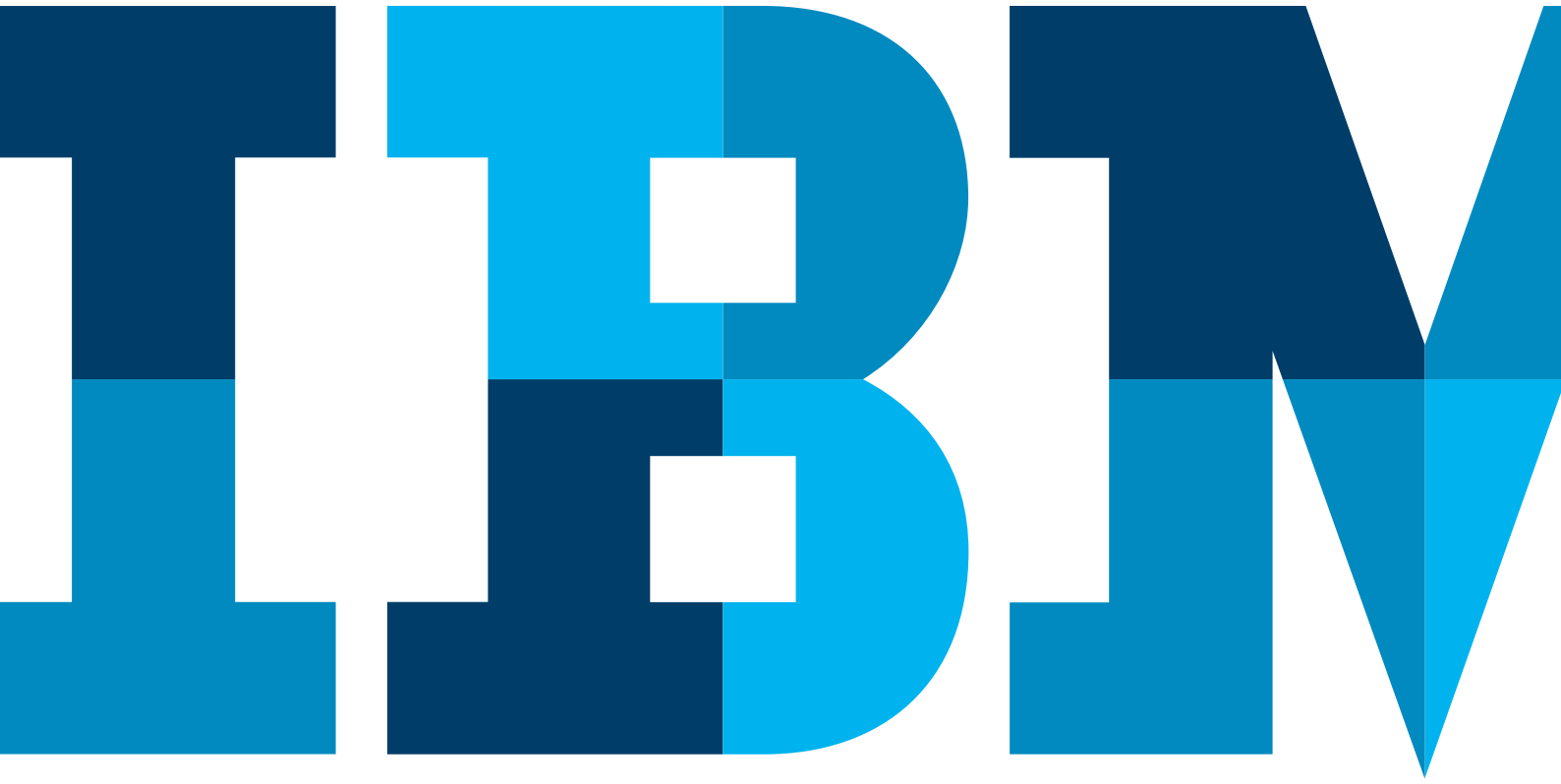
Better decision making under uncertain conditions using Monte Carlo Simulation

Monte Carlo simulation and risk analysis techniques in IBM SPSS Statistics



Contents

Introduction

IBM® SPSS® Statistics is one of the world's leading statistical software solutions, providing predictive models and advanced analytics to help solve business and research problems. For many businesses, research institutions and statisticians, it is the de facto standard for statistical analysis. Organizations use SPSS Statistics to:

- Understand data.
- Analyze trends.
- Forecast and plan.
- Validate assumptions.
- Drive accurate conclusions.

The SPSS Statistics environment offers a wide range of multivariate procedures for investigating complex data relationships. A number of procedures include advanced models such as general and generalized linear modeling capabilities. With general linear models, you can model relationships and interactions between many factors. The general linear model incorporates a number of different statistical models: analysis of variance (ANOVA), multivariate analysis of variance (MANOVA), analysis of covariance (ANCOVA), repeated measures and more. General linear modeling is well suited for statisticians who analyze data with unique characteristics (for example, nested-structure data) or who describe relationships between a dependent and a set of independent variables to discover whether random effects introduce correlations between cases.

Regression models (for continuous dependent variables) are a family of classical predictive techniques, all of which involve fitting (or regressing) a line or curve to a series of observations to model effects or predict outcomes. With SPSS Statistics, you can also use regression models to predict categorical outcomes for more than two categories, easily classify data into two groups, accurately model non-linear relationships, find the best predictor from dozens of possibilities and more.

Predictive models, such as linear regression, require a set of known inputs to predict an outcome. In many real-world applications, however, inputs are not known with certainty, and users are interested in accounting for that uncertainty in their models. For example, when given a profit model that includes cost of materials as an input, users also want to account for uncertainty in materials cost and determine the likelihood that profit will fall below a target value. To deal with uncertainty in future input values, statisticians turn to simulation.

SPSS Statistics includes a simulation module designed to account for uncertainty in inputs to predictive models. This paper describes Monte Carlo simulation, the value of using Monte Carlo simulation for risk analysis and how SPSS Statistics and its new simulation module can help businesses use Monte Carlo simulation for risk analysis.

Uncertainty in inputs into models

The development of a forecasting model requires making certain assumptions. These might be assumptions about the investment return on a portfolio, the cost of a construction project or how long it will take to complete a certain task. Because these are projections, it is not possible to know with certainty what the actual value will be, but based on historical data, expertise in the field or past experience, it is possible to estimate. Although this estimate is useful for developing a model, it contains some inherent uncertainty and risk, because the estimate is an unknown value.

Traditionally, companies address uncertainty in one of three ways:

- *Point estimates* use the most likely values for the uncertain variables. These estimates are the easiest, but they can return very misleading results. For example, consider what might happen if you decide to cross a river because its average depth is three feet. Or, suppose, because you have calculated that it takes an average of 25 minutes to get to the airport, you leave 25 minutes before your flight takes off. How likely are you to miss it?
- *Range estimates* typically calculate three scenarios: the best case, the worst case and the most likely case. These types of estimates can show you the range of outcomes, but not the probability of any of these outcomes. This approach also considers only a few discrete outcomes, ignoring hundreds of thousands of others. Simply put, it gives equal weight to each outcome, so there is no attempt to assess the likelihood of each outcome. Interdependence between inputs and the impact of different inputs relative to the outcome are ignored, oversimplifying the model and reducing its accuracy.
- *What-if scenarios* are usually based on the range estimates and typically are about exploring the effect of things you can control. What is the worst case? What if sales are best case but expenses are the worst case? What if sales are average, but expenses are the best case? What if sales are average, expenses are average, but sales for the next month are flat? This form of analysis can be time consuming, and it results in a great deal of data.

With each of the three traditional approaches, it is not possible to determine the probability of achieving different outcomes.

Addressing uncertainty with simulation

A Monte Carlo simulation is a computer experiment involving random sampling from probability distributions of the inputs to obtain approximate solutions to problems, especially in the case of a range of values where each has a calculated probability of being the solution. When statisticians use the term "simulation," they usually mean Monte Carlo simulation.

In this approach, uncertain inputs are modeled with probability distributions (such as the triangular distribution), and simulated values for those inputs are generated by drawing from those distributions. The simulated values are then
used in a predictive model to generate an outcome. The process is repeated many times (typically thousands or tens of thousands of times), resulting in a distribution of outcomes that can be used to answer questions of a probabilistic nature to determine behavior, to analyze risk and more.

One of the strengths of Monte Carlo simulation is that it makes it possible to account for risk in quantitative analysis and decision making. Historical simulation (usually used in computing value at risk in financial scenarios, for example) consists of generating scenarios by sampling historical data associated with each risk factor included in the problem. This approach doesn't require any distributional assumptions.

Running simulations from known distributions based on historical data produces accuracy. For example, stochastic risk analysis uses a model and Monte Carlo simulation to analyze the effect of varying inputs on outputs of the model. It defines probability distributions to express the possible variation of the model input variables and uses the Monte Carlo simulation technique to calculate the effect of uncertainty on the model's key outputs. Stochastic analysis can be an invaluable decision-making tool for investment appraisal, business and strategic planning, marketing and sales forecasting, pricing models, along with many scientific applications.

If historical data exists, you still might not want to use historical simulation because Monte Carlo simulation has the advantage of allowing for a wider variety of scenarios than the rather limited results that historical data can provide. Therefore, you can fit probability distributions to the data and use it as the basis for input distributions for Monte Carlo simulation. For example, someone might have collected historical data on a product price and might want to create a distribution of possible future prices that is based on the data.

One of the reasons why risk analysis was not frequently applied in the past is that computers were not powerful enough to handle the demanding needs of Monte Carlo simulation. In addition, for each case, you had to develop a custom project appraisal computer model that represented the relationships between input and output variables using a combination of functions, formulas and data. Now, however, most of the computers and processors can handle intensive computation from Monte Carlo simulation.

The value of Monte Carlo simulation for risk analysis

Existing statistical and modeling software solutions have different methods for addressing risk analysis. Some modeling solutions work in the Microsoft Excel environment. The advantage of this approach is that companies can easily incorporate risk management in everyday processes at all levels of their organization because spreadsheet models have been created for them.

Other statistical software products can act as standalone solutions for risk management, but they are designed mostly for banks and other types of financial institutions. They provide risk management tools for the management and control of market risk, credit risk, operational risk and liquidity risk, but it is not clear how a model is built within their framework.

Monte Carlo simulation solutions either provide functions to simulate from all of the standard probability distributions or offer two-dimensional Monte Carlo simulations. Econometric tools are available for performance and risk analysis, and other software has functions for calculating a risk model. Custom coding is necessary if you want to use these solutions for Monte Carlo simulation and risk analysis together.

Risk analysis can greatly benefit from Monte Carlo simulation. Consider a simple example of monthly sales as a function of advertising expenditures, the number of sales agents and the consumer confidence index (called "cci" in the model), which is an indicator of the general health of the economy. This process has several stages:

1. Set up the model and identify risk variables

Figure 1 demonstrates the data represented over 5 years of monthly sales data that was used to create the linear regression model, assuming that there are no time dynamics in the data and no correlation in the error terms.

advert	cci	agents	sales
11332.00	55.00	109.00	6709410.00
69477.00	93.00	113.00	7829320.00
59087.00	73.00	105.00	6943660.00
36364.00	53.00	118.00	7257860.00
50401.00	42.00	115.00	6481200.00
75382.00	102.00	107.00	7476770.00
75892.00	68.00	122.00	7364630.00
37461.00	73.00	116.00	6980820.00
73628.00	73.00	116.00	6840060.00
22951.00	91.00	98.00	6943570.00
55946.00	60.00	112.00	7149030.00

Figure 1: The data for monthly sales

Figure 2 shows the linear regression model generated after fitting the model with sales as a dependent (target) variable and advert, cci and agents as independent (predictor) variables. In the example, the advert, cci and agent variables are those for which the projected value is both probable and potentially damaging to the monthly sales value.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.711[a]	.505	.479	275285.8758

a. Predictors: (Constant), agents, advert, cci

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.330E+12	3	1.443E+12	19.047	.000[b]
	Residual	4.244E+12	56	75782313427		
	Total	8.574E+12	59			

a. Dependent Variable: sales

b. Predictors: (Constant), agents, advert, cci

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	4627254.938	501272.059		9.231	.000	3623086.337	5631423.538
	advert	6.619	1.890	.331	3.502	.001	2.833	10.405
	cci	10331.400	1705.754	.573	6.057	.000	6914.363	13748.437
	agents	12747.541	4260.719	.284	2.992	.004	4212.295	21282.788

a. Dependent Variable: sales

Figure 2: Linear regression model with three risk variables

2. Specify probability distributions.

The specification of a probability distribution for each selected risk variable involves setting up a range of values and allocating probability weights. In general, analysts for the project should rely on judgment and subjective factors for determining the range of values and probabilities. Figure 3 illustrates some of the probability distributions used in the application of risk analysis.

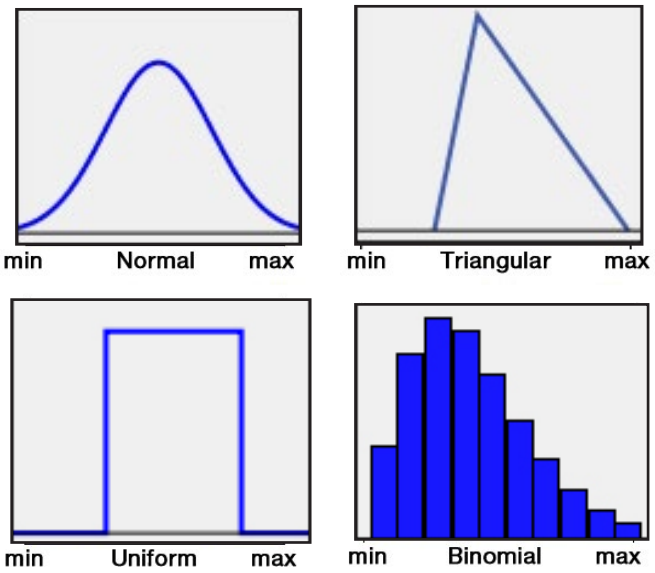


Figure 3: Probability distributions used in risk analysis

If historical data exists, you can skip this stage and conduct historical simulation directly. Or, you can fit distributions to the data for Monte Carlo simulation.

3. Run simulations

The model is processed and repeated until enough results are gathered to make up a representative sample of the near infinite number of possible combinations. During a simulation, the values of the risk variables are generated randomly according to specified probability distributions. In the example, the goal of the simulation was to look at the probability of reaching a monthly target sales goal of $7,000,000. The results of the model (the "monthly sales" in the example) for each run are computed and stored away for statistical analysis (the final stage of the risk analysis process). Figure 4 shows the cumulative distribution function of monthly sales (based on 100,000 records of simulated data) where you can observe that the probability of reaching the monthly target sales goal of $7,000,000 is 62 percent.

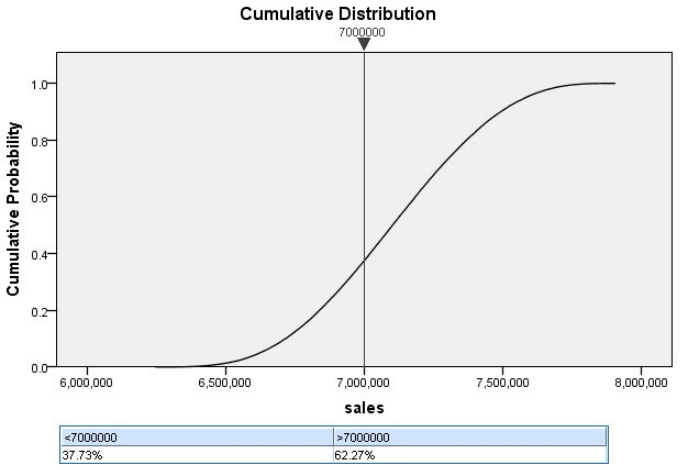
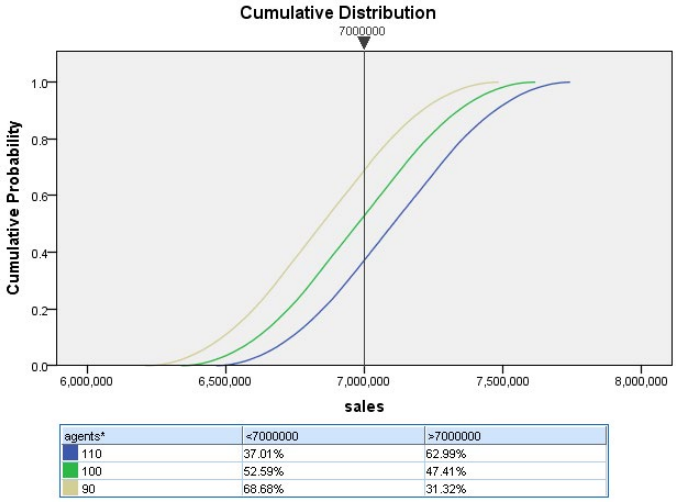


Figure 4: Cumulative distribution function of monthly sales

The example also looked at the effect of varying the number of sales agents over three fixed values: 110, 100 and 90 to analyze the effect on monthly sales of reducing sales staff. Varying an input over a set of fixed values, as done here for number of sales agents, is referred to as "sensitivity analysis." Figure 5 shows the sensitivity analysis of varying the number of sales agents with respect to monthly sales.



agents*	<7000000	>7000000
110	37.01%	62.99%
100	52.59%	47.41%
90	68.68%	31.32%

Figure 5: Sensitivity analysis

You can see that when you reduce your sales staff from 110 to 90, you decrease your chances of meeting your goal from 63 percent to 31 percent.

Next, the company decides to analyze the impact of customer satisfaction on monthly sales. It has obtained monthly satisfaction data from both a formal survey of existing customers and analysis of social media content. This data has been spread across five categories: Strong Negative, Somewhat Negative, Neutral, Somewhat Positive and Strongly Positive.

Specifically, the company wants to determine whether the satisfaction level measured from social media data has an impact on the sales target. An additional step can be taken to confirm this hypothesis.

4. Fitting a categorical distribution

The original sales model can be expanded by including monthly satisfaction data. In the original story, the company wanted to know how likely it was to achieve a monthly sales target of $7,000,000. For inputs with a categorical distribution, you can automatically compute a multiway contingency table from the historical data that describes the associations between those inputs. The contingency table is then used when data are generated for those inputs. After fitting categorical distributions and an associated contingency table to the monthly satisfaction level data, the simulation is rerun to generate scatter plots of the target with the inputs. Scatter plots that contain a categorical target and/or categorical input are displayed as heat maps, as shown in Figure 6.

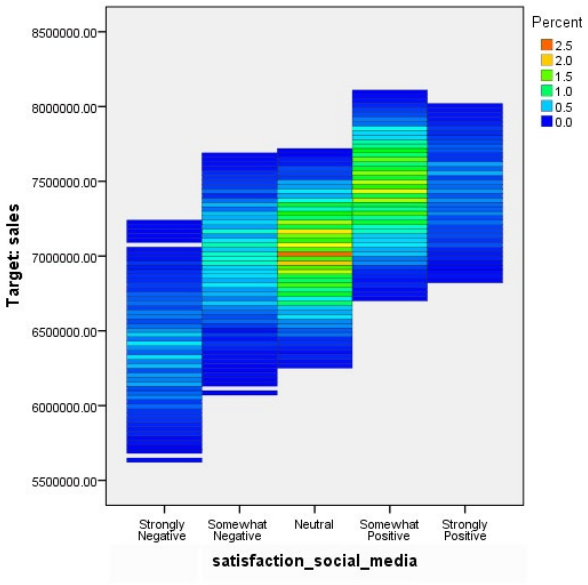


Figure 6: The expanded model that includes satisfaction data shows that when the satisfaction level measured from social media data is "Neutral," the monthly sales numbers are spread roughly evenly around the $7,000,000 target. However, when the social media satisfaction level is "Somewhat Positive" the sales distribution shifts so that the bulk of the distribution (for that satisfaction level) exceeds the target. This suggests that customer satisfaction level as measured by social media data is an integral part of attaining the sales target.

SPSS Statistics: Building better simulations and assessing risk with automation

The simulation functionality in SPSS Statistics is designed to account for the uncertainty of values of the inputs in predictive models. It helps users:

- Design a simulation. Users can specify all details required to run a simulation, such as the distributions for the simulated predictors and correlations for those predictors. When historical data is present, distributions and correlations for simulated predictors can be automatically determined from that data.
- Save specifications for a simulation to a simulation plan file.
- Run a simulation. The specifications for the simulation can come from a loaded simulation plan file or the user can simply provide the specifications in the associated user interface and run the simulation from that user interface.
- Load a simulation plan, modify any aspect of the plan, optionally run the modified simulation and optionally save the changes.

In addition to the procedures described in the previous section, SPSS Statistics includes a range capabilities to help users build better simulations, including:

Simulating strings. Non-numeric variables such as "male" and "female" can be simulated without recoding them as numeric variables. The software also supports fitting a categorical distribution to a string field in the active dataset. For example, if a model contains gender as an input, the user can load the model, fit the model inputs to the active dataset, and SPSS Statistics will fit a categorical distribution for string fields.

Support for Automatic Linear Modeling (ALM). Users can export models from ALM and use them as the starting point for the simulation.

Association between categorical inputs. SPSS Statistics can automatically determine and use associations between categorical inputs when generating data for those inputs. A multiway contingency table is computed for all inputs that are fit to a categorical distribution. This table is then used when generating data for the inputs.

Ability to generate data in the absence of a predictive model. In this case, the user simply specifies which variables to simulate and either fits them to the active dataset or manually specifies their distributions.

How does it work? Here are two examples.

Modeling impact on profit due to uncertainty of cost of materials

Paul is an analyst at a large manufacturing company and is responsible for financial modeling and forecasting. Given the current instability in markets, Paul's managers want him to include risk in the cost of materials in his profit model. For this scenario, many statisticians use a relatively simple profit model: estimates of the maximum, minimum and most likely value for cost of materials. This has the limitations described previously for range estimates.

Using the new simulation dialog in SPSS Statistics, Paul enters his profit model in the expression builder and specifies "cost of materials" (one of the predictors in the model) as a simulated predictor. Paul only has estimates of the maximum, minimum and most likely cost, so he chooses a triangular distribution to model his cost variable. Because there is some uncertainty in his estimates of the maximum, minimum and most likely cost, Paul plans to run his simulation multiple times, varying the parameters of his cost distribution in each simulation. He is able to save his specifications in a simulation plan so that his scenario analysis does not have to be completed in just one session, and he is able to vary the parameters for his triangular distribution more easily after reloading his simulation plan.

Modeling energy usage accounting for uncertainty in temperature

It can be a challenge to model or predict energy usage needs for a utility company that purchases energy from other providers when demand exceeds on-site production capacity. Consider Pamela, who is responsible for that task in her utility company. She models energy needs with a regression model that she has built in SPSS Statistics. Pamela has saved the model to an XML file, which enables her to quickly apply the model to a given set of inputs—something that she does daily.

One of the primary sources of uncertainty in Pamela's model is temperature. Although she can easily use what-if analysis to get point estimates of energy needs for specific temperatures, she is most interested in the likelihood that energy needs will exceed production capacity on a given day, necessitating the purchase of additional energy from an outside provider.

Pamela has access to historical temperature data that she plans to use to model uncertainty in daily temperature, so she starts up SPSS Statistics and loads that data.

She opens the simulation dialog, loads her regression model and specifies that temperature is a simulated predictor. She then chooses to automatically fit (autofit) the temperature distribution from the historical data. Because she also has temperature forecasts for the current day, she can customize the parameters of the autofit distribution if she chooses. She specifies values for the fixed predictors and obtains a distribution of energy usage, from which she can easily determine the probability that energy needs will exceed on-site production capacity by a given amount.

Conclusion

Monte Carlo simulation helps address the challenges of dealing with uncertainty in predictive and forecasting models and assessing risk. SPSS Statistics is designed to help you use Monte Carlo simulation in risk analysis. Using the simulation module in SPSS Statistics, you can simulate data according to parameters you specify and then use that simulated data as input for predicting an outcome. You can also modify the parameters used to simulate the data and compare outcomes. For example, you can simulate various advertising budget amounts and see how that affects total sales. Based on the outcome of the simulation, you might decide to spend more on advertising to meet your total sales goal. With automation, features for saving simulation plans and support for predictive modeling, the simulation module in SPSS Statistics smoothly combines risk analysis and Monte Carlo simulations in one software solution.

About IBM Business Analytics

IBM Business Analytics software delivers data-driven insights that help organizations work smarter and outperform their peers. This comprehensive portfolio includes solutions for business intelligence, predictive analytics and decision management, performance management, and risk management.

Business Analytics solutions enable companies to identify and visualize trends and patterns in areas, such as customer analytics, that can have a profound effect on business performance. They can compare scenarios, anticipate potential threats and opportunities, better plan, budget and forecast resources, balance risks against expected returns and work to meet regulatory requirements. By making analytics widely available, organizations can align tactical and strategic decision-making to achieve business goals. For further information please visit **ibm.com**/business-analytics.

Request a call

To request a call or to ask a question, go to **ibm.com**/business-analytics/contactus. An IBM representative will respond to your inquiry within two business days.

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
May 2012

YTW03231-GBEN-01