for ease of comparison.

The Excel equivalent of Willemain's (1994) Lotus 1-2-3 univariate formula, applied to our example, to resample treasury bill return rates would be as follows:

$$\text{M11=LOOKUP(INT(RAND()*n)+1), B\$11:B\$110, D\$11:D\$110),} \quad (1)$$

where n is the number of values in the original sample, B11:B110 is the indexRange and D11:D110 is the valueRange. (Explanations are provided in Appendix 1 for those who are less familiar with Excel functions.) The first argument in (1) is a formula that randomly generates an integer between 1 and n. Cell M11 in Figure 1 would be copied and pasted in the other rows below it, invoking Excel's relative referencing property, to replicate more resampled values.

Woodroof's (2000) multivariable formula in Excel would be:

$$\text{I11} = \text{INT(RAND())*n+1)} \quad (2)$$

$$\text{M11} = \text{LOOKUP(I11, B\$11:B\$110, D\$11:D\$110)} \quad (3a)$$

$$\text{N11} = \text{LOOKUP(I11, B\$11:B\$110, E\$11:E\$110)} \quad (3b)$$

$$\text{O11} = \text{LOOKUP(I11, B\$11:B\$110, F\$11:F\$110).} \quad (3c)$$

The essential difference here is that I11 is a common value shared by resampling formulas (3a)–(3c). Each time resampling is done, a whole row in the original data sample table is selected. This approach retains the statistical dependency among the different variables.

Instead of resampling from the original sample, an approach commonly encountered in textbooks (recent ones include Moore and Weatherford 2001, Gips 2003, Hillier and Hillier 2004, Winston 2004, and Powell and Baker 2004) uses LOOKUP functions to resample from probability tables. Such tables compress the sample data by presenting data values against their relative frequencies. Extra computational work is required to construct the tables (a step often skipped by the textbooks). Relative frequencies are taken to be good estimates of the probabilities of occurrences. Examples of such probability tables are shown in Table 1.

From the relative frequencies, students can compute the cumulative relative frequencies (as the estimated cumulative probabilities). From this, we can shift the values down one row (to be explained later) to obtain the lookupRange. To generate a random value according to the distribution of the data, we can use

$$\text{resampledValue} = \text{LOOKUP(RAND(), lookupRange, valueRange).} \quad (4)$$

Here, the LOOKUP function seeks out the largest value in lookupRange that is less than or equal to the value generated by RAND and remembers its relative position in the lookupRange. It then returns the value in the same relative position in the valueRange.

Figure 1 **Data and Basic Model**

	B	C	D	E	F	G	H	I	J	K	L	M	N	O
9			Raw Annual Returns									Resampled Annual Returns		
10	#	Year	T.Bills	T.Bonds	Stocks		#					T.Bills	T.Bonds	Stocks
11	1	1972	4.0%	2.8%	18.8%		1							
12	2	1973	5.1%	3.7%	-14.3%		2							
13	3	1974	7.5%	2.0%	-25.9%		3							
14	4	1975	7.2%	3.6%	37.0%		4							
15	5	1976	5.4%	16.0%	23.8%		5							
16	6	1977	4.4%	1.3%	-7.0%		6							
17	7	1978	6.1%	-0.8%	6.5%		7							
18	8	1979	9.1%	0.7%	18.5%		8							
19	9	1980	12.0%	-3.0%	31.7%		9							
20	10	1981	15.5%	8.2%	-4.7%		10							

Table 1 Examples of Probability Tables

Panel a

Number of children in household	Frequency	Rel Freq	LookupRange
0	5	0.1786	0.0000
1	6	0.2143	0.1786
2	9	0.2857	0.3929
3	5	0.2143	0.6786
≥ 4	3	0.1071	0.8929

Panel b

Time between Arrivals (mins)	Probability	LookupRange
5	0.25	0.00
10	0.50	0.25
15	0.25	0.75

Panel c

Demand per week	Probability	LookupRange
10,000	0.13	0.00
20,000	0.25	0.13
30,000	0.38	0.38
40,000	0.19	0.75
50,000	0.06	0.94

Formulation (4) works best for discrete data with small finite ranges, as in Table 1(a).

4. New Resampling Approaches

To resample discrete data with small ranges, still referring to Figure 1, we can use

$$\text{M11} = \text{LOOKUP}(\text{RANDBETWEEN}(1, n), \text{B\$11:B\$110}, \text{D\$11:D\$110}). \quad (5)$$

This formula is modified from (1), making only a minor change-replacing RAND with RANDBETWEEN. RANDBETWEEN(1, n) also uniformly samples from $\{1, 2, \ldots, n\}$, but this function is easier to comprehend. The Analysis Toolpak add-in needs to be activated before RANDBETWEEN can be used.

We further simplify (5) to propose the following:

$$\text{M11} = \text{INDEX}(\text{D\$11:D\$110}, \text{RANDBETWEEN}(1, n)). \quad (6)$$

This formula also directly resamples, with equal probability, the n values of the original sample, stored in valueRange D11:D110. It is simpler than (1) and (5) because it no longer requires the indexRange B11:B110. The values resampled would appear with probabilities that match their relative frequencies in the original sample.

Other possible equivalent formulations are

$$\text{M11} = \text{SMALL}(\text{D\$11:D\$110}, \text{RANDBETWEEN}(1, n)) \quad (7)$$

and

$$\text{M11} = \text{LARGE}(\text{D\$11:D\$110}, \text{RANDBETWEEN}(1, n)). \quad (8)$$

Functions SMALL and LARGE in (7) and (8) implicitly sort the values in the valueRange D11:D110 in ascending and descending order, respectively. Sorting may appear to be unnecessary, in contrast to (1), (5), and (6); nonetheless, it costs negligible computation time in Excel and is no more difficult to read than (5). The explanation for (7) goes as follows: assign 1/n probability to each of the sorted data values and then compute their cumulative probabilities. Resampling using (7) is thus equivalent to sampling from the population's inverse empirical distribution. Sampling from an inverse distribution is a well-accepted approach for random values generation, and in using the empirical distribution (7) does so without needing to justify a theoretical distribution. Formulas (5)–(8) are improvements over (4), resampling directly from the original sample data, as the effort to construct the probability table is saved. Overall, the preferred approach is the SMALL formulation (7), as it parallels the PERCENTILE formulation, which (to be explained next) resamples for the continuous data case.

5. New Resampling with Interpolation

For resampling formulas in §§3 and 4 to be effective, we should strictly require all the possible data values to be represented in the sample. That is, if the variable ranges from 1 to 25, there must be at least one count of 1, at least one count of 2, …, and at least one count of 25 in the sample. Satisfying this requirement could mean very large samples when the discrete variable has a nonuniform distribution or a large range. Applying these formulas for discrete data

with large ranges or even for continuous data will yield subranges of discontinuities in the resampled values.

To mitigate this imperfection, we propose in this section more new formulations http://archive.ite.journal.informs.org/Vol7No3/Leong/Resampling.xls. We focus first on continuous variables and later elaborate how this formulation can be modified for discrete variables with large ranges.

The proposed formula for resampling continuous data is

$$\text{M11} = \text{PERCENTILE(D\$11:D\$110, RAND())}. \quad (9)$$

Similar to the description made for the SMALL formulation (7), this formula samples from the inverse empirical distribution. The additional advantage here is that although the empirical distribution in (7) is a series of staircase steps, this empirical distribution is piecewise linear (joining consecutive values in the original sample) and thus smoother. The random values it generates include the original sample data values and all values linearly interpolated in between. Again, no probability table needs to be created and there is no loss of information from aggregating the data.

Formulation (9) looks deceptively simple, but it stealthily does all the massive sorting (in ascending order) and interpolation work. With a small modification, it turns into the formula for resampling discrete data with large ranges. This is done by incorporating the ROUND function in (9) as follows:

$$\text{M11} = \text{ROUND(PERCENTILE(D\$11:D\$110, RAND()),0)}. \quad (10)$$

When (9) is used on discrete data, the generated random value, as a result of the implicit interpolation, can be nondiscrete. Applying the ROUND function converts this continuous value into a discrete value.

By visual inspection (see Appendices 3 and 4), we can see that cumulative relative frequency plots of the interpolated resampled values have smaller maximum absolute deviations from the cumulative relative frequency of the original sample than resampling without interpolation. It is thus an improvement in both form and outcome.

6. Frequency Table Bias

When constructing probability tables for continuous data (e.g., Table 1(b)) or discrete data with large ranges (e.g., Table 1(c)), the relative frequencies are given for data intervals, typically referred to as "bins." In Excel, the binRange contains the upper limits of the intervals. On the other hand, the LOOKUP function needs the lower limits of the intervals as input. The shifting down of the cumulative relative frequency by one row, to give the lookupRange, corrects for this. The same needs to done for the discrete data case. We have found that students are either not adept in doing this or have tremendous difficulty accepting the reasons they should. "N/A" errors will arise sporadically when the cumulative relative frequency data range is used as the lookupRange because LOOKUP cannot find relative positions for RAND values smaller than the first cumulative relative frequency value in the table.

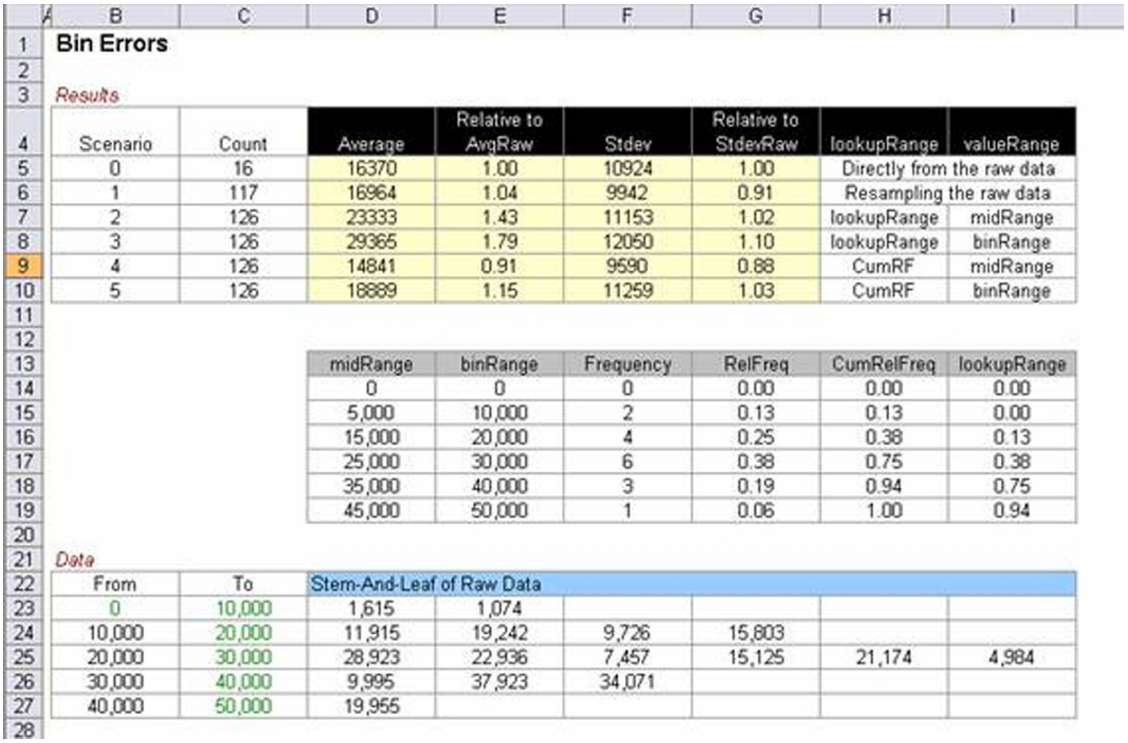
Furthermore, to avoid bias, the binRange should not be used in (4) as the valueRange, a common mistake made by students and textbook writers alike. The bins' midpoint values should be used instead. Using the data in Table 1(c), we will illustrate the bias that can arise for all of these reasons. We first arbitrarily simulate a sample of 16 raw data points (shown as the Stem-and-Leaf table in Figure 2), uniformly distributed within the bin intervals, with frequency occurrences that approximately yield the relative frequencies of Table 1(c). From this raw sample, we can compute its mean and standard deviation. Data are then simulated by resampling according to five other scenarios, described in cells H6:I10 of Figure 2. For each of these scenarios, 126 data values were resampled. We computed the mean and standard deviation of these five data sets and compared them against that of the raw data.

The result of one typical representative sample run shows that the sample mean under Scenario 1 (resampling using (10) directly from the sample data) is only 4% larger than the mean of the raw data. The sample mean for Scenario 2 (using bins, with the correct lookupRange and midRange as valueRange) is 43% larger. It was 79% larger when the binRange was incorrectly used as the valueRange (in Scenario 3). The deviation of sample mean for Scenario 4 (using cumulative relative frequency as lookupRange and midRange as valueRange) at 9% is about the same as that of Scenario 5 (also using cumulative relative frequency as lookupRange, but now with binrange as valueRange) at 15%, both of which are inherently incorrect representations of the data distribution. The standard deviations for all the scenarios are comparable. In all, because data are aggregated into bins, resampling using (4) requires more work and has a large bias, and not using the midpoints of bins as resampled values generally makes it worse.

7. Multivariate Resampling

The new formulations discussed so far are for univariate cases. We can apply them without change to multivariate cases if we assume that the variables are statistically independent. For the general multivariate

Figure 2 Bin and Other Errors



Bin Errors

Results

Scenario	Count	Average	Relative to AvgRaw	Stdev	Relative to StdevRaw	lookupRange	valueRange
0	16	16370	1.00	10924	1.00	Directly from the raw data	
1	117	16964	1.04	9942	0.91	Resampling the raw data	
2	126	23333	1.43	11153	1.02	lookupRange	midRange
3	126	29365	1.79	12050	1.10	lookupRange	binRange
4	126	14841	0.91	9590	0.88	CumRF	midRange
5	126	18889	1.15	11259	1.03	CumRF	binRange

midRange	binRange	Frequency	RelFreq	CumRelFreq	lookupRange
0	0	0	0.00	0.00	0.00
5,000	10,000	2	0.13	0.13	0.00
15,000	20,000	4	0.25	0.38	0.13
25,000	30,000	6	0.38	0.75	0.38
35,000	40,000	3	0.19	0.94	0.75
45,000	50,000	1	0.06	1.00	0.94

Data

From	To	Stem-And-Leaf of Raw Data					
0	10,000	1,615	1,074				
10,000	20,000	11,915	19,242	9,726	15,803		
20,000	30,000	28,923	22,936	7,457	15,125	21,174	4,984
30,000	40,000	9,995	37,923	34,071			
40,000	50,000	19,955					

case, resampling (without interpolation) can be done by refining fomulations (2)–(3c) as follows:

$$\text{I11} = \text{RANDBETWEEN}(1, \text{n}) \quad (11)$$

$$\text{M11} = \text{SMALL}(\text{D\$11:D\$110, I11}) \quad (12a)$$

$$\text{N11} = \text{SMALL}(\text{E\$11:E\$110, I11}) \quad (12b)$$

$$\text{O11} = \text{SMALL}(\text{F\$11:F\$110, I11}). \quad (12c)$$

The data for the variables are stored in columns, one for each variable: D11:D110 for treasury bills, E11:E110 for treasury bonds and F11:F110 for common stocks' annual return rates. In multidimensional space, each row represents a data point, so randomly sampling the rows preserves the correlation among the data values from the different variables. Again, this approach is only appropriate for discrete variables with small ranges.

For continuous data and discrete data with large ranges, we propose the following resampling (with interpolation) formulas:

$$\text{I11} = \text{RAND}() \quad (13)$$

$$\text{J11} = \text{RANDBETWEEN}(1, \text{n}-1) \quad (14)$$

$$\text{K11} = \text{MATCH}(\text{SMALL}(\text{F\$11:F\$110, J11}), \text{F\$11:F\$110,0}) \quad (15a)$$

$$\text{L11} = \text{MATCH}(\text{SMALL}(\text{F\$11:F\$110,J11+1}), \text{F\$11:F\$110,0}) \quad (15b)$$

$$\text{M11} = \text{I11*INDEX}(\text{D\$11:D\$110,K11}) + (1\text{-I11}) \text{*INDEX}(\text{D\$11:D\$110,L11}) \quad (16a)$$

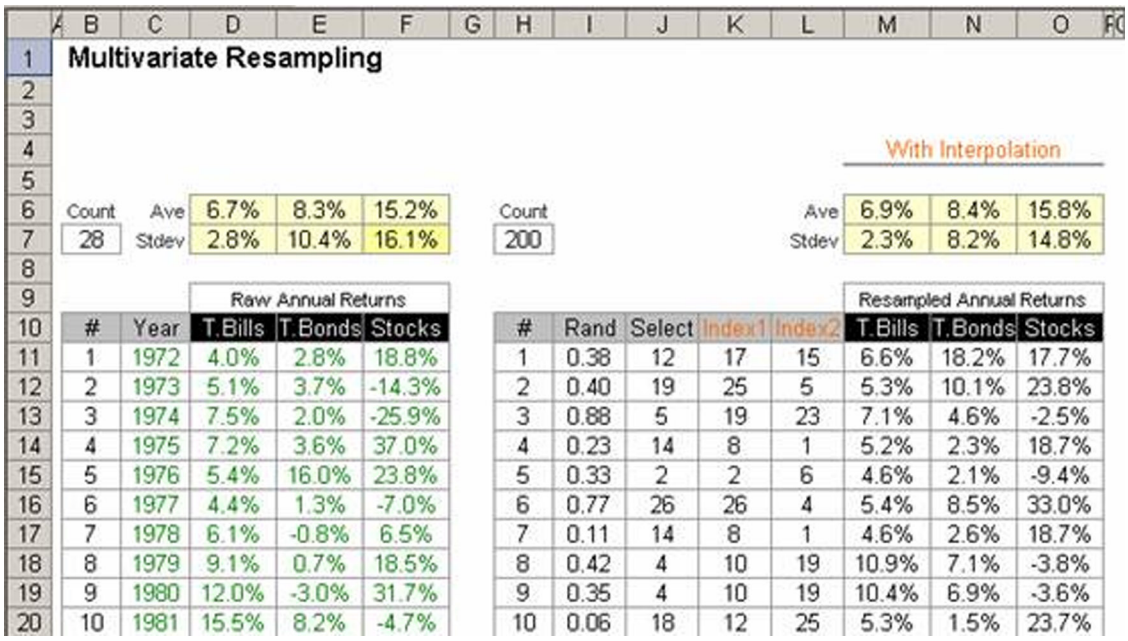
$$\text{N11} = \text{I11*INDEX}(\text{E\$11:E\$110,K11}) + (1\text{-I11}) \text{*INDEX}(\text{E\$11:E\$110,L11}) \quad (16b)$$

$$\text{O11} = \text{I11*INDEX}(\text{F\$11:F\$110,K11}) + (1\text{-I11}) \text{*INDEX}(\text{F\$11:F\$110,L11}). \quad (16c)$$

Value range F11:F110 in (15a) and (15b) could be replaced by any of the other variables' value ranges (i.e., E11:E110 or D11:D110). We call the variable selected for this role the *pivot variable*. Equation (14) randomly selects an (ascending) ordered position in the pivot variable. Equations (15a) and (15b) compute the relative (nonsorted) positions in the pivot variable that correspond to the selected ordered position J11 and its adjacent ordered position J11 + 1. Adjacency of data points here is defined with respect to the sorted order of the pivot variable. It then takes, in Equations (16a)–(16c), a weighted average of the two values in the relative positions K11 and L11 for each variable, using uniform random [0, 1) value in I11 as a weight.

These are complex formulas but not too difficult to explain and, if done systematically, are easily constructed in Excel. The end result is shown in Figure 3. The resampled data can be used in a class exercise to determine the optimal portfolio mix http://archive.

Figure 3 **Multivariate Resampling with Interpolation**



Multivariate Resampling

															With Interpolation			
	Count	Ave	6.7%	8.3%	15.2%		Count					Ave	6.9%	8.4%	15.8%			
	28	Stdev	2.8%	10.4%	16.1%		200					Stdev	2.3%	8.2%	14.8%			
			Raw Annual Returns												Resampled Annual Returns			
	#	Year	T.Bills	T.Bonds	Stocks		#	Rand	Select	Index1	Index2		T.Bills	T.Bonds	Stocks			
	1	1972	4.0%	2.8%	18.8%		1	0.38	12	17	15		6.6%	18.2%	17.7%			
	2	1973	5.1%	3.7%	-14.3%		2	0.40	19	25	5		5.3%	10.1%	23.8%			
	3	1974	7.5%	2.0%	-25.9%		3	0.88	5	19	23		7.1%	4.6%	-2.5%			
	4	1975	7.2%	3.6%	37.0%		4	0.23	14	8	1		5.2%	2.3%	18.7%			
	5	1976	5.4%	16.0%	23.8%		5	0.33	2	2	6		4.6%	2.1%	-9.4%			
	6	1977	4.4%	1.3%	-7.0%		6	0.77	26	26	4		5.4%	8.5%	33.0%			
	7	1978	6.1%	-0.8%	6.5%		7	0.11	14	8	1		4.6%	2.6%	18.7%			
	8	1979	9.1%	0.7%	18.5%		8	0.42	4	10	19		10.9%	7.1%	-3.8%			
	9	1980	12.0%	-3.0%	31.7%		9	0.35	4	10	19		10.4%	6.9%	-3.6%			
	10	1981	15.5%	8.2%	-4.7%		10	0.06	18	12	25		5.3%	1.5%	23.7%			

ite.journal.informs.org/Vol7No3/Leong/OptimalPortfolio.xls of the three groups of investment instruments. Remember to "freeze" the resampled values once they have been computed by the cell formulas (by Copy and PasteSpecial Values to itself in Excel). Otherwise, Excel's Solver will not be able to find the optimal solution, because the data values would be ever changing while Solver is running. The scattergrams and cumulative relative frequency charts of the illustrative example (in Appendices 3 and 4) show that the proposed formulas are effective in resampling data.

In our limited computational experiments, we found that choosing the variable with the largest standard deviation (in this case common stocks) as the pivot variable would give the best variance reduction effect, compared with the cases when the other variables are used as pivot variables. The intuition behind this is that this variable has the flattest marginal cumulative distribution function (CDF). Then randomly choosing any two adjacent data points along this profile will give a relatively larger interval to interpolate in, to generate new data points. These two data points are only adjacent along the pivot variable's marginal CDF but may be far apart with respect to the other variables' marginal CDFs. Because the other variables' marginal CDFs are steeper, any new point created by interpolating between two nonadjacent points along them would not deviate too far from their profiles. Detailed theoretical and computational analyses to support this argument and also to show the effectiveness of the resampling formulas are left for future research.

8. Comments

The major contributions of this paper are in highlighting bias that arises from both proper and improper use of frequency tables in resampling, providing new streamlined formulations in Excel for resampling (with and without replacement) for both univariate and multivariate cases. As typical of resampling methods, no assumption is made here of the underlying distributional form or its parameter values. Therefore, there is no need to test goodness of fit and accuracy of parameter estimates and no need to keep a ready set of required inverse distribution formulas for generating the random values.

We found that distribution fitting and the use of inverse distribution formulas are really challenging for the undergraduates whom we teach, and we would have required additional class time to discuss them. Though available commercial add-ins can ease the pain, the use of unfamiliar add-ins puts new stress on students, and class time is again diverted to learn and develop competency to use them. To use resampling to generate data for our classroom model-building exercises, we need to prepare sample data sets ahead of time. For project work, students do field work to collect the data. Now with resampling, even students with statistics phobia can do Monte Carlo simulation modeling.

Another reason that add-ins are used is to make it easier to replicate the simulations and collate the results. Excel's DataTable feature (not found in older Excel versions and not to be confused with a table of data) can now be used to easily automate replications (refer to Appendix 1 for details). Because computing

DataTables would have been taught in the first few weeks of the term, its application in Monte Carlo simulation reinforces learning. However, DataTables is computationally intensive and thus can completely slow down all the opened spreadsheets. It should therefore be avoided whenever possible. For fast calculations, we recommend that relative referenced formulas be set up as base equations in the top row of the simulation table and copied down to other rows to complete the model. It is, however, appropriate to use DataTable to extract and tabulate the Monte Carlo simulation summary results. Even so, remember to "freeze" the values (i.e., Copy and PasteSpecial Values into itself) after they are collected.

Using only native Excel features, the undergraduates in our business-modeling course learn to build reasonably sophisticated spreadsheet models of problems in operations, marketing, and finance business areas. Through the problem-based pedagogy and studio classroom environment, students develop good intimate knowledge of Excel spreadsheet use, and more importantly, can apply them effectively to business problems. This is assessed by having student teams do real-world projects in the community, using what they have learned in the course. Typically, the Excel workbooks the teams generate are of high quality and win the praise of companies and organizations for which they were developed.

Appendices

A. Appendix 1—Excel Functions

RAND

RAND( ) returns with equal probability a random value in [0, 1), i.e., between 0 and 1, inclusive of 0 but not inclusive of 1. This can be used in formulas to generate random values of other distributions and interval values. For example, to generate a random real value between 5 and 9, we can use $5+\text{RAND}(\,)*(9-5)$. So when RAND is 0, it returns 5; when RAND is 1, it returns 9; and when RAND is a value between 0 and 1, it generates the interpolated value between 5 and 9, with uniform probability density.

RANDBETWEEN

RANDBETWEEN(a, b) returns with equal probability a random integer in {a, a+1, . . . , b}, i.e., integers between a and b, a < b. For example, RANDBETWEEN(1, 6) can be used to simulate the toss of a die. This function is only available when the Analysis Toolpak (a standard add-in) has been

Figure A2-1 **Resampling Without Replacement Worksheet (Partial)** ResamplingWithoutReplacement.xls http:// archive.ite.journal.informs.org/Vol7No3/Leong/ResamplingWithoutReplacement.xls

Resampling Without Replacement

Data

#	Children in House	Rand	Sorted
1	0	0.5285	0.0170
2	0	0.0767	0.0259
3	0	0.0170	0.0767
4	0	0.6094	0.0921
5	0	0.5639	0.1151
6	1	0.9145	0.1537
7	1	0.8212	0.1990
8	1	0.0259	0.3291
9	1	0.8319	0.3558
10	1	0.5930	0.5021
11	1	0.0921	0.5156
12	2	0.5156	0.5285
13	2	0.5553	0.5390
14	2	0.5021	0.5553
15	2	0.6861	0.5639
16	2	0.9161	0.5930
17	2	0.5390	0.5930
18	2	0.9037	0.6094
19	2	0.7947	0.6413
20	2	0.5930	0.6861
21	3	0.3558	0.7344
22	3	0.1537	0.7947

Resampling

Visit	Children in House
1	2
2	0
3	0
4	2
5	2
6	≥ 4
7	3
8	0
9	3
10	2
11	0
12	1
13	2
14	1
15	2
16	≥ 4
17	2
18	≥ 4
19	3
20	2
21	1
22	1

activated in Excel. You can do this by choosing Tools/Add-ins/Analysis Toolpak in Excel's main menu.

LOOKUP

LOOKUP(lookupValue, lookupRange, valueRange) returns the value in valueRange that has the same relative position that lookupValue has in lookupRange. The values in lookupRange must be in ascending order. If LOOKUP cannot find lookupValue in lookupRange, it matches the largest value in lookupRange that is less than or equal to lookupValue. For example, LOOKUP(6, A1:A6, B5:B11) with (3, 5, 7, 9, 11, 12) in A1:A6 and (2, 4, 6, 8, 10, 12) in B5:B11 would return the value of 4, as the value 5 in A1:A6—the largest value less than or equal to 6—is in the second position, and the second value in B5:B11 is 4. LOOKUP(RAND(), lookupRange, mid-binValueRange) returns a random value, according to the distribution of the values listed in the table defined by the mid-bin value and lookup ranges. Excel has other lookup functions that behave similarly.

INDEX

INDEX(valueRange, rowIndex) returns the value found in the relative rowIndex position in valueRange. This is easier to understand than LOOKUP because there is no need to have an index range.

MATCH

MATCH(cellRange, valueRange, 0) returns the relative position that the value in cellRange is in valueRange. The 0 in the last argument denotes that an exact match between the cellRange value and a value in valueRange is required.

SMALL

SMALL(valueRange, k) returns the kth smallest value in the data set specified by valueRange. For example, SMALL(A1:A20, 1) returns the smallest value in A1:A20 (same result as MIN(A1:A20)), SMALL(A1:A20, 2) returns the second smallest value, SMALL(A1:A20, 3) returns the third smallest value, and so on. When used against a column with running serial numbers 1, 2, 3, . . ., it can be used to sort a set of numbers in ascending order. SMALL(valueRange, RANDBETWEEN(1, n)) resamples with equal probability the n sample data values stored in valueRange.

LARGE

LARGE(valueRange, k) returns the kth largest value in the data set specified by valueRange. For example, LARGE(A1:A20, 1) returns the largest value in A1:A20 (same result as MAX(A1:A20)), LARGE(A1:A20, 2) returns the second largest value, LARGE(A1:A20, 3) returns the third largest value, and so on. When used against a column with running serial numbers 1, 2, 3, . . ., it can be used to sort a set of numbers in descending order. LARGE(valueRange, RANDBETWEEN(1, n)) also resamples with equal probability the n sample data values stored in valueRange.

PERCENTILE

PERCENTILE(valueRange, k) returns the kth fractile from among the data in valueRange. For example, PERCENTILE(A1:A5, 0.5) with (91, 33, 52, 45, 67) in cell range A1:A5 returns the 50th percentile value of 52. PERCENTILE (valueRange, RAND()) works like SMALL(valueRange, RANDBETWEEN(1, n)), but PERCENTILE also interpolates when the fractile required does not coincide with one of the original sample data points, and it returns a continuous value.

DATATABLE

DataTable is not really a function, but rather a computational feature in Excel. To set up a one-dimensional DataTable, construct a table such that the first column contains the candidate input values, with the other columns

Figure A3-1 Scattergram of T-Bond vs. T-Bill

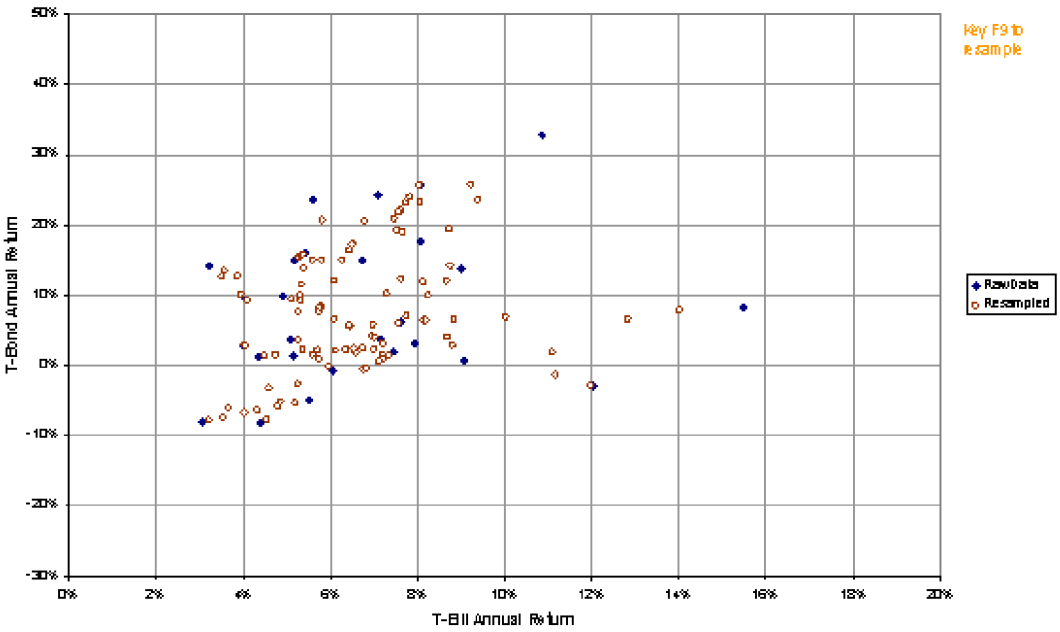
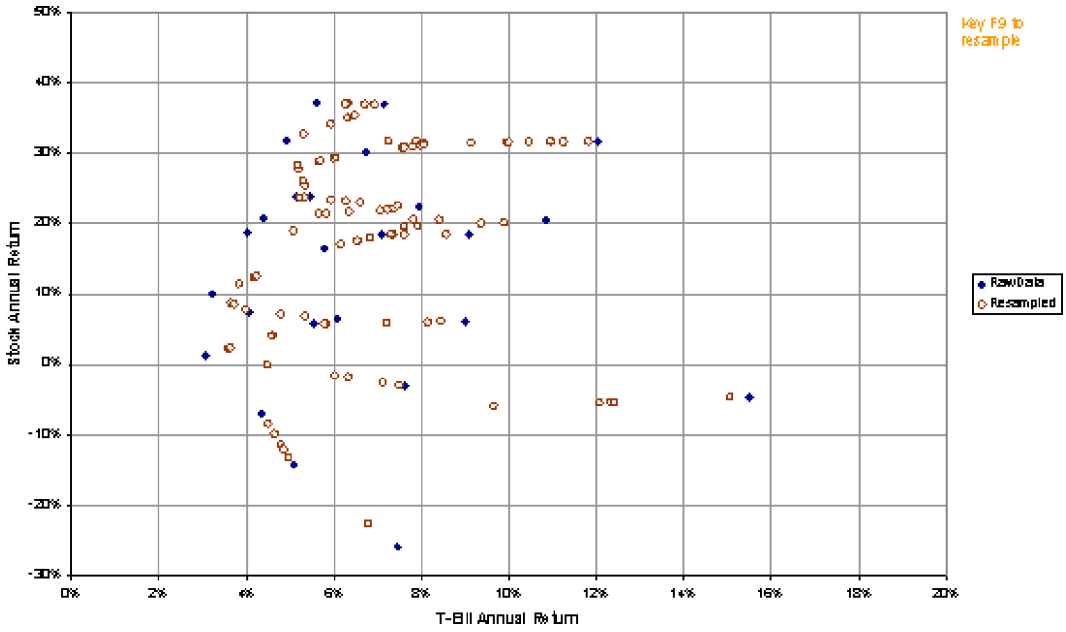


Figure A3-2 Scattergram of Stock vs. T-Bill



designated to store the desired output values after they are computed. Leave the row just below the header empty. For cells in column 2 onward in this row and under the corresponding headers, enter the required output formulas or link them to where the required computed results are found. Select the whole table (less the headers) and then activate the computation by selecting Data/Table from the main menu. In the pop-up box, enter the cell reference where the column input values should be applied. The input values will be then entered successively into the column input cell and the results pulled into the table in the corresponding rows under the appropriate headers to complete it.

To use DataTable to collect replication results in a Monte Carlo simulation, simply use column one to represent the replication counter and use any arbitrary empty cell as the column input cell.

Appendix 2—Resampling Without Replacement http://archive.ite.journal.informs.org/Vol7No3/Leong/ResamplingWithoutReplacement.xls

It is not too difficult to resample without replacement in Excel, but it is meaningful only for discrete variables with small ranges. Otherwise, the sample may not contain all the possible values from the population.

Figure A3-3 Scattergram of Stock vs. T-Bond

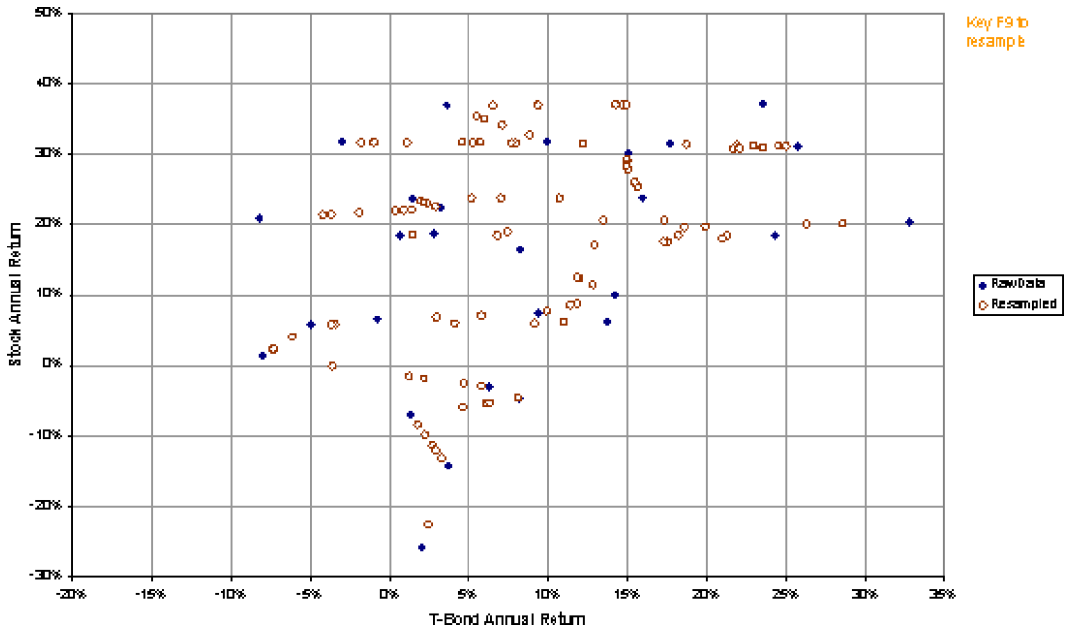
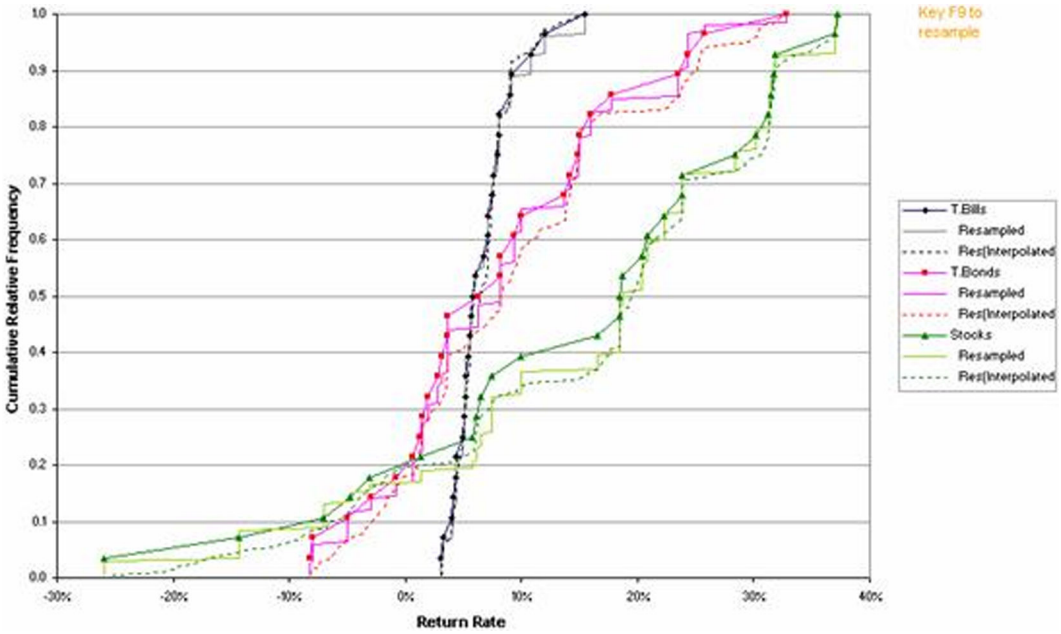


Figure A4-1 Cumulative Relative Frequencies of Univariate Resampled Data



The data from the table above are elaborated from Table 1(a) of the main text. The frequency of the number of children in each household is now reflected in the table by the number of rows of that entry. An explanation on how to resample without replacement goes like this: for each data point, a random value is generated. The data points are to be resampled in ascending order of their associated random values. To sort the values in Excel, the SMALL function is used, as this is interactive, unlike the SORT feature found in Excel's main menu.

The formulas for the relevant key cells are given below:

Serial numbers B5:B32 <Input>, 1, 2, . . . , 28
Children in house hold data C5:C32
<Input> [One for each occurrence]
Random values D5:D32 =RAND()
Sorted random value E5 =SMALL(D$5:D$32,B5)
Resampled no. of children H5
=INDEX($C$5:$C$32,MATCH(D5,$E$5:$E$32,0))

Figure A4-2 Cumulative Relative Frequencies of Multivariate Resampled Data

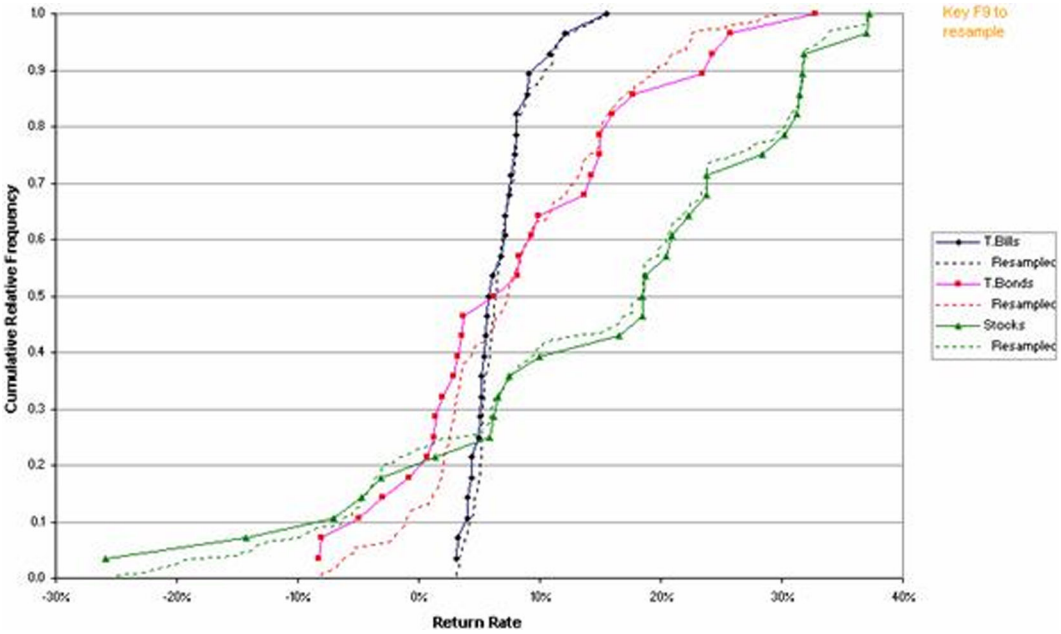
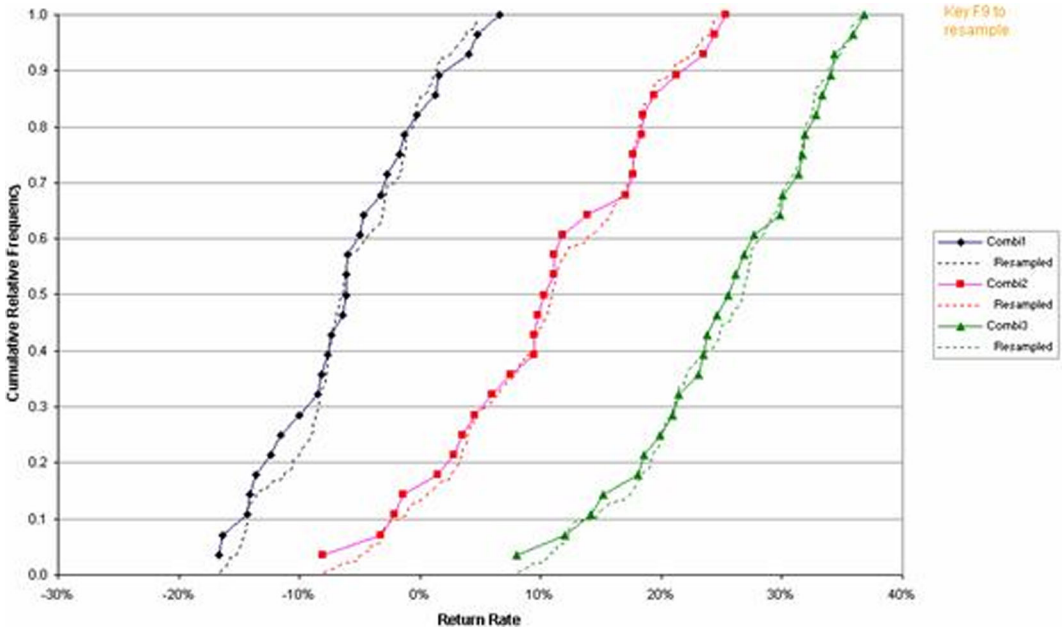


Figure A4-3 Cumulative Relative Frequencies of Linear Combinations of Variables



Appendix 3—Scattergram

These three scattergrams show the data points in three-dimensional space, two dimensions at a time. The axes represent the annual return rates of the investment instrument types, namely T-Bill, T-Bond, and stocks. The dark filled diamond markers indicate the positions of the sample data points. The unfilled diamond markers are the resampled (with interpolation) data points. Resampling without interpolation would result in only placing the unfilled markers over the filled markers, and nowhere else. For the small number of data points resampled, some of the filled markers will not be covered by unfilled markers, and others may even be covered more than once. Depending on the Monte Carlo simulation model built, the nonsmooth distribution of the resampling data points without interpolation may or may not pose problems to the model.

In the Excel workbook, keying F9 will recalculate the worksheets and thereby regenerate a new set of resampled data points. For resampling with interpolation, the unfilled markers will move almost Brownian motion-like within the convex hull formed by the filled markers, each time F9 is keyed. This is true for all three charts. Because the unfilled markers can only be found within the convex hull of the filled markers and the data sample, represented by filled markers, does not necessarily contain all the extreme data points in the true population, the marginal standard deviations of the resampled data are expected to be slightly smaller than those in the population. They would, however, converge to the population values as the sample size increases.

Appendix 4—Cumulative Relative Frequency Chart

The three charts in this appendix show the data points in their cumulative relative frequencies. The purpose is to demonstrate visually how well the resampled data in the various situations conform to the distributional properties of the given data sample. The cumulative relative frequency curves in Figure A4-1 come in triplets: one for the raw data sample, one for data resampled without interpolation, and one for data resampled with interpolation. Though a comprehensive analysis was not done, we can see that under the univariate (or equivalently independent multivariate) case the resampled data fit the empirical distribution fairly well with fairly small maximum absolute deviations.

Figure A4-2 shows how well resampling works for the general multivariate (dependent variable) case. The resampled marginal distribution of the three variables fits the distribution of the sample data very well. In fact, it seems to do better than the independent variable case. This suggests that interpolating across multiple variables may be better than interpolating within a variable itself, which in turn means that multivariate resampling with interpolation approach is applicable for all multivariate situations, whether or not the variables are dependent.

Often, stochastic values are used in combination with other stochastic values. The situation here is that a portfolio consists of a combination of T-Bills, T-Bonds, and common stocks. The relative holding proportion of these instruments would determine the final return of the invested funds. Thus, it would be useful to see if the distributional properties of linear combination of the resampled values would also stay faithful to that of the linear combination of the raw data sample. The visual conclusion from Figure A4-3 is a resounding yes.

References

AnalyCorp. 2006. XLSim and Insight, http://www.analycorp.com/ (last accessed on March 15, 2006).

Barreto, H., F. Howland. 2005. *Introductory Econometrics: Monte Carlo Simulation Using Excel*, Cambridge University Press, Cambridge, UK.

Decisioneering. 2006. Crystal Ball, http://www.decisioneering.com/ (last accessed on March 15, 2006).

Duckworth, W. M., W. R. Stephenson. 2003. Resampling methods: Not just for statisticians anymore. *2003 Joint Statistical Meetings*. San Francisco.

Efron, B. 1979. Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26.

Efron, B., R. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall Ltd.

Gips, J. 2003. *Mastering Excel: A Problem-Solving Approach*, 2nd ed. Wiley, New York.

Hillier, F. S., M. S. Hillier. 2004. *Introduction to Management Science: A Modeling and Case Studies Approach with Spreadsheets*, 2nd ed. McGraw Hill, New York.

Leong, T.-Y. 2007. Simpler spreadsheet simulation of multi-server queues. *INFORMS Trans. Ed.* **7**(1), http://ite.pubs.informs.org/Vol7No2/Leong/.

Leong, T.-Y., M. L. F. Cheong. 2006. Teaching business modeling using spreadsheets. Working paper, Singapore Management University, Singapore.

Moore, J. H., L. R. Weatherford. 2001. *Decision Modeling with Microsoft Excel*, 6th ed. Prentice Hall, Upper Saddle River, NJ.

Myerson, R. 2006. Simtools.xla, http://home.uchicago.edu/~rmyerson/addins.htm (last accessed on March 22, 2006).

Palisade Corporation. 2006. @Risk, http://www.palisade.com.au/, (last accessed on March 15, 2006).

Powell, G. S. 1997. From intelligent consumer to active modeler, Two MBA success stories. *Interfaces* **27**(3) 88–99.

Powell, S. G., K. R. Baker. 2004. *The Art of Modeling with Spreadsheets: Management Science, Spreadsheet Engineering, and Modeling Craft*. Wiley, New York.

Resampling Stats. 2006. Resampling Stats for Excel 3.2, http://www.resample.com/content/software/excel/index.shtml, (last accessed on March 22, 2006).

Rogers, J. L. 1999. The bootstrap, the Jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Res.* **34**(4) 441–456.

Savage, S. L. 2002. *Decision Making with Insight*. Duxbury Press, Pacific Grove, CA.

Willemain, T. R. 1994. Bootstrap on a shoestring: Resampling using spreadsheets. *The Amer. Statistician* **48**(1) 40–42.

Winston, W. L. 2004. *Microsoft Excel: Data Analysis and Business Modeling*. Microsoft Press, Redmond, VA.

Wittwer, J. W. 2004. Generating random numbers in excel for monte carlo simulation. *Vertex42.com* (June 1), http://vertex42.com/ExcelArticles/mc/SalesForecast.html.

Woodroof, J. 2000. Bootstrapping: As easy as 1-2-4. *J. Appl. Statist.* **27**(4) 509–517.

Yu, C. H. 2003. Resampling methods: Concepts, applications, and justification. Practical assessment. *Res. Evaluation* **8**(19).