Chapter 1 Introduction

1.1 Introduction

As the title of this book clearly indicates, the purpose of this book is to provide a practical guide for using the SAS System to conduct Monte Carlo simulation studies to solve many practical problems encountered in different disciplines. The book is intended for quantitative researchers from a variety of disciplines (e.g., education, psychology, sociology, political science, business and finance, marketing research) who use the SAS System as their major tool for data analysis and quantitative research. With this audience in mind, we assume that the reader is familiar with SAS and can read and understand SAS code.

Although a variety of quantitative techniques will be used and discussed as examples of conducting Monte Carlo simulation through the use of the SAS System, quantitative techniques per se are not intended to be the focus of this book. It is assumed that readers have a good grasp of the relevant quantitative techniques discussed in an example such that their focus will not be on the quantitative techniques, but on how the quantitative techniques can be implemented in a simulation situation.

Many of the quantitative techniques used as examples in this book are those that investigate linear relationships among variables. Linear relationships are the focus of many widely used quantitative techniques in a variety of disciplines, such as education, psychology, sociology, business and finance, agriculture, etc. One important characteristic of these techniques is that they are all fundamentally based on the least-squares principle, which minimizes the sum of residual squares. Some examples of these widely used quantitative methods are regression analysis, univariate and multivariate analysis of variance, discriminant analysis, canonical correlation analysis, and covariance structure analysis (i.e., structural equation modeling).

Before we begin our detailed discussion about how to use the SAS System to conduct Monte Carlo studies, we would like to take some time to discuss briefly a few more general but relevant topics. More specifically, we want to discuss the following:

- What is a Monte Carlo study?
- Why are Monte Carlo studies often necessary?
- What are some typical situations where Monte Carlo simulation is needed?
- Why use the SAS System for conducting Monte Carlo studies?

1.2 What Is a Monte Carlo Study?

What is a Monte Carlo study? According to Webster's dictionary, Monte Carlo relates to or involves "the use of random sampling techniques and often the use of computer simulation to obtain approximate solutions to mathematical or physical problems especially in terms of a range of values each of which has a calculated probability of being the solution" (Merriam-Webster, Inc., 1994, pp. 754-755). This definition provides a concise and accurate description for Monte Carlo studies. For those who are not familiar with Monte Carlo studies, a simple example below will give you a good sense of what a Monte Carlo study is.

1.2.1 Simulating the Rolling of Two Dice

Suppose that we are interested in knowing what the chances are of obtaining two as the sum from rolling a dice twice (assuming a fair dice, of course). There are basically three ways of obtaining an answer to our question. The first is to do it the hard way, and you literally roll a dice twice tens of thousands of times so that you could reasonably estimate the chances of obtaining two as the sum of rolling a dice twice.

Another way of estimating the chance for this event (i.e., obtaining two as the sum from rolling a fair dice twice) is to rely on theoretical probability theory. If you do that, you will reason as follows: to obtain a sum of two from rolling a fair dice twice necessarily means you obtain one in each roll. The probability of obtaining one from rolling the dice once is 1/6 (0.167). The probability of obtaining one from another rolling of the same dice is also 1/6. Because each roll of the dice is independent of another, according to probability theory, the joint probability of obtaining one from both rolls is the product of two—that is, $0.167 \times 0.167 \approx 0.028$. In other words, the chances of obtaining the sum of two from rolling a fair dice twice should be slightly less than 3 out of 100, a not very likely event. In the same vein, the chances of obtaining the sum of 12 from rolling a fair dice twice can also be calculated to be about 0.028. Although it is relatively easy to calculate the theoretical probability of obtaining two as the sum from rolling a fair dice twice, it is more cumbersome to figure out the probability of obtaining, say, seven as the sum from rolling the dice twice, because you have to consider multiple events (6+1, 5+2, 4+3, 3+4, 2+5, 1+6) that will sum up to be seven. Because each of these six events has the probability of 0.028 to occur, the probability of obtaining the sum of seven from rolling a dice twice is $6 \times 0.028 = 0.168$.

Instead of relying on actually rolling a dice tens of thousands of times, or on probability theory, we can also take an empirical approach to obtain the answer to the question without actually rolling a dice. This approach entails a Monte Carlo simulation (MCS) in which the outcomes of rolling a dice twice are *simulated*, rather than actually rolling a dice twice. This approach is only possible with a

computer and some appropriate software, such as SAS. The following (Program 1.1) is an annotated SAS program that conducts an MCS to simulate the chances of obtaining a certain sum from rolling a dice twice.

Program 1.1 *Simulating the Rolling of Two Dice*

```
                    *** simulate the rolling of two dice and the distribution;
                    *** of the sum of the two outcomes;

DATA DICE(KEEP=SUM) OUTCOMES(KEEP=OUTCOME);
     DO ROLL=1 TO 10000;                    *** roll the two dice 10,000 times.;
        OUTCOME1=1+INT(6*RANUNI(123));      *** outcome from rolling the first dice;
        OUTCOME2=1+INT(6*RANUNI(123));      *** outcome from rolling the second dice;
        SUM=OUTCOME1+OUTCOME2;              *** sum up the two outcomes.;
        OUTPUT DICE;                        *** save the sum.;
        OUTCOME=OUTCOME1; OUTPUT OUTCOMES; *** save the first outcome.;
        OUTCOME=OUTCOME2; OUTPUT OUTCOMES; *** save the second outcome.;
     END;
RUN;

PROC FREQ DATA=DICE;          *** obtain the distribution of the sum.;
     TABLE SUM;
RUN;

PROC FREQ DATA=OUTCOMES;     *** check the uniformity of the outcomes.;
     TABLE OUTCOME;
RUN;
```

Output 1.1a presents part of the results (the sum of rolling a dice twice) obtained from executing the program above. Notice that the chances of obtaining two as the sum from rolling a dice twice (2.99%) is very close to what was calculated according to probability theory (0.028). In the same vein, the probability of obtaining the sum of 7 is almost identical to that based on probability theory (16.85% from MCS versus 0.168 based on probability theory).

Output 1.1b presents the estimated chances of obtaining an outcome from rolling a dice once. Note that the chances of obtaining 1 though 6 are basically equal from each roll of the dice, as theoretically expected if the dice is fair.

Output 1.1a *Chances of Obtaining a Sum from Rolling a Dice Twice*

SUM	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2	299	2.99	299	2.99
3	534	5.34	833	8.33
4	811	8.11	1644	16.44
5	1177	11.77	2821	28.21
6	1374	13.74	4195	41.95
7	1685	16.85	5880	58.50
8	1361	13.61	7241	72.41
9	1083	10.83	8324	83.24
10	852	8.52	9176	91.76
11	540	5.40	9716	97.10
12	284	2.84	10000	100.00

Output 1.1b *Chances of Obtaining an Outcome from Rolling a Dice Once*

OUTCOME	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	3298	16.49	3298	16.49
2	3367	16.84	6665	33.33
3	3362	16.81	10027	50.14
4	3372	16.86	13399	67.00
5	3341	16.71	16740	83.70
6	3260	16.30	20000	100.00

Some readers may have some trouble understanding all the elements in the program presented in Program 1.1. We elaborate on the details of the program in later sections. The basic idea of this program is to use a computer to simulate the process of rolling two dice independently, and then sum up the outcomes of the two dice. After 10,000 replications (each consisting of rolling two dice), we obtain 10,000 sums, each of which is based on rolling two dice. By using the SAS FREQ procedure, we obtain the percentage associated with each sum (2 through 12), and this percentage represents the chance of obtaining a specific sum from rolling two dice.

As implied from the above, Monte Carlo simulation offers researchers an alternative to the theoretical approach. There are many situations where the theoretical approach is difficult to implement, much less to find an exact solution. An empirical alternative like the one above is possible because of technological developments in the area of computing. As a matter of fact, with computing power becoming increasingly cheap and with powerful computers more widely available than ever, this computing-intensive approach is becoming more popular with quantitative researchers. In a nutshell, MCS simulates the sampling process from a defined population repeatedly by using a computer instead of actually drawing multiple samples (i.e., in this context, actually rolling dice) to estimate the sampling distributions of the events of interest. As we will discuss momentarily, this approach can be applied to a variety of situations in different disciplines.

1.3 Why Is Monte Carlo Simulation Often Necessary?

After going over the example provided in the previous section, some readers may ask the question: Why is MCS needed or necessary? After all, we already have probability theory which allows us to figure out the chances of any outcome as the sum from rolling a dice twice, and using probability theory is relatively efficient, obviously more so than writing the SAS program presented in Program 1.1. For the situation discussed above, it is true that using probability theory will be more efficient than using the MCS approach to provide the answer to our question. But please keep in mind that the example provided in Program 1.1 is for illustration purposes only, and there are many situations where MCS is needed, or where MCS is the only viable approach to providing analytic solutions to some quantitative research questions.

Although statistical theories are efficient, the validity of any statistical theory is typically contingent upon some theoretical assumptions. When the assumptions of a theory are met by the data that we have in hand, the statistical theory provides us with valid and efficient estimates of sampling distribution characteristics for a statistic of our interest. On the other hand, when the assumptions of a theory are violated in the data that we have, the validity of the estimates about certain sampling distribution characteristics based on the theory is often compromised and uncertain; consequently, we are often at a loss about how much we can trust the theoretical estimates, or about how erroneous our conclusion might be if we blindly rely on the theory, even if some crucial assumptions of the theory

have been violated. It is in these kind of analytic situations that MCS becomes very useful to quantitative researchers, because this approach relies on *empirical* estimation of sampling distribution characteristics, rather than on *theoretical* expectations of those characteristics. With a large number of replications, the empirical results should asymptotically approach the theoretical results, and this can be demonstrated when the theoretical results can be obtained.

In addition to the situations discussed above in which the assumptions of statistical theories may not be met by the data we have at hand, and where consequently, MCS becomes an empirical alternative to theoretical approach, there are some other situations where statistical theories are either so weak that they can not be fully relied upon, or statistical theories simply do not exist. In these situations, MCS may be the only viable approach to providing answers to a variety of questions quantitative researchers may have.

Such situations abound. For example, the distributional characteristics of sample means are well known (e.g., unbiased, with mean equal to μ and standard deviation equal to $\sigma/\sqrt{N}$). But how about the distributional characteristics of sample medians? Is a sample median an unbiased estimate? What is the expected standard deviation of a distribution of sample medians? Does the central limit theorem, which is so important for the distribution of sample means, apply to the distribution of sample medians? These and other similar questions may not be answered from statistical theory, because it is an area where theory is weak or nonexistent. As a result, these questions may need to be answered empirically by conducting MCS, and the distributional characteristics of sample medians can be examined empirically, rather than theoretically based on statistical theory.

1.4 What Are Some Typical Situations Where a Monte Carlo Study is Needed?

As the brief discussion in the previous section indicates, for quantitative researchers in a variety of disciplines, there are two typical situations in which MCS may be called for: when theoretical assumptions of a statistical theory may not hold; and when statistical theory is either weak or nonexistent. In this section, we will discuss some typical situations in which MCS becomes relevant or necessary.

1.4.1 Assessing the Consequences of Assumption Violations

As is well known, statistical techniques can generally be classified into two broad categories: parametric and non-parametric. Most popular statistical techniques belong to the category of parametric statistics. A common characteristic for all parametric statistics is that there are certain assumptions about the distribution of the data. If the assumptions are violated, the validity of the results derived from applying these techniques may be in question. However, statistical theory itself does not usually provide any indication about what, if any, the consequences are, and how serious the consequences will be. If a quantitative researcher wonders about these questions, MCS becomes, in many situations, the only viable approach to obtaining answers to these questions.

For example, for the very popular statistical technique of analysis of variance (ANOVA), which is designed to test the hypothesis of equal means on the dependent variable from two or more groups, a fundamental assumption for the validity of the probability statement from ANOVA is that the groups involved come from populations with equal population variances on the variable of interest (homogeneity of variance assumption). What happens if, in reality, the populations that the groups are from do not have equal population variances on the variable of interest? To what extent is the probability statement from ANOVA invalid? How robust is the ANOVA technique in relation to the violation of this equal variance assumption?

To answer these and other similar questions, we may want to design a MC study in which we intentionally manipulate the variances of different population groups, draw samples from these populations, and apply ANOVA to test the hypothesis that the groups have equal means. Over repeated replications, we will be able to derive an empirical distribution of any sample statistic of our interest. Based on these distributions, we will be able to provide some answers to the questions that cannot be addressed by the statistical theory. Researchers have long used MCS to examine these issues related to ANOVA. (For a very early review, see Glass, Peckham, & Sanders 1972.)

For many popular statistical techniques, data normality is an important assumption. For example, for regression analysis, which is used in almost all disciplines, the tests for regression model parameters, both for the overall regression model fitted to the sample data and for the individual regression coefficients, it is assumed that the data are normally distributed. What are the consequences if the data are not normally distributed as assumed? How extreme should the non-normality condition be before we discount the regression analysis results as invalid? These are only a few of the potential questions quantitative researchers may ask. As discussed before, the answers to these questions may be provided by MCS, because statistical theory only stipulates what the condition should be, and it does not provide a clear indication of what the reality would be if the conditions were not met by the data.

1.4.2 Determining the Sampling Distribution of a Statistic That Has No Theoretical Distribution

In some situations, due to the complexity of a particular statistic, a theoretical sampling distribution of the statistic may not be available. In such situations, if one is interested in understanding how the statistic will vary from sample to sample, i.e., the sampling distribution of the statistic, MCS becomes one viable and realistic approach to obtaining such information.

For example, discriminant analysis and canonical correlation analysis are two multivariate statistical techniques widely used in different disciplines. In both of these techniques, there are (discriminant and canonical) function coefficients which are analogous to regression coefficients in regression analysis, and also, there are (discriminant and canonical) structure coefficients which are the correlations between the measured variables and the (discriminant and canonical) functions. Because of the complexity of these statistics, theoretical distributions are not available for these coefficients (both function and structure coefficients). In the case of discriminant or canonical correlation analysis, there has been a lot of debate about which type of coefficients, function or structure, is more stable across samples (Stevens 1996). Because theoretical sampling distributions are not available for these two type of coefficients, it is not possible to answer the question from any theoretical perspective. Faced with this lack of theoretical sampling distributions, Thompson (1991) conducted a Monte Carlo

study in which the sampling distributions of these two types of coefficients were empirically generated, and based on these empirical sampling distributions, this issue was empirically investigated.

The same situation exists for exploratory factor analysis, a popular statistical technique widely used in psychometrics and in social and behavioral science research in general. In factor analysis, factor pattern coefficients play an important role. Unfortunately, the theoretical sampling distributions of factor pattern coefficients are not available. The lack of theoretical sampling distributions for factor pattern coefficients makes it difficult to assess the importance of a variable in relation to a factor. In practice, such assessment often relies on half guess work and half common sense. It is often suggested that factor pattern coefficients smaller than 0.30 be discounted. Ideally, such an assessment should be made by taking into consideration the sampling variability of the factor pattern coefficient. If one wants to get some idea about the sampling variability of such factor pattern coefficients, in the absence of the theoretical sampling distribution, MCS becomes probably the only viable approach. Quantitative researchers have utilized MCS to investigate this issue in factor analysis. (For examples, see Stevens 1996, pp. 370-371.)

In the past two decades, covariance structure analysis, more commonly known as structural equation modeling (SEM), has become a popular analytic tool for quantitative researchers. In SEM analysis, a group of descriptive model fit indices have been developed to supplement the model fit information provided by the χ^2 test, or to compensate for the widely perceived limitations of the χ^2 test in SEM, that is, it is heavily influenced by the sample size used in testing the model fit (Fan & Wang, 1998). These descriptive fit indices, however, have unknown theoretical sampling distributions, so it is not clear how these fit indices will vary from sample to sample. Again, MCS becomes the primary tool for providing the information about the variability of these fit indices, and many researchers have used this approach in their research (e.g., Fan, Thompson, & Wang 1999; Fan & Wang 1998; Marsh, Balla, & Hau 1996).

1.5 Why Use the SAS System for Conducting Monte Carlo Studies?

As discussed above, Monte Carlo simulation has been an important research area for quantitative researchers in a variety of disciplines. Because MCS is computation-intensive, it is obvious that MCS research typically requires programming capabilities. Furthermore, because many MC studies involve some type of statistical techniques and/or mathematical functions, statistical/mathematical capabilities are also essential. The SAS System has the combination of a powerful variety of built-in statistical procedures (e.g., in SAS/STAT and SAS/ETS software), mathematical functions, and the versatile programming capabilities (in base SAS, the SAS Macro Facility, and SAS/IML software). This combination makes the SAS System ideal for conducting Monte Carlo simulation research, especially research related to statistical techniques. Such a combination of built-in statistical procedures and versatile programming capabilities makes it much more convenient for MCS researchers to get their job done. Without such a combination of statistical capabilities and programming capabilities within the same system, an MCS researcher may have to deal with different systems, and consequently worry about the interface among different systems.

For example, some MCS researchers use the Fortran language for programming their Monte Carlo simulations. Because there are no built-in statistical procedures, any statistical analysis will either have to be programmed by the researchers themselves (a formidable task if one is dealing with a

complicated quantitative technique), or some other system has to be used for the purpose (e.g., IMSL: International Mathematical & Statistical Libraries, a package of mathematical routines). In the latter case, the interface between different programs in the programming process may become cumbersome and difficult.

By relying on the SAS System for statistical simulation, almost all statistical procedures are already built in, and statistical analysis results are easily obtained either through the built-in statistical procedures, or through programming using the powerful interactive matrix language (PROC IML) under the SAS System. In either case, both the statistical computation and programming are highly integrated within the same system, which considerably simplifies the tasks of Monte Carlo researchers. In addition, the SAS System offers great flexibility in data generation, data transformation, obtaining and saving simulation results, etc. The completeness and the flexibility of the SAS System have convinced us that currently no other system makes Monte Carlo research, especially research involving statistical techniques, easier and more efficient than the SAS System does.

1.6 About the Organization of This Book

This book has nine chapters. The first two chapters provide an overview of the Monte Carlo research process. Starting with the third chapter, we lead the readers through a step-by-step process of conducting a Monte Carlo simulation. The third chapter discusses data generation by using different random number generators that are available in base SAS. This chapter lays the foundation for Chapter 4, which focuses on generating multiple variables that are correlated and that have different population characteristics (e.g., variables that deviate from the theoretical normal distribution to different degrees). As a matter of fact, data generation is so crucial that it is no exaggeration to say that the success of Monte Carlo simulation research hinges on the correct data generation process.

Once readers understand the data generation process in Monte Carlo simulation research, the next chapter, Chapter 5, discusses an important programming aspect of a Monte Carlo study: automation of the simulation process. Because a Monte Carlo study usually involves a large number (e.g., thousands, or hundreds of thousands) of replications (i.e., repeatedly drawing samples from a specified statistical population, and obtaining and analyzing the sample statistic of interest), unless the process can be automated through programming, MCS would be almost impossible to do in practice. Chapter 5 provides a detailed practical guide for automating the MCS process in SAS.

Chapter 6 and Chapter 7 present some Monte Carlo simulation examples involving both univariate and multivariate statistical techniques widely used by researchers in different fields. The examples in these two chapters integrate what has been discussed up to Chapter 5. Quantitative researchers who are interested in conducting Monte Carlo simulation involving statistical techniques will find these two chapters very useful and practical. For each of the examples used, a problem is presented, and the rationale for conducting a Monte Carlo simulation study is provided. Then, the SAS program and explanatory comments are presented step by step. Finally, some selected results of the simulation are presented. Thus, each example provides a complete examination of a Monte Carlo study.

In Chapter 8, our focus shifts a little, and we discuss Monte Carlo simulation examples related to the financial industry. As the examples in this chapter clearly indicate, the issues addressed by Monte Carlo simulation tend to be quite different from those in Chapters 6 and 7. For this reason, we present these examples from the financial industry in this separate chapter. Lastly, Chapter 9 provides discussion about implementing a Monte Carlo simulation study using techniques that involve SAS/ETS software. Examples related to time series analysis are presented in Chapter 9 as well.

Combined, the chapters in this book provide a systematic and practical guide to conducting Monte Carlo simulation studies in SAS. In our presentation of the examples, if a quantitative technique is involved, the quantitative technique per se is not our focus; instead, we focus more on the programming aspects of the Monte Carlo study, and the quantitative technique is presented as an example. Because of this, we provide little elaboration on the mathematical or statistical aspects of the quantitative techniques used as examples, and we assume that readers who are interested in the quantitative techniques will consult other relevant sources.

1.7 References

Fan, X., B. Thompson, and L. Wang. 1999. "The Effects of Sample Size, Estimation Methods, and Model Specification on SEM Fit Indices." *Structural Equation Modeling: A Multidisciplinary Journal* 6:56-83.

Fan, X., and L. Wang. 1998. "Effects of Potential Confounding Factors on Fit Indices and Parameter Estimates for True and Misspecified SEM Models." *Educational and Psychological Measurement* 58:699-733.

Glass, G. V., P. D. Peckham, and J. R. Sanders. 1972. "Consequences of Failure to Meet Assumptions Underlying the Fixed-Effects Analysis of Variance and Covariance." *Review of Educational Research* 42:237-288.

Marsh, H. W., J. R. Balla, and K. T. Hau. 1996. "An Evaluation of Incremental Fit Indices: A Clarification of Mathematical and Empirical Properties." In *Advanced Structural Equation Modeling: Issues and Techniques,* ed. G. A. Marcoulides and R. E. Schumacker, 315-353. Mahwah, NJ: Lawrence Erlbaum Associates.

Merriam-Webster, Inc. 1994. *Merriam-Webster's Collegiate Dictionary.* 10th ed. Springfield, MA: Merriam-Webster, Inc.

Stevens, J. 1996. *Applied Multivariate Statistics for the Social Sciences*. 3d ed. Mahwah, NJ: Lawrence Erlbaum Associates.

Thompson, B. 1991. "Invariance of Multivariate Results: A Monte Carlo Study of Canonical Function and Structure Coefficients." *Journal of Experimental Education* 59:367-382.