

Spreadsheets in Education (eJSiE)

Volume 3, Issue 1

2008

Article 6

Spreadsheet Data Resampling for Monte-Carlo Simulation

Thin Yin Leong*

Wee Leong Lee[†]

*Singapore Management University,

[†]Singapore Management University, wlee@smu.edu.sg

Copyright 2008. All rights reserved. This paper is posted at ePublications@bond.

Spreadsheet Data Resampling for Monte-Carlo Simulation

Thin Yin Leong Dr and Wee Leong Lee Dr

Abstract

The pervasiveness of spreadsheet software resulted in its increased application as a simulation tool for business analysis. Random values generation supporting such evaluations using spreadsheets are simple and yet powerful. However, the typical approach to Monte-Carlo simulations, which is what simulations with stochasticity are called, requires significant amount of time to be spent on data collection and distribution function fitting. In fact, the latter can be overwhelming for undergraduate students to do properly in a short time. Resampling eliminates both the need to fit distributions to the sample data, and to perform the ensuing tests of goodness of fit, where sufficiently large data sets are necessary to achieve satisfactory levels of statistical confidence. In contrast, resampling methods can be used even with small data sets. This not only saves class time required to teach statistical data fitting; by generating random values, students also need not learn to use the more complex distribution function inversion method and can better focus on learning business modeling and analysis.

KEYWORDS: resampling, monte carlo simulation

Spreadsheet data resampling for Monte-Carlo simulation

Thin Yin Leong

School of Information Systems, Singapore Management University

tyleong@smu.edu.sg

Wee Leong Lee

School of Information Systems, Singapore Management University

wlee@smu.edu.sg

Abstract

The pervasiveness of spreadsheets software resulted in its increased application as a simulation tool for business analysis. Random values generation supporting such evaluations using spreadsheets are simple and yet powerful. However, the typical approach to Monte-Carlo simulations, which is what simulations with stochasticity are called, requires significant amount of time to be spent on data collection, data collation, and distribution function fitting. In fact, the latter can be overwhelming for undergraduate students to learn and do properly in a short time. Resampling eliminates both the need to fit distributions to the sample data, and to perform the ensuing tests of goodness-of-fit, where sufficiently large data sets are necessary to achieve satisfactory levels of statistical confidence. In contrast, resampling methods can be used even with small data sets. This not only saves class time required to teach statistical data fitting; by generating random values, students also need not learn to use the more complex inverse distribution function method and can better focus on learning business modeling and analysis.

Keywords: Resampling, Monte-Carlo Simulation, Spreadsheet

1. Introduction

Increasingly spreadsheets have been used as a teaching tool to perform Monte-Carlo simulation due to its ease of use, intuitive environment and user friendliness as compared to learning a simulation application package. Craft [1] used Microsoft Excel™ to teach Monte-Carlo experiments to undergraduates in an econometrics course. He commented that most students have experience with spreadsheets and could easily pick up new modeling techniques. Judge [2] presented an exercise to help students understand the meaning of the sampling distribution of a least squares regression estimator, and the way in which the properties of the sampling distribution reflect the characteristics of the regression model itself. He also highlighted that students should be aware of the limitations of using a spreadsheet package for large scale Monte-Carlo simulations and recognized the benefits of using dedicated statistics and econometrics software tools for more advanced modeling.

The exercises described in this article were one of the many exercises taught over a period of thirteen weeks (three hours each week) to mostly second year undergraduate students in a business modeling course.

2. Exercises

The exercises described in this article illustrate several alternatives to resample raw data. Unlike bootstrapping which actually resamples data as a method of drawing statistical inference about a data set, we resample to generate representative data as inputs for Monte-Carlo simulations. From past teaching experience, we observed that students find it easier and quicker to apply data collected directly into simulation exercises without spending too much effort pre-processing the raw data. The pre-processing work includes tabulating the data into frequency tables, plotting histogram or cumulative graphs, selecting and fitting suitable theoretical distribution functions, and performing goodness-of-fit tests. Class time can be spent more effectively on experimenting and analyzing different business scenarios rather than overcoming complicated statistical goodness-of-fit concepts and using inverse distribution functions to simulate the data. The three types of methods deliberately demonstrated in the exercises, all contained within an Excel workbook, are: 1) resampling from frequency bins, 2) resampling from discrete raw data, and 3) resampling from continuous raw data with interpolation.

2.1 Resampling from frequency bins

In this exercise, we demonstrate how resampling can be applied using a frequency table with simple Excel functions. With some minor adjustments, this is easily executed using the LOOKUP function. In Excel, the binRange column array in a frequency table contains the upper limits of data value intervals while the LOOKUP function needs the lower limits of these intervals as input. For the LOOKUP function to perform properly, the cumulative relative frequency (CumRF) column array referred to has to be off-shifted up by one row as shown in Figure 1.

Step 1: Compute the Cumulative Relative Frequency (CumRF)

CumRF is given by $I_i = \text{SUM}(E\$8:E_i) / I\5 for the i^{th} row

where $E\$8:E_i$ is the sum of frequencies up to the i^{th} bin, and

$I\$5 = \text{SUM}(E8:E12)$ is total of all the frequency counts.

Step 2: Add a random variable

Create a random function at $I16 = \text{RAND}()$.

Apply the same function to range $I17:I25$.

Step 3: Compute the number of children encountered in the 10 households

The number of children (resampledValue) can be computed in a formula using a LOOKUP function as follows:

$J16 = \text{LOOKUP}(I16, I\$8:I\$12, H\$9:H\$13)$

where I16 has the random value generated by the RAND function,
 I\$8:I\$12 is the CumRF range as the lookupArray, and
 H\$9:H\$13 is the number of children as the corresponding valueArray.

The LOOKUP function seeks out the largest value in the lookupArray that is less than or equal to the random value generated by the RAND function and returns the resampledValue in the corresponding relative position in the value array. The function RAND generates a random real number greater or equal to zero and less than one.

This method is not as easy to teach as envisaged, because of the need to do the off-shifting of cumulative relative frequency column array and other complications to be explained, although many textbooks on business analysis using spreadsheets apply it. The approach is however often applied incorrectly when the data values are discrete values with ranges larger than a handful. The frequency table for variables with larger ranges, which in itself requires some effort to put up, uses intervals instead of single values in the bins in order to summarize and reduce the data into interval categories. This is so that intervals will have some frequency counts, instead of being just a series of zeros and ones. Students must be careful to use the interval mid-values, not the bin values, as the valueArray as they tend to do. With the interval binning, there is also some degree of information loss. (Leong [3] provided further details on this.) The method of resampling from frequency bins is shown to the class for completeness rather than as a useful and practical general approach. . This method of resampling is equivalent to sampling of the discrete variable from the sample's inverse empirical distribution.

2.2 Resampling from discrete raw data

In student projects, the quantity of data points collected is often small due to the limited time available to complete the assignment. The inverse distribution function method may not be appropriate for small data sets since curve fitting to establish the right distribution as discussed earlier would not be possible. To resample discrete data with small ranges in Excel, a simple way is to make use of the SMALL function incorporated with the RANDBETWEEN function as shown in Figure 2. Hurley [3] employed a similar approach with the SMALL function but used $(\text{INT}(\text{COUNT}(\text{data_range}) * \text{RAND}()) + 1)$ in place of RANDBETWEEN function for the same discrete random number generation purpose.

This exercise involves computing the number of bowls of noodles consumed by a person based on a collected sample data set. It is important to note that in Microsoft Windows™ versions of Excel the RANDBETWEEN is only available after the AnalysisToolpak is activated. To do this, select "Tools" and then "Add-ins" from the main menu and check the "Analysis Toolpak" option. Excel running on Apple Mac personal computers may not require any special activation.

Step 1: Compute the probable number of bowls of noodles consumed in a large party based on a sample collected from a small gathering

Number of bowls of noodles is given by

$J8=SMALL(\$B\$8:\$F\$12,RANDBETWEEN(1,\$J\$6))$

where $\$B\$8:\$F\12 is the valueRange of the sample data set,

$\$J\6 contains the count of the values in the sample data set,

$RANDBETWEEN(1,\$J\$6)$ generates a random integer between 1 and sample data set number count, and

$SMALL(\text{valueRange}, k)$ returns the k^{th} smallest value in the valueRange.

Step 2: Replicate the formula in cell J8 to all cells in range J8: S27

The SMALL function implicitly sorts values in the valueRange $\$B\$8:\$F\12 in ascending order while the RANDBETWEEN function returns a random integer within the range specified by its two arguments. The SMALL function uses this integer value to select the resampling position in the array of (ascending order) sorted sample data values. This method of resampling is also equivalent to sampling of the discrete variable from the sample's inverse empirical distribution.

2.3 Resampling from continuous raw data with interpolation

For the two resampling methods discussed in the above sections, the possible outcome of the resampled data set should be values represented in the raw data set. As a consequence, to get reasonable resampled values, a rather large data sample with many of the data values represented would be required if the data population has a wide range. In a nutshell, the two approaches described previously are suitable for discrete data sets that do not require any data value that lies in between any pair of adjacent sample data points. To overcome this deficiency, an alternate formulation proposed by Leong [3] is applied here as shown in Figure 3, modified from the approach applicable for resampling discrete data. Leong [3] provided broader in-depth discussion of resampling approaches, covering even resampling with interpolation for the dependent multiple variable case.

Our exercise in this section involves computing the weight of pasta consumed by a person based on a collected data set.

Step 1: Compute the probable weight of pasta consumed in a large party based on a sample collected from a small gathering

Weight of pasta is given by

$I8=PERCENTILE(\$B\$8:\$F\$12,RAND())$

where $\$B\$8:\$F\12 is the valueRange of the existing data set.

Step 2: Replicate the formula in cell I8 to all cells in range I8: R27

In the same manner, this formula samples the inverse empirical distribution from the data set. Unlike the SMALL function where the distribution form has staircase steps, the PERCENTILE function generates a distribution which is piecewise linear. That is, it effectively joins the consecutive points in the data set in a linear fashion resulting in a continuous albeit crinkly 'curve'. As such, this formulation is able to associate random variable values from the distribution to any random values generated by the RAND function, including those absent from the sample data set, by implicitly performing linear interpolations between adjacent sample data points.

Despite the simplicity of the formulation, a considerable amount of computation to do the sorting and interpolation work is effectively performed by the PERCENTILE function. A consistency check of the data interpolation is shown in Figure 4 which shows a scatter plot of the empirical data overlapped with data points generated from resampling.

By adding a ROUND function, the formulation can be easily adapted to resample a large discrete data set as follows:

```
I8=ROUND(PERCENTILE($B$8:$F$12,RAND()),0)
```

The ROUND function translates the interpolated values (generated by the PERCENTILE function by taking two consecutive points) to the nearest data point that is a member of the data set. This approach is correct and valid for resampling discrete data with distribution over a close interval, except possibly over-representing the data points at the two extreme ends if these ends are finite.

3. Comments

Students' responses to the exercises are dramatically different to those we got when we in our earlier offerings of the course applied the inverse distribution function method. To begin with, students are really appreciative of not having to deal with complex theoretical distributions. On the contrary, they were amazed at how effortlessly they could generate new data values that mimic the behavioral pattern of the raw data set and in doing so inject more realism to their simulation models. After overcoming initial difficulties, we witnessed students showing greater enthusiasm in applying Monte-Carlo simulations to help them solve business problems.

Students are beginning to experiment with different what-if scenarios and are learning to truly appreciate the potential of using spreadsheets to perform Monte-Carlo simulations, something that they previously did not think possible. However, students still lack experience using stochastic simulations to know when it is not necessary to use Monte-Carlo simulations. This may be due to the examples used to demonstrate resampling. In finding the bowls of noodles or amount of pasta needed, the desired

results may be computed for the required “no-shortage” probability by just applying the PERCENTILE function on the data samples. However, this may confuse the student on the use of this function as a resampler or statistical confidence limit finder. Of course, the use of such simple examples is a compromise as using more complex examples would take the focus away from the key lesson of resampling.

4. Conclusion

We have been teaching a course on spreadsheets business modeling to undergraduates, most of whom are second-year students from the business, accounting, economics, information systems, and social sciences majors. A few thousand students have taken the course to date. In this course, amongst many other topics, Monte-Carlo simulation is taught at an introductory level over two weeks (i.e., six hours of class time).

From our collective past experiences, we discovered that undergraduates generally consider it challenging to do distribution function fitting of the raw data and subsequent application of the inverse distribution formulation, which results in more class time than necessary spent to get students up to speed on this topic. To overcome this problem, we introduced the resampling approaches described in this paper and have been impressed with how quickly students picked up resampling techniques to generate input data for their Monte-Carlo simulations. The three resampling methods presented here utilize only standard Excel functions, without additional add-ins, which can reduce the stress on students having to familiarize themselves with new software add-ins. Due to the ease of generating new data sets from sampled data, students can quickly move on to building more sophisticated spreadsheets simulation models and applying them to solve a variety of business problems. Using resampling techniques, even students who are weak in statistics can construct Monte-Carlo simulation models for this class.

References

- [1] Craft, R.K. (2003) “Using Spreadsheets to Conduct Monte Carlo Experiments for Teaching Introductory Econometrics,” *Southern Economic Journal* 69(3):726-735.
- [2] Judge, G. (1999) “Simple Monte-Carlo Studies on a Spreadsheet,” *Computers in Higher Education Economics Review (CHEER)* 13(2).
- [3] Leong, T-Y (2007) “Monte-Carlo Spreadsheet Simulation using Resampling,” *INFORMS Transactions on Education* 7(3).
- [4] Hurley, W.J. (2000, September/October) “Resampling Calculations in a Spreadsheet,” *Decision Line* 31(5):9-10.

RESAMPLING

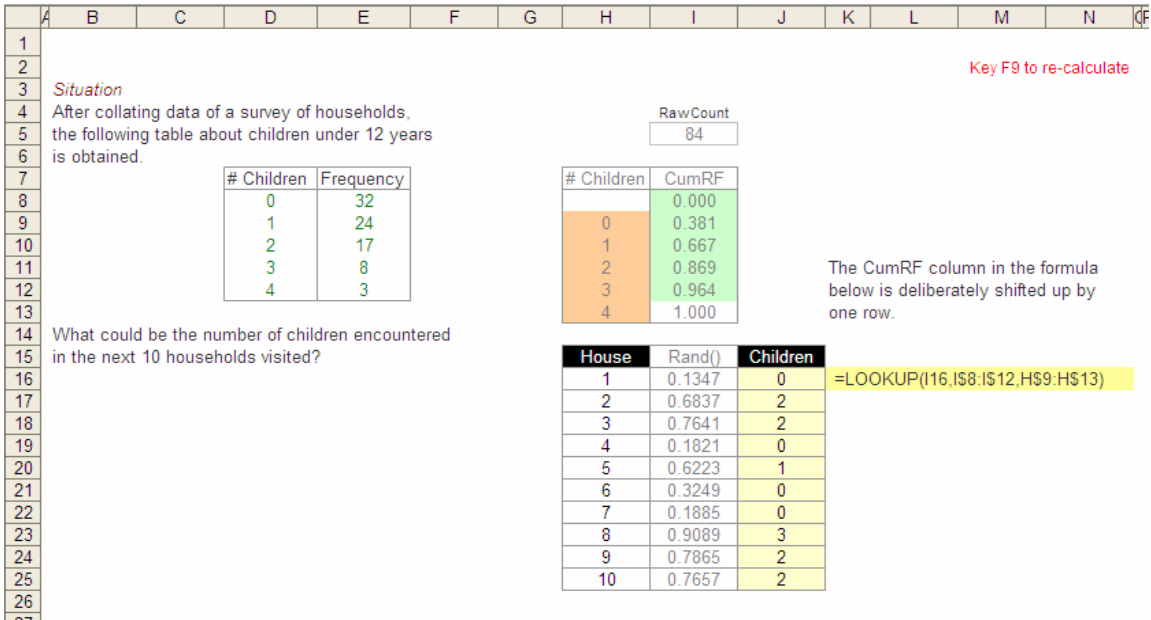


Figure 1: Simulating data from frequency bins

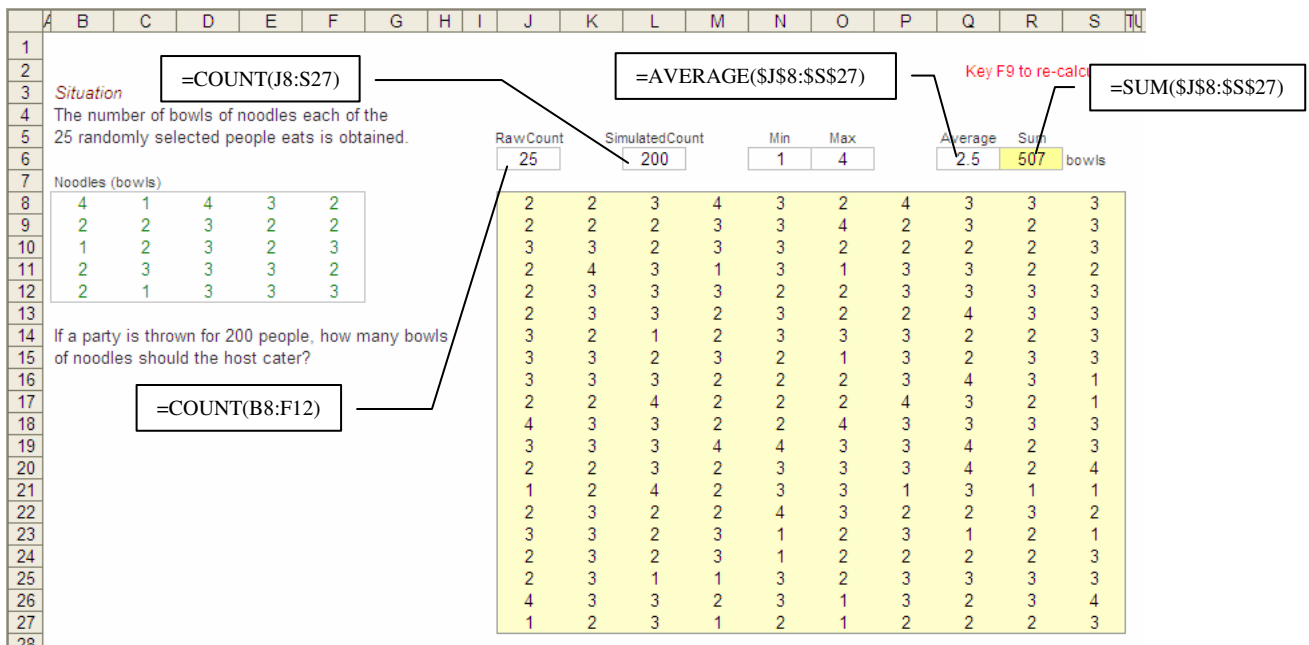


Figure 2: Simulating data by resampling (discrete) raw data

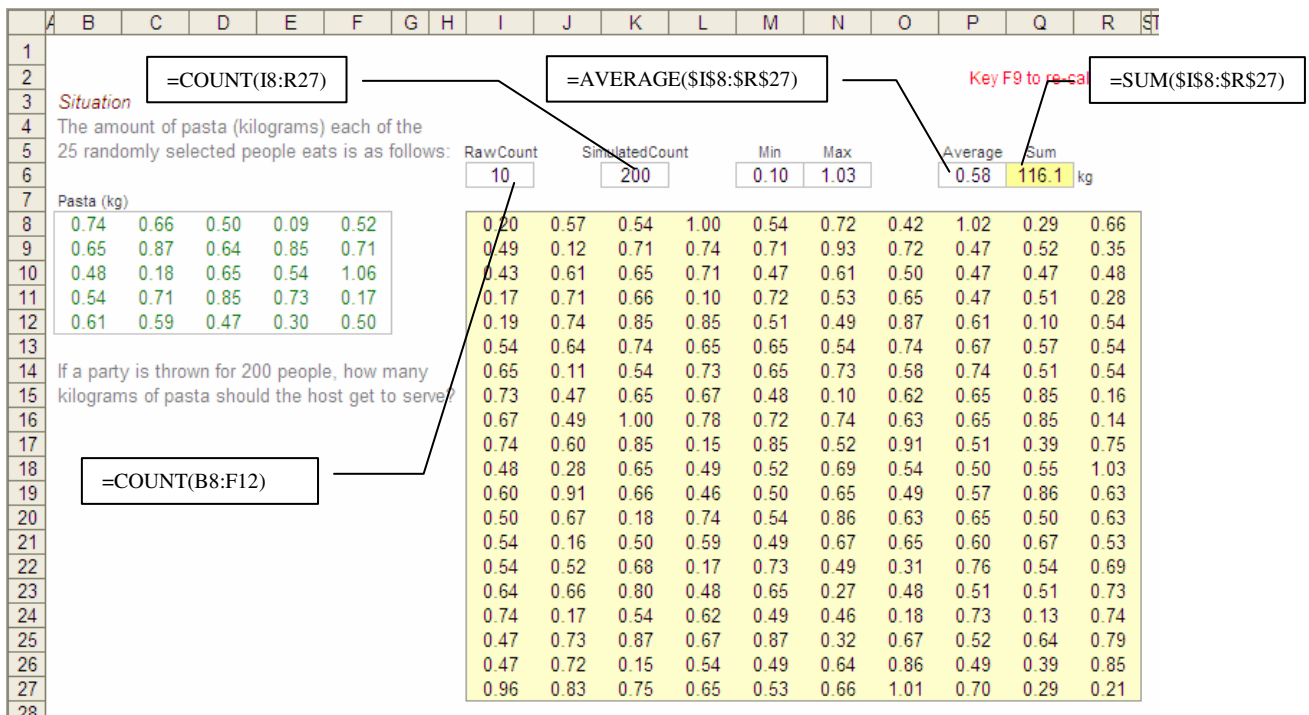


Figure 3: Simulating data by resampling (continuous) raw data

RESAMPLING

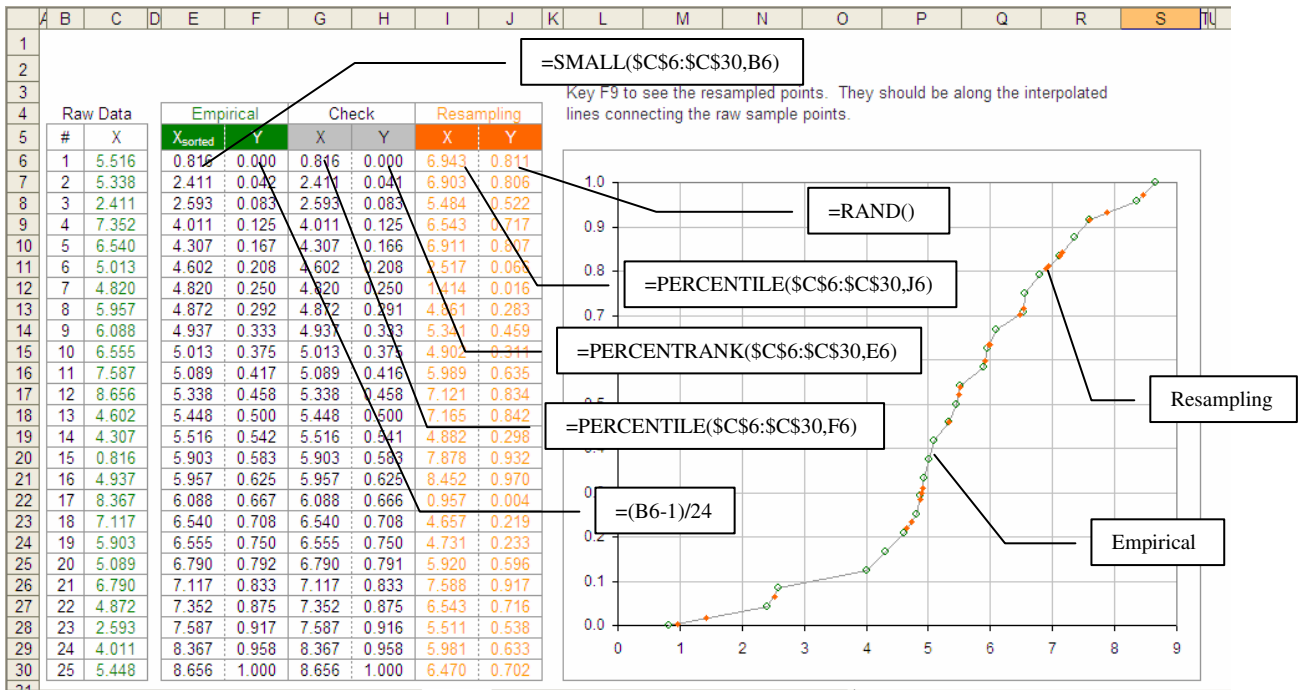


Figure 4: Resampling (continuous) raw data check