Risk Simulation:

The Basics of Quantitative Risk Analysis and Simulation

Short Examples Series

using

Risk Simulator

For more information please visit:
www.realoptionsvaluation.com
or contact us at:
admin@realoptionsvaluation.com

Introduction to Risk Simulator

This section also provides the novice risk analyst an introduction to the *Risk Simulator* software for performing Monte Carlo simulation, where a 30-day trial version of the software is included in the book's DVD. This section starts off by illustrating what Risk Simulator does and what steps are taken in a Monte Carlo simulation as well as some of the more basic elements in a simulation analysis. It then continues with how to interpret the results from a simulation and ends with a discussion of correlating variables in a simulation as well as applying precision and error control. Software versions with new enhancements are released continually. Please review the software's user manual and the software download site (www.realoptionsvaluation.com) for more up-to-date details on using the latest version of the software. See *Modeling Risk: Applying Monte Carlo Simulation, Real Options Analysis, Stochastic Forecasting, and Portfolio Optimization* (Wiley 2007) also by the author, for more technical details on using Risk Simulator.

Risk Simulator is a Monte Carlo simulation, forecasting, and optimization software. It is written in Microsoft .NET C# and functions with Excel as an add-in. This software is compatible and often used with the Real Options SLS software used in Part II of this book, also developed by the author. Standalone software applications in C++ are also available for implementation into other existing proprietary software or databases. The different functions or modules in both software applications are briefly described next. The Appendix provides a more detailed list of all the functions, tools, and models.

- The *Simulation Module* allows you to:
    - Run simulations in your existing Excel-based models
    - Generate and extract simulation forecasts (distributions of results)
    - Perform distributional fitting (automatically finding the best-fitting statistical distribution)
    - Compute correlations (maintain relationships among simulated random variables)
    - Iidentify sensitivities (creating tornado and sensitivity charts)
    - Test statistical hypotheses (finding statistical differences between pairs of forecasts)
    - Run bootstrap simulation (testing the robustness of result statistics)
    - Run custom and nonparametric simulations (simulations using historical data without specifying any distributions or their parameters for forecasting without data or applying expert opinion forecasts)

- The *Forecasting Module* can be used to generate:
  - Automatic time-series forecasts (with and without seasonality and trend)
  - Automatic ARIMA (automatically generate the best-fitting ARIMA forecasts)
  - Basic Econometrics (modified multivariate regression forecasts)
  - Box-Jenkins ARIMA (econometric forecasts)
  - GARCH Models (forecasting and modeling volatility)
  - J-Curves (exponential growth forecasts)
  - Markov Chains (market share and dynamics forecast)
  - Multivariate regressions (modeling linear and nonlinear relationships among variables)
  - Nonlinear extrapolations (curve fitting)
  - S-Curves (logistic growth forecasts)
  - Spline Curves (interpolating and extrapolating missing values)
  - Stochastic processes (random walks, mean-reversions, jump-diffusion, and mixed processes)
- The *Optimization Module* is used for optimizing multiple decision variables subject to constraints to maximize or minimize an objective. It and can be run as a static optimization, as a dynamic optimization under uncertainty together with Monte Carlo simulation, or as a stochastic optimization. The software can handle linear and nonlinear optimizations with integer and continuous variables.
- The *Real Options Super Lattice Solver* (*SLS*) is another standalone software that complements Risk Simulator, used for solving simple to complex real options problems.

To install the software, insert the accompanying CD-ROM, click on the *Install Risk Simulator* link, and follow the onscreen instructions. You will need to be online to download the latest version of the software. The software requires Windows 2000/XP/Vista, administrative privileges, and Microsoft .Net Framework 1.1 installed on the computer. Most new computers come with Microsoft .NET Framework 1.1 already preinstalled. However, if an error message pertaining to requiring .NET Framework 1.1 occurs during the installation of Risk Simulator, exit the installation. Then, install the relevant .NET Framework software also included in the CD (found in the *DOT NET Framework* folder). Complete the .NET installation, restart the computer, and then reinstall the Risk Simulator software. Version 1.1 of the .NET Framework is required even if your system has version 2.0/3.0 as they work independently of each other. You may also download this software on the Download page of www.realoptionsvaluation.com.

Once installation is complete, start Microsoft Excel. If the installation was successful, you should see an additional *Simulation* item on the menu bar in Excel and a new icon bar, as shown in Figure 1. Figure 2 shows the icon toolbar in more detail. You are now ready to start using the software.

Please note that Risk Simulator supports multiple languages (e.g., English, Chinese, Japanese and Spanish) and you can switch among languages by going to **Risk Simulator** | **Languages**.

There is a default 30-day trial license file that comes with the software. To obtain a full corporate license, please contact the author's firm, Real Options Valuation, Inc. at admin@realoptionsvaluation.com. If you are using Windows Vista, make sure to disable User Access Control before installing the software license. To do so: Click on **Start | Control Panel | Classic View** (on the left panel) | **User Accounts** | **Turn User Account Control On or Off** and **uncheck** the option, **Use User Account Control (UAC)**, and restart the computer. When restarting the computer, you will get a message that UAC is turned off. You can turn this message off by going to the **Control Panel | Security Center | Change the Way Security Center Alerts Me | Don't Notify Me and Don't Display the Icon**.

The sections that follow provide step-by-step instructions for using the software. As the software is continually updated and improved, the examples in this book might be slightly different from the latest version downloaded from the Internet.

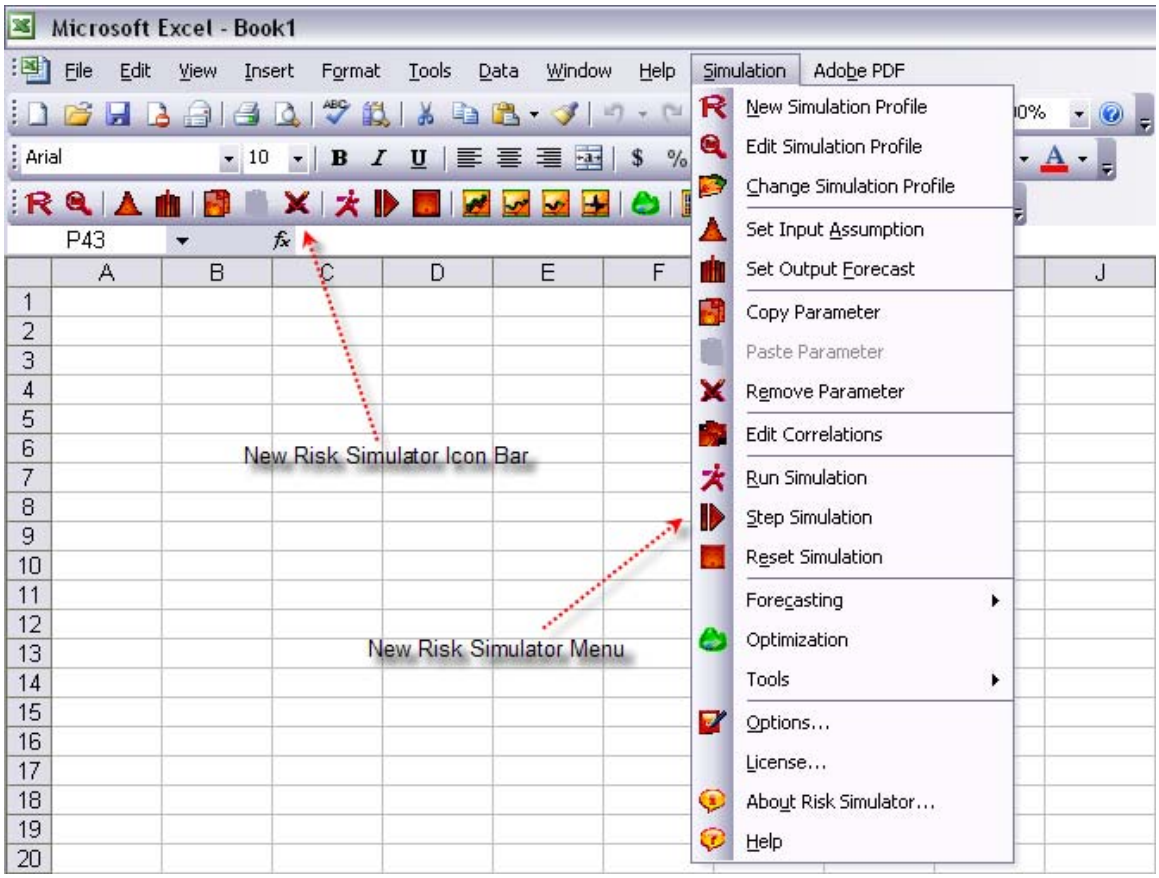


Figure 1A: Risk Simulator menu and icon toolbar in Excel XP and Excel 2003

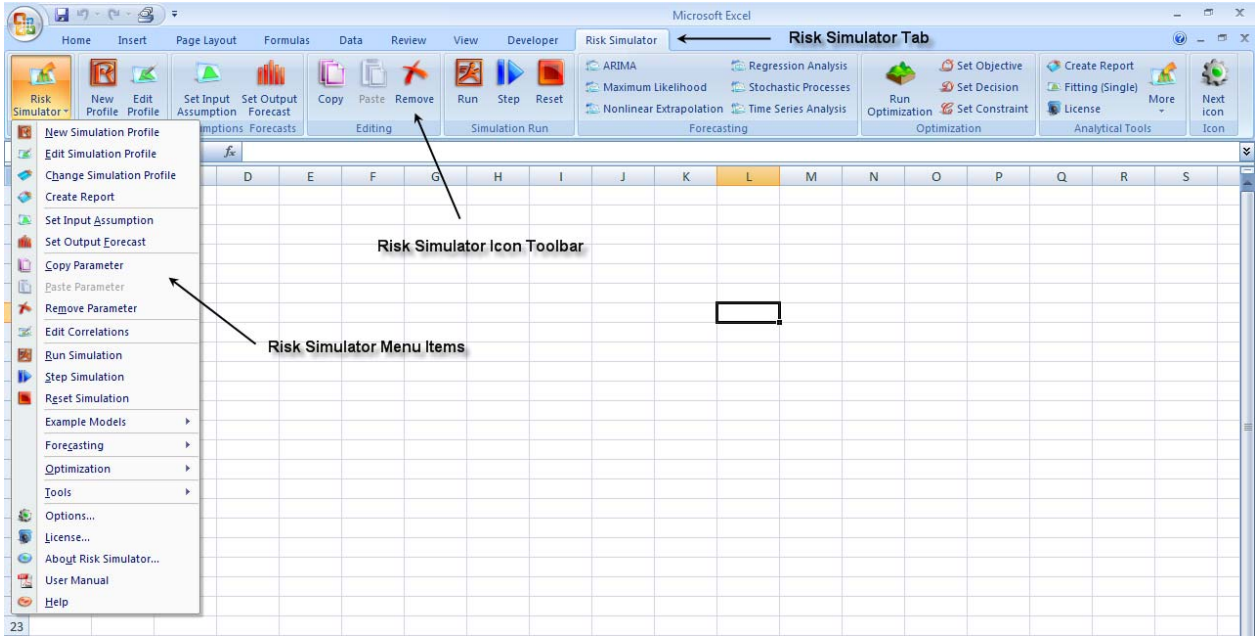


Figure 1B: Risk Simulator menu and icon toolbar in Excel 2007

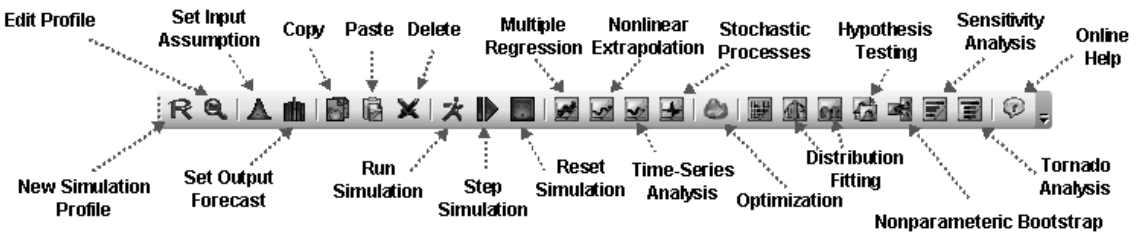


Figure 2A: Risk Simulator icon toolbar in Excel XP and Excel 2003

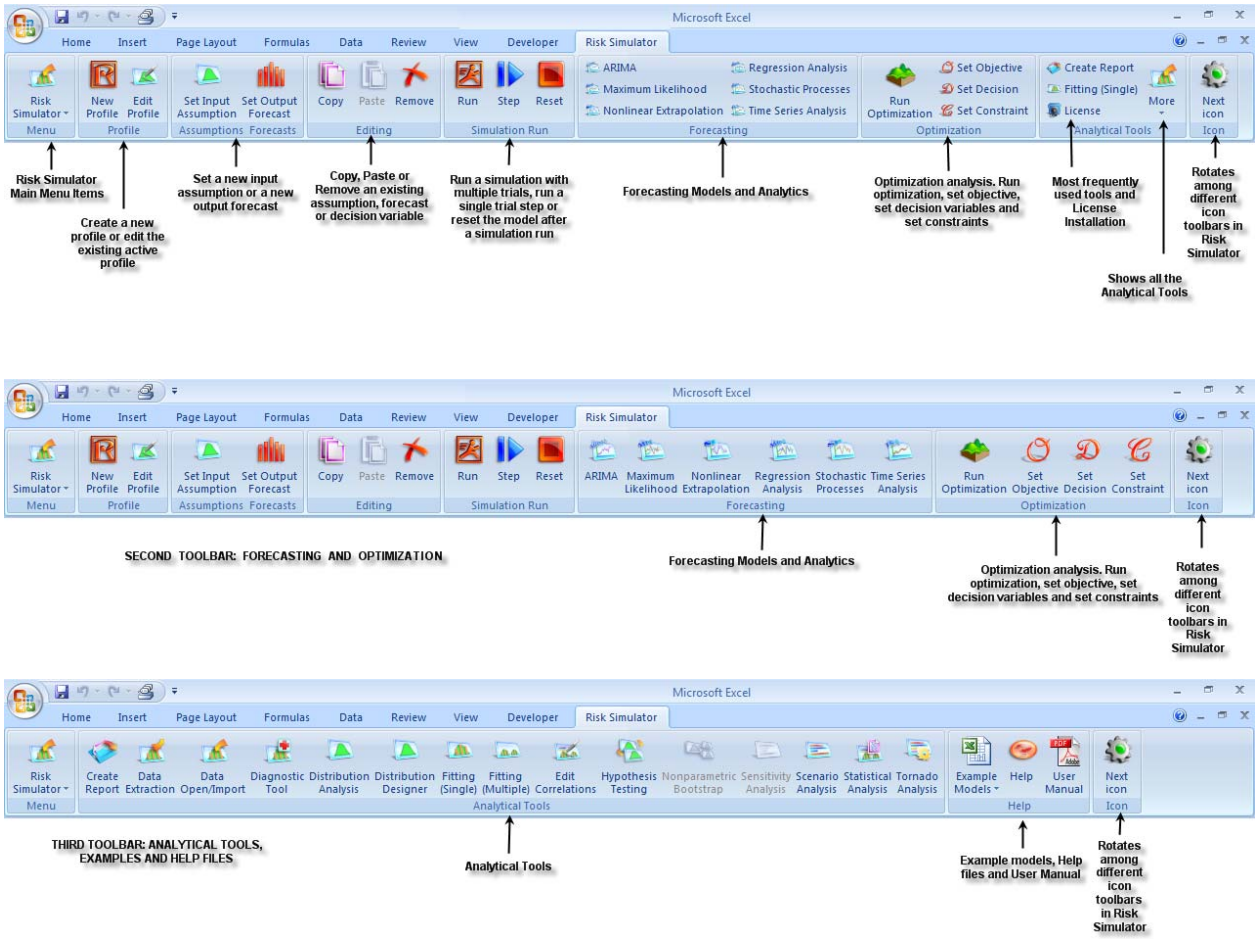


Figure 2B: Risk Simulator icon toolbars in Excel 2007

Running a Monte Carlo Simulation

Typically, to run a simulation in your existing Excel model, you must perform these five steps::

1. **Start a new or open an existing simulation profile**
2. **Define input assumptions in the relevant cells**
3. **Define output forecasts in the relevant cells**
4. **Run the simulation**
5. **Interpret the results**

If desired, and for practice, open the example file called *Basic Simulation Model* and follow along the examples on creating a simulation. The example file can be found on the start menu at **Start | Real Options Valuation | Risk Simulator | Examples.**

1. Starting a New Simulation Profile

To start a new simulation, you must first create a simulation profile. A simulation profile contains a complete set of instructions on how you would like to run a simulation; it contains all the assumptions, forecasts, simulation run preferences, and so forth. Having profiles facilitate creating multiple scenarios of simulations; that is, using the same exact model, several profiles can be created, each with its own specific simulation assumptions, forecasts, properties, and requirements. The same analyst can create different test scenarios using different distributional assumptions and inputs or multiple users can test their own assumptions and inputs on the same model. Instead of having to make duplicates of the model, the same model can be used and different simulations can be run through this model *profiling* process.

Start a new simulation profile by performing these steps:

- **Start Excel** and create a new or open an existing model. You can use the *Basic Simulation Model* example to follow along: **Risk Simulator | Examples | Basic Simulation Model**.
- Click on **Risk Simulator | New Simulation Profile**.
- Enter a title for your simulation including all other pertinent information (Figure 3).

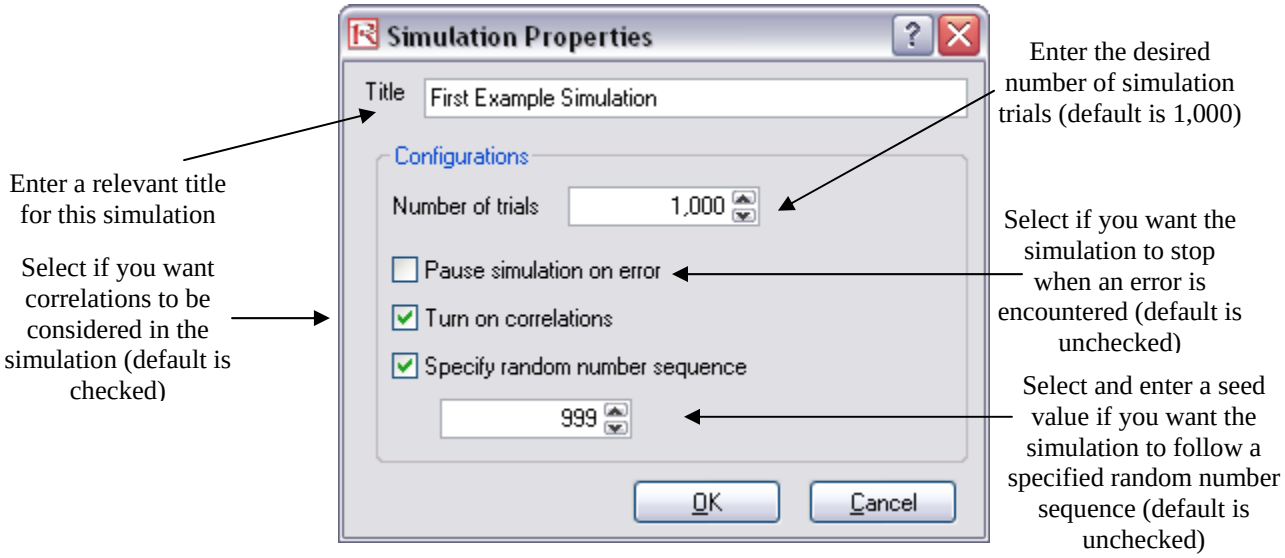


Figure 3: New simulation profile

The elements in the new simulation profile dialog shown in Figure 3 include:

- *Title*: Specifying a simulation profile name or title allows you to create multiple simulation profiles in a single Excel model. By so doing, you can save different simulation scenario profiles within the same model without having to delete existing assumptions and change them each time a new simulation scenario is required.
- *Number of trials*: Enter the number of simulation trials required. Running 1,000 trials means that 1,000 different iterations of outcomes based on the input assumptions will be generated. You can change this number as desired but the input has to be positive integers. The default number of runs is 1,000 trials.
- *Pause on simulation error*: If checked, the simulation stops every time an error is encountered in the Excel model; that is, if your model encounters a computational error (e.g., some input values in a simulation trial may yield a divide-by-zero error in a spreadsheet cell), the simulation stops. This feature is important to help audit your model to make sure there are no computational errors in your Excel model. However, if you are sure the model works, there is no need for you to check this preference.
- *Turn on correlations*: If checked, correlations between paired input assumptions will be computed. Otherwise, correlations will all be set to zero and a simulation is run assuming no cross-correlations between input assumptions. Applying correlations will yield more accurate results if correlations do indeed exist and will tend to yield a lower forecast confidence if negative correlations exist.
- *Specify random number sequence*: By definition, a simulation yields slightly different results every time it is run by virtue of the random number generation routine in Monte Carlo simulation. This is a theoretical fact in all random number generators. However, when making presentations, sometimes you may require the same results. For example, during a live presentation you may like to shown the same results being generated as are in the report; when you are sharing models with others, you also may want the same results to be obtained every time. If that is the case, check this preference and enter in an initial seed number. The seed number can be any positive integer. Using the same initial seed value, the same number of trials, and the same input assumptions always will yield the same sequence of random numbers, guaranteeing the same final set of results.

Note that once a new simulation profile has been created, you can come back later and modify your selections. In order to do this, make sure that the current active profile is the profile you wish to modify, otherwise, click on **Risk Simulator | Change Simulation Profile,** select the profile you wish to change

and click **OK** (Figure 4 shows an example where there are multiple profiles and how to activate, duplicate or delete a selected profile). Then, click on **Risk Simulator | Edit Simulation Profile** and make the required changes.

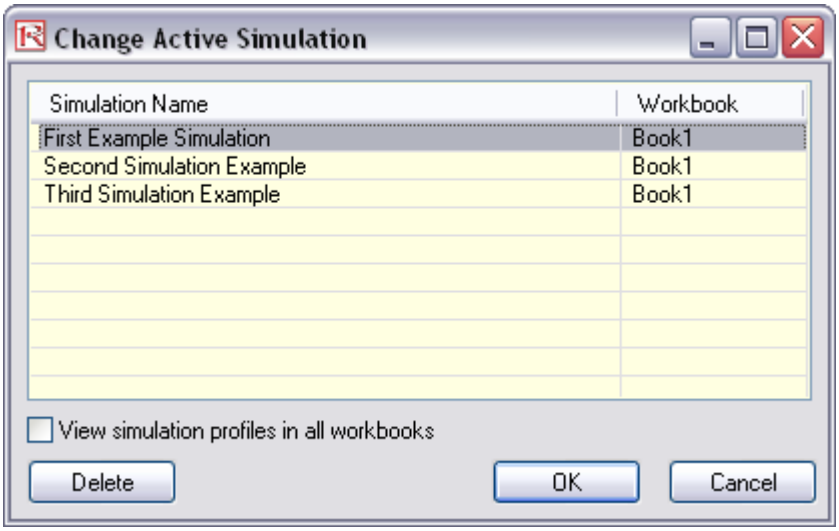


Figure 4: Change active simulation

2. Defining Input Assumptions

The next step is to set input assumptions in your model. Note that assumptions can be assigned only to cells without any equations or functions (i.e., typed-in numerical values that are inputs in a model), whereas output forecasts can be assigned only to cells with equations and functions (i.e., outputs of a model). Recall that assumptions and forecasts cannot be set unless a simulation profile already exists. Follow these steps to set new input assumptions in your model:

- Select the cell you wish to set an assumption on (e.g., cell G8 in the Basic Simulation Model example).
- Click on **Risk Simulator | Set Input Assumption** or click the *Set Assumption* icon in the Risk Simulator icon toolbar.
- Select the relevant distribution you want, enter the relevant distribution parameters, and hit **OK** to insert the input assumption into your model (Figure 5).

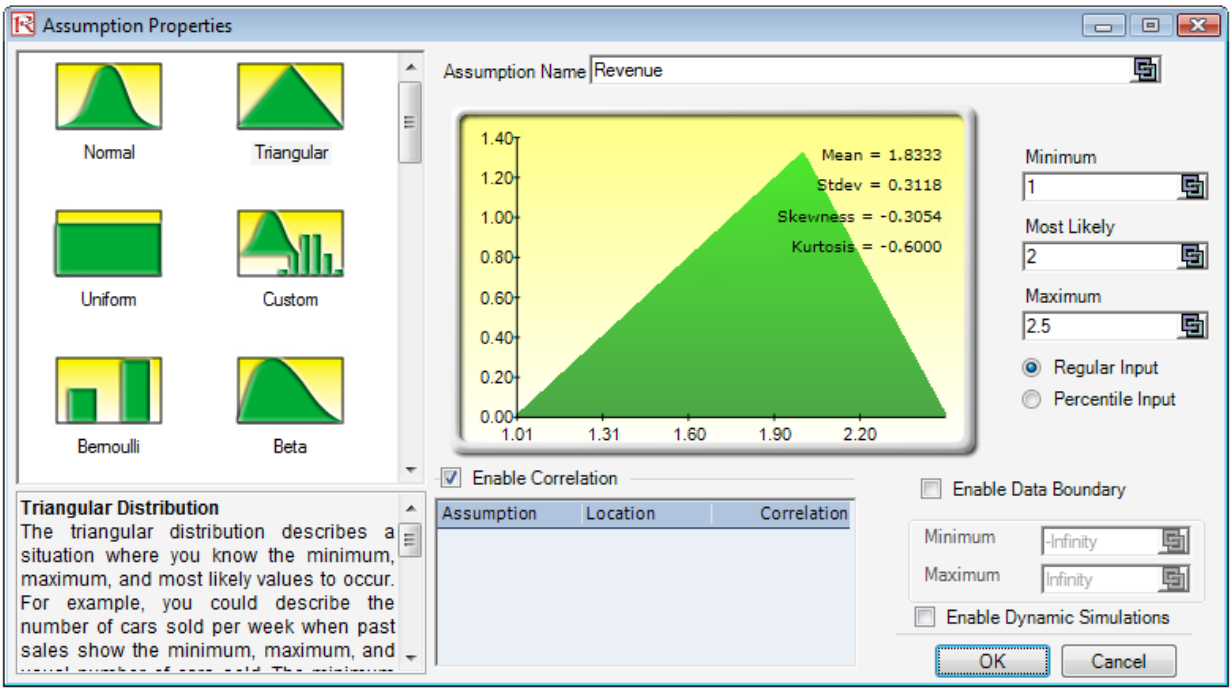


Figure 5: Setting an input assumption

Several key areas are worthy of mention in the Assumption Properties. Figure 6 shows the different areas:

- *Assumption Name:* This optional area allows you to enter in unique names for the assumptions to help track what each of the assumptions represents. Good modeling practice is to use short but precise assumption names.
- *Distribution Gallery:* This area to the left shows all of the different distributions available in the software. To change the views, right click anywhere in the gallery and select large icons, small icons, or list. More than two dozen distributions are available.
- *Input Parameters:* Depending on the distribution selected, the required relevant parameters are shown. You may either enter the parameters directly or link them to specific cells in your worksheet. Click on the link icon to link an input parameter to a worksheet cell. Hard coding or typing the parameters is useful when the assumption parameters are assumed not to change. Linking to worksheet cells is useful when the input parameters themselves need to be visible on the worksheets or can be changed, as in a dynamic simulation (where the input parameters themselves are linked to assumptions in the worksheets creating a multidimensional simulation or simulation of simulations).
- *Data Boundary:* Typically, the average analyst does not use distributional or data boundaries truncation, but they exist for truncating the distributional assumptions. For instance, if a normal distribution is selected, the theoretical boundaries are between negative infinity and positive

infinity. However, in practice, the simulated variable exists only within some smaller range. This range can be entered to truncate the distribution appropriately.

- *Correlations:* Pairwise correlations can be assigned to input assumptions here. If assumptions are required, remember to check the *Turn on Correlations* preference by clicking on *Risk Simulator | Edit Simulation Profile*. See the discussion on correlations later in this chapter for more details about assigning correlations and the effects correlations will have on a model.
- *Short Descriptions:* These exist for each of the distributions in the gallery. The short descriptions explain when a certain distribution is used as well as the input parameter requirements. See the section in the appendix, *Understanding Probability Distributions*, in *Modeling Risk: Applying Monte Carlo Simulation, Real Options Analysis, Stochastic Forecasting, and Portfolio Optimization* (Wiley 2006) also by the author, for details about each distribution type available in the software.

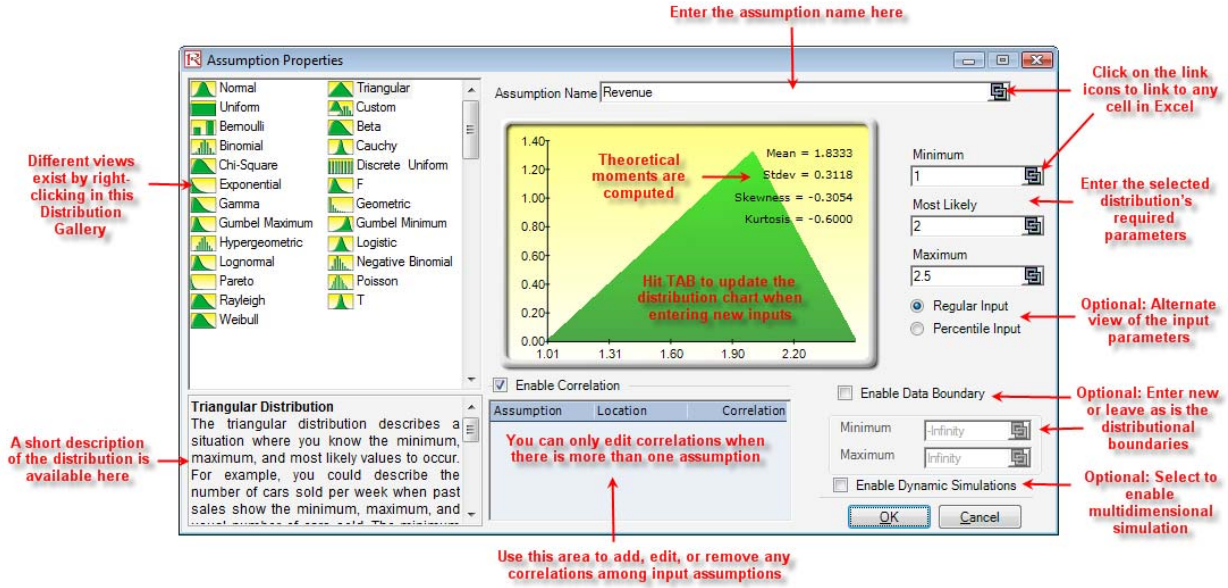


Figure 6: Assumption properties

Note: If you are following along with the example, continue by setting another assumption on cell G9. This time use the Uniform distribution with a minimum value of 0.9 and a maximum value of 1.1. Then, proceed to defining the output forecasts in the next step.

3. Defining Output Forecasts

The next step is to define output forecasts in the model. Forecasts can be defined only on output cells with equations or functions.

Use these steps to define the forecasts:

- Select the cell on which you wish to set an assumption (e.g., cell G10 in the Basic Simulation Model example).
- Click on **Risk Simulator | Set Output Forecast** or click on the set forecast icon on the Risk Simulator icon toolbar.
- Enter the relevant information and click **OK**.

Figure 7 illustrates the set forecast properties, which include:

- *Forecast Name:* Specify the name of the forecast cell. This is important because when you have a large model with multiple forecast cells, naming the forecast cells individually allows you to access the right results quickly. Do not underestimate the importance of this simple step. Good modeling practice is to use short but precise assumption names.
- *Forecast Precision:* Instead of relying on a guesstimate of how many trials to run in your simulation, you can set up precision and error controls. When an error-precision combination has been achieved in the simulation, the simulation will pause and inform you of the precision achieved. Thus the number of simulation trials is an automated process; you do not have to guess the required number of trials to simulate. Review the section on error and precision control for more specific details.
- *Show Forecast Window:* This property allows you to show or not show a particular forecast window. The default is to always show a forecast chart.

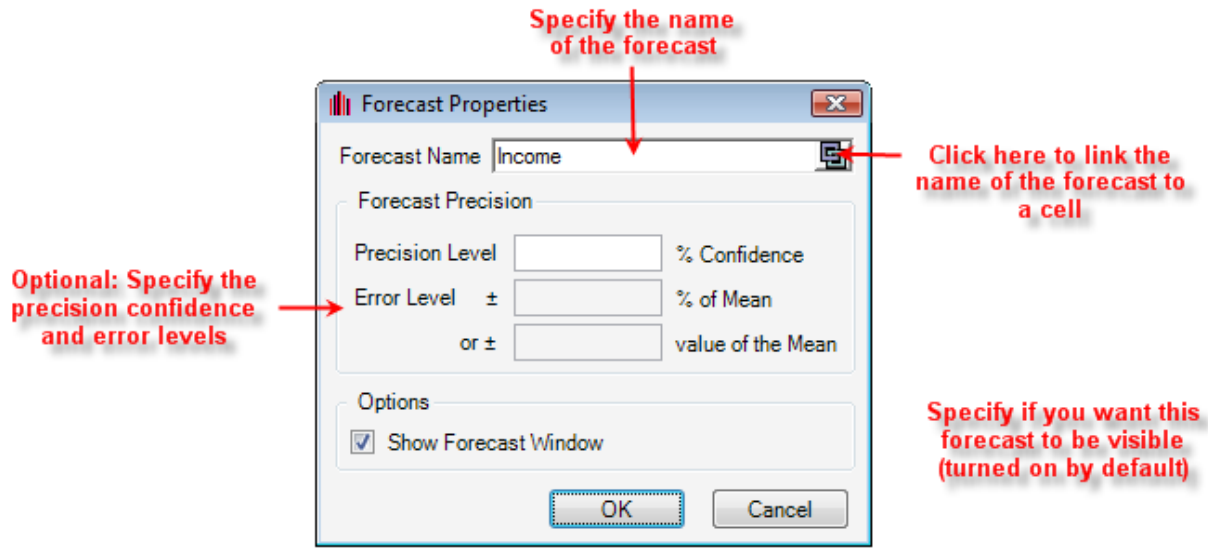


Figure 7: Set output forecast

4. Run Simulation

If everything looks right, click on **Risk Simulator | Run Simulation** or click on the **Run** icon on the Risk Simulator toolbar, and the simulation will proceed. You may also reset a simulation after it has run to rerun it (**Risk Simulator | Reset Simulation** or the *Reset* icon on the toolbar), or to pause it during a run. Also, the step function (**Risk Simulator | Step Simulation** or the *Step* icon on the toolbar) allows you to simulate a single trial, one at a time, which is useful for educating others on simulation (i.e., you can show that at each trial, all values in the assumption cells are replaced and the entire model is recalculated each time).

5. Interpreting the Forecast Results

The final step in Monte Carlo simulation is to interpret the resulting forecast charts. Figures 8 to 15 show the forecast chart and the statistics generated after running the simulation. Typically, these sections on the forecast window are important in interpreting the results of a simulation:

- **Forecast Chart:** The forecast chart shown in Figure 8 is a probability histogram that shows the frequency counts of values occurring and the total number of trials simulated. The vertical bars show the frequency of a particular *x* value occurring out of the total number of trials, while the cumulative frequency (smooth line) shows the total probabilities of all values at and below *x* occurring in the forecast.
- **Forecast Statistics:** The forecast statistics shown in Figure 9 summarizes the distribution of the forecast values in terms of the four moments of a distribution. You can rotate between the histogram and statistics tab by depressing the space bar.

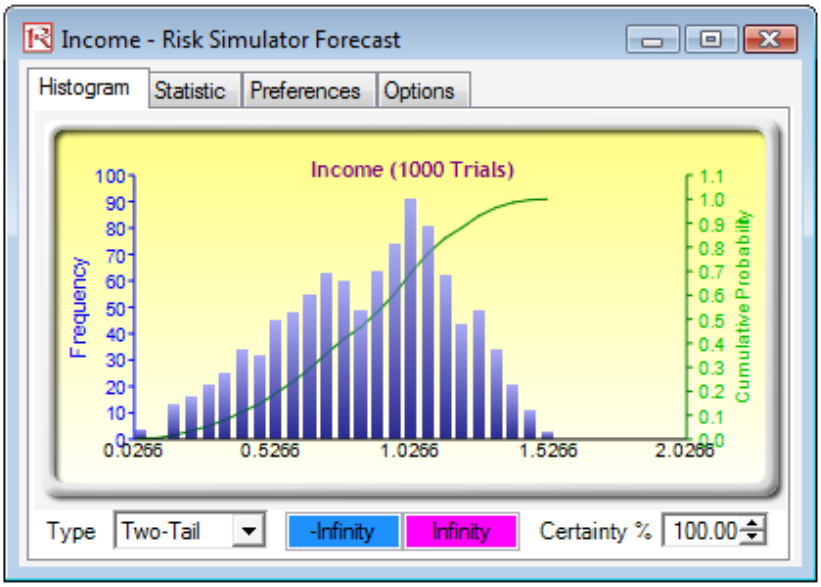


Figure 8: Forecast chart

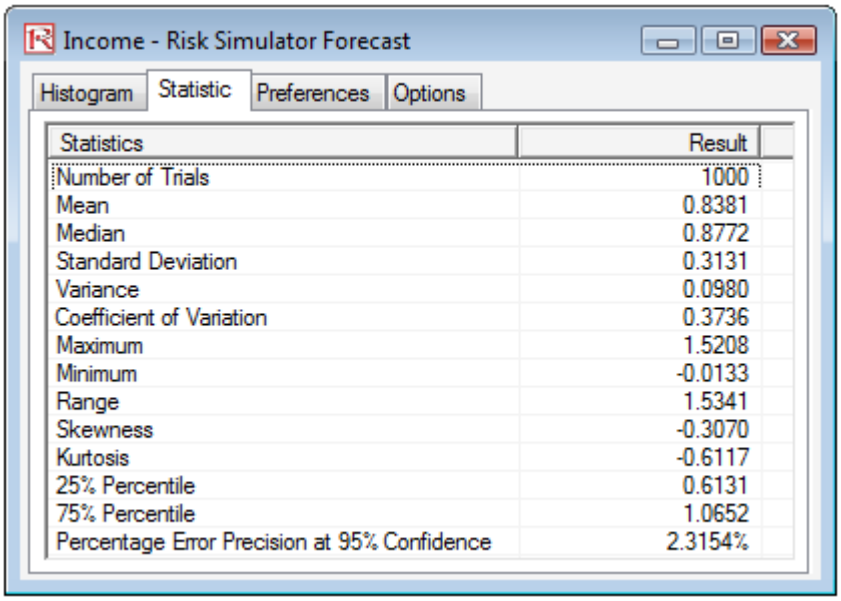


Figure 9: Forecast statistics

- **Preferences:** The preferences tab in the forecast chart (Figure 10) allows you to change the look and feel of the charts. For instance, if *Always Show Window On Top* is selected, the forecast charts will always be visible regardless of what other software is running on your computer. The *Semitransparent When Inactive* is a powerful option used to compare or overlay multiple forecast charts at once (e.g., enable this option on several forecast charts and drag them on top of one another to visually see the similarities or differences). *Histogram Resolution* allows you to change the number of bins of the histogram, anywhere from 5 bins to 100 bins. Also, the *Update Data Interval* section allows you to control how fast the simulation runs versus how often the forecast chart is updated. That is, if you wish to see the forecast chart updated at almost every trial, this will slow down the simulation as more memory is being allocated to updating the chart versus running the simulation. This is merely a user preference and in no way changes the results of the simulation, just the speed of completing the simulation.

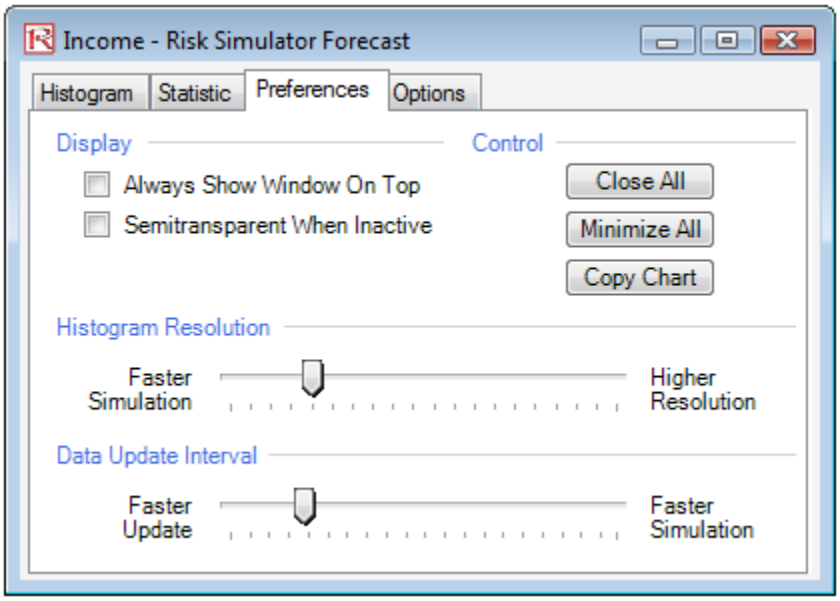


Figure 10: Forecast chart preferences

- **Options:** This forecast chart option allows you to show all the forecast data or to filter in or out values that fall within some specified interval, or within some standard deviation that you choose. Also, you can set the precision level here for this specific forecast to show the error levels in the statistics view. See the section on precision and error control for more details.

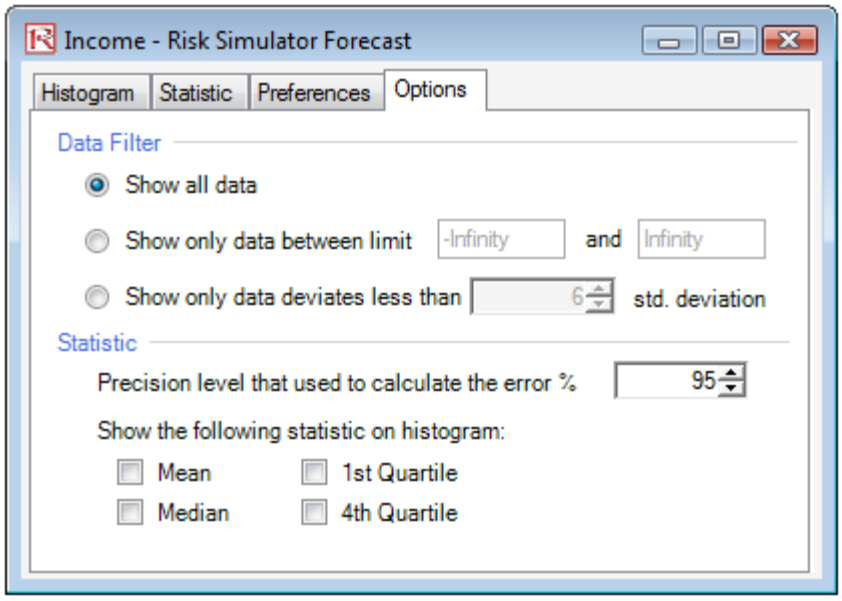


Figure 11: Forecast chart options

Using Forecast Charts and Confidence Intervals

In forecast charts, you can determine the probability of occurrence called *confidence intervals*; that is, given two values, what are the chances that the outcome will fall between these two values? Figure 12 illustrates that there is a 90% probability that the final outcome (in this case, the level of income) will be between $0.2781 and $1.3068. The two-tailed confidence interval can be obtained by first selecting *Two-Tail* as the type, entering the desired certainty value (e.g., 90) and hitting **Tab** on the keyboard. The two computed values corresponding to the certainty value then will be displayed. In this example, there is a 5% probability that income will be below $0.2781 and another 5% probability that income will be above $1.3068; that is, the two-tailed confidence interval is a symmetrical interval centered on the median or 50th percentile value. Thus, both tails will have the same probability.

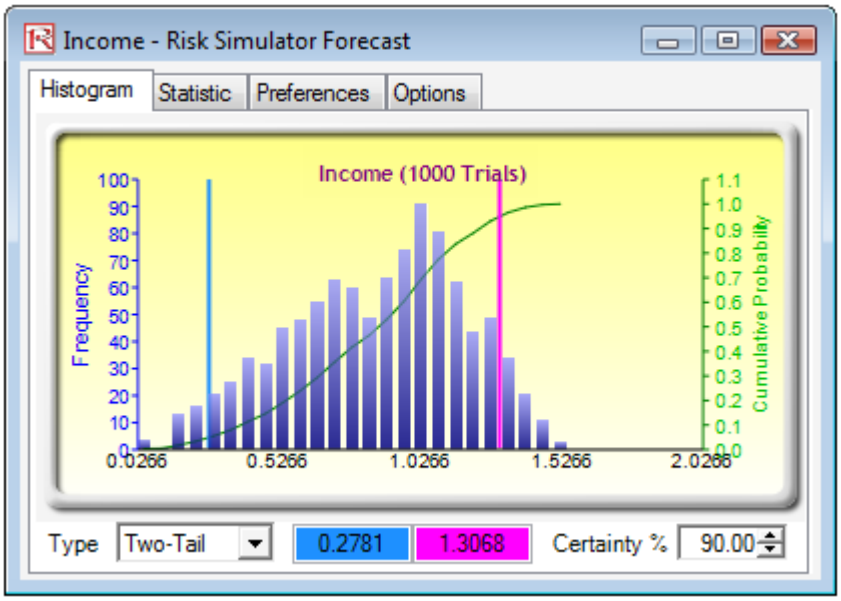


Figure 12: Forecast chart two-tailed confidence interval

Alternatively, a one-tail probability can be computed. Figure 13 shows a *Left-Tail* selection at 95% confidence (i.e., choose *Left-Tail* as the type, enter 95 as the certainty level, and hit *Tab* on the keyboard). This means that there is a 95% probability that the income will be below $1.3068 (i.e., 95% on the left tail of $1.3068) or a 5% probability that income will be above $1.3068, corresponding perfectly with the results seen in Figure 12.

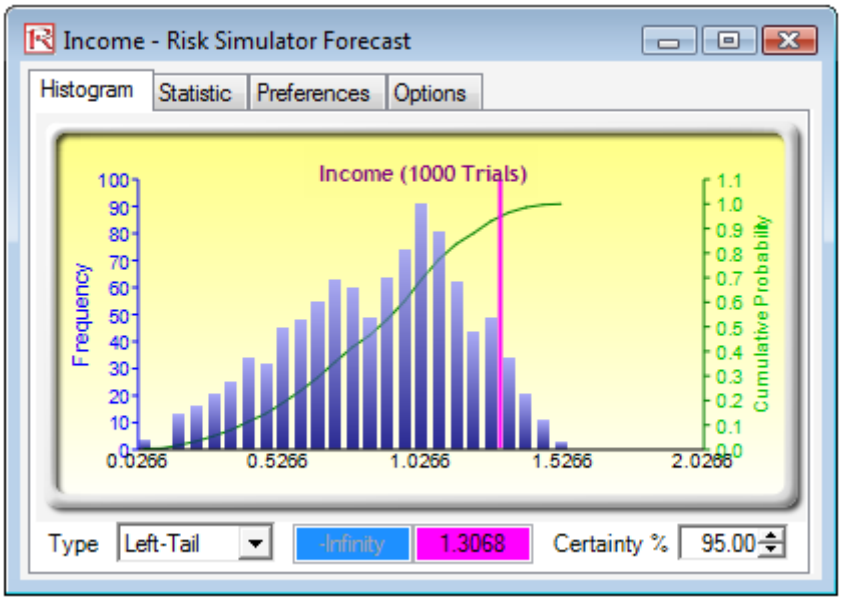


Figure 13: Forecast chart one-tailed confidence interval

In addition to evaluating the confidence interval (i.e., given a probability level and finding the relevant income values), you can determine the probability of a given income value (Figure 14). For instance, what is the probability that income will be less than $1? To do this, select the *Left-Tail* probability type, enter 1 into the value input box and hit **Tab**. The corresponding certainty will then be computed (in this case, there is a 64.80% probability income will be below $1).

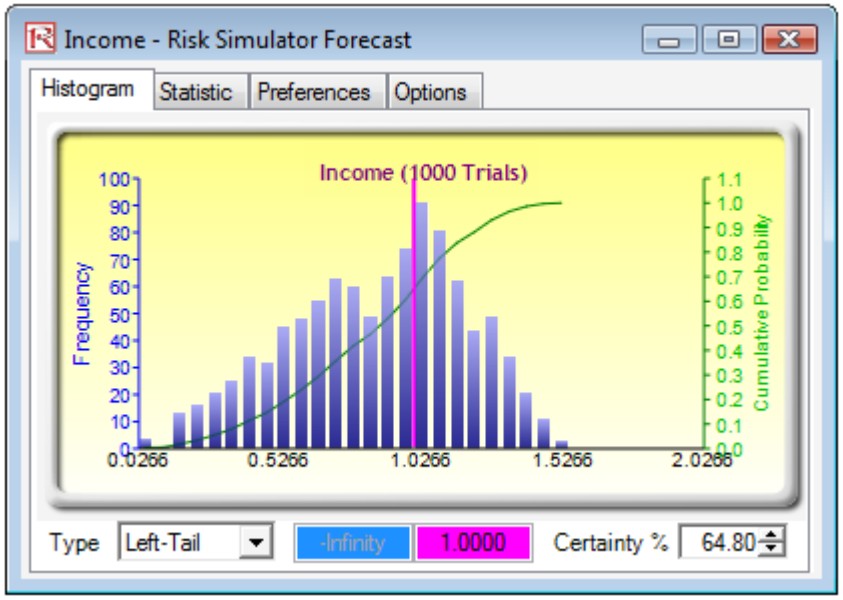


Figure 14: Forecast chart left tail probability evaluation

For the sake of completeness, you can select the *Right-Tail* probability type and enter the value 1 in the value input box, and hit **Tab** (Figure 15). The resulting probability indicates the right-tail probability past the value 1, that is, the probability of income exceeding $1 (in this case, we see that there is a 35.20% probability of income exceeding $1).

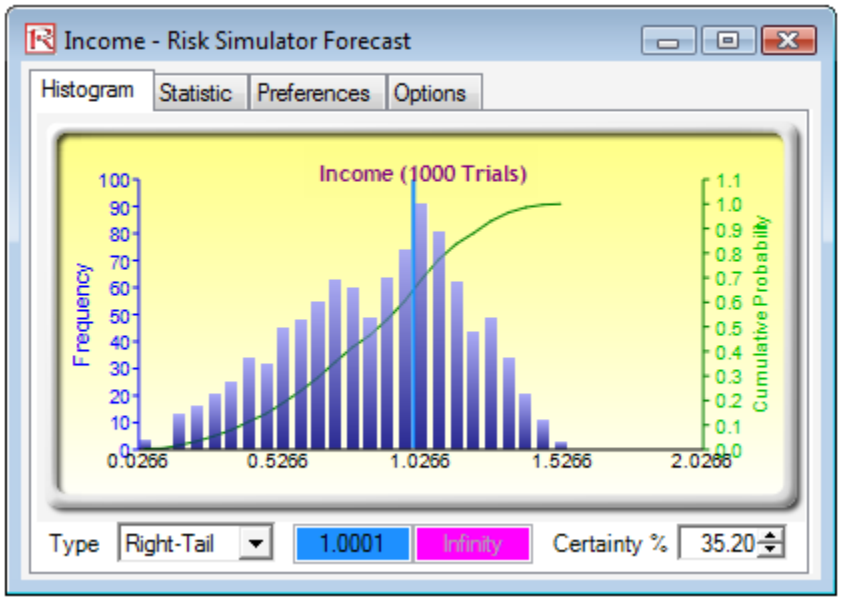


Figure 15: Forecast chart right-tail probability evaluation

Note that the forecast window is resizable by clicking on and dragging the bottom right corner of the window. Finally, it is always advisable that before rerunning a simulation, you reset the current simulation by selecting **Risk Simulator | Reset Simulation**.

Correlations and Precision Control

The correlation coefficient is a measure of the strength and direction of the relationship between two variables, and can take on any values between –1.0 and +1.0; that is, the correlation coefficient can be decomposed into its direction or sign (positive or negative relationship between two variables) and the magnitude or strength of the relationship (the higher the absolute value of the correlation coefficient, the stronger the relationship).

The correlation coefficient can be computed in several ways. The first approach is to manually compute the correlation coefficient r of a pair of variables x and y using:

$$r_{x,y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

The second approach is to use Excel's *CORREL* function. For instance, if the 10 data points for *x* and *y* are listed in cells A1:B10, then the Excel function to use is *CORREL (A1:A10, B1:B10)*.

The third approach is to run Risk Simulator's *Multi-Variable Distributional Fitting Tool* and the resulting correlation matrix will be computed and displayed.

It is important to note that correlation does not imply causation. Two completely unrelated random variables might display some correlation but this does not imply any causation between the two (e.g., sunspot activity and events in the stock market are correlated but there is no causation between the two).

There are two general types of correlations: parametric and nonparametric correlations. Pearson's correlation coefficient is the most common correlation measure; and usually is referred to simply as the correlation coefficient. However, Pearson's correlation is a parametric measure, which means that it requires both correlated variables to have an underlying normal distribution and that the relationship between the variables is linear. When these conditions are violated, which is often the case in Monte Carlo simulations, the nonparametric counterparts become more important. Spearman's rank correlation and Kendall's tau are the two nonparametric alternatives. The Spearman correlation is used most commonly and is most appropriate when applied in the context of Monte Carlo simulation—there is no dependence on normal distributions or linearity, meaning that correlations between different variables with different distribution can be applied. In order to compute the Spearman correlation, first rank all the *x* and *y* variable values and then apply the Pearson's correlation computation.

Risk Simulator uses the more robust nonparametric Spearman's rank correlation. However, to simplify the simulation process and to be consistent with Excel's correlation function, the correlation user inputs required are the Pearson's correlation coefficient. Risk Simulator then applies its own algorithms to convert them into Spearman's rank correlation, thereby simplifying the process.

Applying Correlations in Risk Simulator

Correlations can be applied in Risk Simulator in several ways:

- When defining assumptions, simply enter the correlations into the correlation grid in the Distribution Gallery.
- With existing data, run the Multi-Variable Distribution Fitting tool to perform distributional fitting and to obtain the correlation matrix between pairwise variables. If a simulation profile exists, the assumptions fitted automatically will contain the relevant correlation values.
- With the use of a direct-input correlation matrix, click on **Risk Simulator | Edit Correlations** to view and edit the correlation matrix used in the simulation.

Note that the correlation matrix must be positive definite; that is, the correlation must be mathematically valid. For instance, suppose you are trying to correlate three variables: grades of graduate students in a particular year, the number of beers they consume a week, and the number of hours they study a week. You would assume that these correlation relationships exist:

Grades and Beer: –	The more they drink, the lower the grades (no show on exams)
Grades and Study: +	The more they study, the higher the grades
Beer and Study: –	The more they drink, the less they study (drunk and partying all the time)

However, if you input a negative correlation between Grades and Study and assuming that the correlation coefficients have high magnitudes, the correlation matrix will be nonpositive definite. It would defy logic, correlation requirements, and matrix mathematics. However, smaller coefficients sometimes still can work even with the bad logic. When a nonpositive definite or bad correlation matrix is entered, Risk Simulator automatically informs you of the error and offers to adjust these correlations to something that is semipositive definite while still maintaining the overall structure of the correlation relationship (the same signs as well as the same relative strengths).

The Effects of Correlations in Monte Carlo Simulation

Although the computations required to correlate variables in a simulation are complex, the resulting effects are fairly clear. Figure 16 shows a simple correlation model (Correlation Effects Model in the example folder). The calculation for revenue is simply price multiplied by quantity. The same model is replicated for no correlations, positive correlation (+0.9), and negative correlation (–0.9) between price and quantity.

Correlation Model

	Without Correlation	Positive Correlation	Negative Correlation
Price	$2.00	$2.00	$2.00
Quantity	1.00	1.00	1.00
Revenue	$2.00	$2.00	$2.00

Figure 16: Simple correlation model

The resulting statistics are shown in Figure 17. Notice that the standard deviation of the model without correlations is 0.1450, compared to 0.1886 for the positive correlation model, and 0.0717 for the negative correlation model. That is, for simple models with positive relationships (e.g., additions and multiplications), negative correlations tend to reduce the average spread of the distribution and create a tighter and more concentrated forecast distribution as compared to positive correlations with larger average spreads. However, the mean remains relatively stable. This implies that correlations do little to change the expected value of projects but can reduce or increase a project's risk. Recall in financial theory that negatively correlated variables, projects, or assets when combined in a portfolio tend to create a diversification effect where the overall risk is reduced. Therefore, we see a smaller standard deviation for the negatively correlated model.

In a positively related model (e.g., A + B = C or A x B = C), a negative correlation reduces the risk (standard deviation and all other second moments of the distribution) of the result (C) whereas a positive correlation between the inputs (A and B) will increase the overall risk. The opposite is true for a negatively related model (e.g., A – B = C or A/B = C), where a positive correlation between the inputs will reduce the risk and a negative correlation increases the risk. In more complex models, as is often the case in real-life situations, the effects will be unknown *a priori* and can be determined only after a simulation is run.

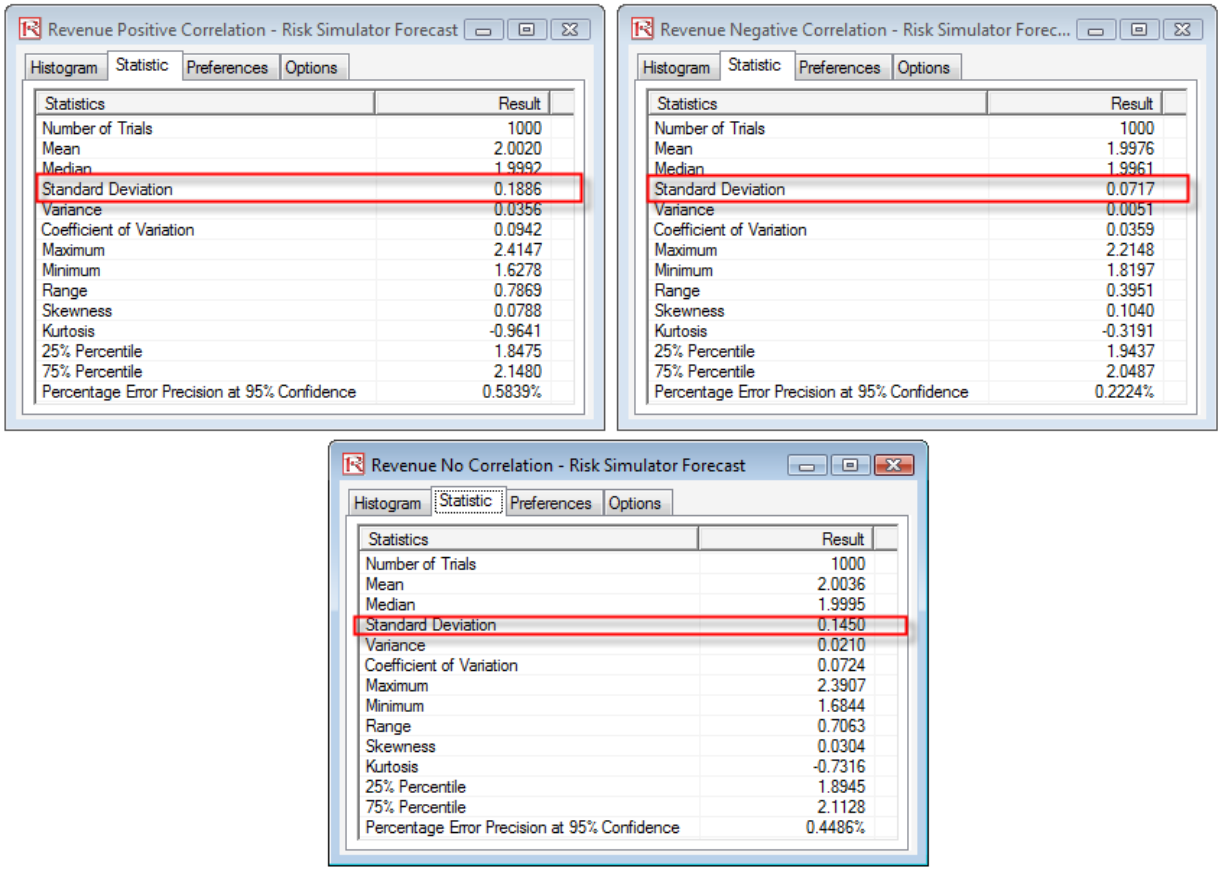


Figure 17: Correlation results

Figure 18 illustrates the results after running a simulation, extracting the raw data of the assumptions, and computing the correlations between the variables. The figure shows that the input assumptions are recovered in the simulation; that is, you enter +0.9 and –0.9 correlations and the resulting simulated values have the same correlations. Clearly there will be minor differences from one simulation run to another but when enough trials are run, the resulting recovered correlations approach those that were inputted.

Spearman's Nonlinear Rank Correlation on Raw Data Extracted from Simulation

Price Negative Correlation	Quantity Negative Correlation	Correlation	Price Positive Correlation	Quantity Positive Correlation	Correlation
676	145	**-0.90**	102	158	**0.89**
368	452		461	515	
264	880		515	477	
235	877		874	833	
122	711		769	792	
490	641		481	471	
336	638		627	446	
495	383		82	190	
241	568		659	674	
651	571		188	286	
854	59		458	439	
66	950		981	972	
707	262		528	569	
943	186		865	812	

Figure 18: Correlations recovered

Tornado and Sensitivity Tools in Simulation

One of the powerful simulation tools is tornado analysis—it captures the static impacts of each variable on the outcome of the model; that is, the tool automatically perturbs each variable in the model a preset amount, captures the fluctuation on the model's forecast or final result, and lists the resulting perturbations ranked from the most significant to the least. Figures 19 through 24 illustrate the application of a tornado analysis. For instance, Figure 19 is a sample discounted cash flow model where the input assumptions in the model are shown. The question is: What are the critical success drivers that affect the model's output the most? That is, what really drives the net present value of $96.63 or which input variable impacts this value the most?

The tornado chart tool can be obtained through **Simulation | Tools | Tornado Analysis**. To follow along the first example, open the **Tornado and Sensitivity Charts (Linear)** file in the examples folder. Figure 19 shows this sample model where cell G6 containing the net present value is chosen as the target result to be analyzed. The target cell's precedents in the model are used in creating the tornado chart. Precedents are all the input and intermediate variables that affect the outcome of the model. For instance, if the model consists of $A = B + C$, and where $C = D + E$, then B, D, and E are the precedents for A (C is not a precedent as it is only an intermediate calculated value). Figure 20 shows the testing range of each precedent variable used to estimate the target result. If the precedent variables are simple inputs, then the testing range will be a simple perturbation based on the range chosen (e.g., the default is ±10%). Each precedent variable can be perturbed at different percentages if required. A wider range is important as it is better able to test extreme values rather than smaller perturbations around the expected values. In certain

circumstances, extreme values may have a larger, smaller, or unbalanced impact (e.g., nonlinearities may occur where increasing or decreasing economies of scale and scope creep in for larger or smaller values of a variable) and only a wider range will capture this nonlinear impact.

Discounted Cash Flow Model

Base Year	*2005*	*Sum PV Net Benefits*	***$1,896.63***
Market Risk-Adjusted Discount Rate	*15.00%*	*Sum PV Investments*	***$1,800.00***
Private-Risk Discount Rate	*5.00%*	*Net Present Value*	***$96.63***
Annualized Sales Growth Rate	*2.00%*	*Internal Rate of Return*	*18.80%*
Price Erosion Rate	*5.00%*	*Return on Investment*	*5.37%*
Effective Tax Rate	*40.00%*		

	2005	*2006*	*2007*	*2008*	*2009*
Product A Avg Price/Unit	$10.00	$9.50	$9.03	$8.57	$8.15
Product B Avg Price/Unit	$12.25	$11.64	$11.06	$10.50	$9.98
Product C Avg Price/Unit	$15.15	$14.39	$13.67	$12.99	$12.34
Product A Sale Quantity ('000s)	50.00	51.00	52.02	53.06	54.12
Product B Sale Quantity ('000s)	35.00	35.70	36.41	37.14	37.89
Product C Sale Quantity ('000s)	20.00	20.40	20.81	21.22	21.65
Total Revenues	***$1,231.75***	***$1,193.57***	***$1,156.57***	***$1,120.71***	***$1,085.97***
Direct Cost of Goods Sold	$184.76	$179.03	$173.48	$168.11	$162.90
Gross Profit	***$1,046.99***	***$1,014.53***	***$983.08***	***$952.60***	***$923.07***
Operating Expenses	$157.50	$160.65	$163.86	$167.14	$170.48
Sales, General and Admin. Costs	$15.75	$16.07	$16.39	$16.71	$17.05
Operating Income (EBITDA)	***$873.74***	***$837.82***	***$802.83***	***$768.75***	***$735.54***
Depreciation	$10.00	$10.00	$10.00	$10.00	$10.00
Amortization	$3.00	$3.00	$3.00	$3.00	$3.00
EBIT	***$860.74***	***$824.82***	***$789.83***	***$755.75***	***$722.54***
Interest Payments	$2.00	$2.00	$2.00	$2.00	$2.00
EBT	***$858.74***	***$822.82***	***$787.83***	***$753.75***	***$720.54***
Taxes	$343.50	$329.13	$315.13	$301.50	$288.22
Net Income	***$515.24***	***$493.69***	***$472.70***	***$452.25***	***$432.33***
Noncash: Depreciation Amortization	$13.00	$13.00	$13.00	$13.00	$13.00
Noncash: Change in Net Working Capital	$0.00	$0.00	$0.00	$0.00	$0.00
Noncash: Capital Expenditures	$0.00	$0.00	$0.00	$0.00	$0.00
Free Cash Flow	***$528.24***	***$506.69***	***$485.70***	***$465.25***	***$445.33***
Investment Outlay	***$1,800.00***				
Financial Analysis					
Present Value of Free Cash Flow	***$528.24***	***$440.60***	***$367.26***	***$305.91***	***$254.62***
Present Value of Investment Outlay	***$1,800.00***	***$0.00***	***$0.00***	***$0.00***	***$0.00***
Net Cash Flows	($1,271.76)	$506.69	$485.70	$465.25	$445.33

Figure 19: Sample discounted cash flow model

Procedure:

Use these steps to create a tornado analysis:

- Select the single output cell (i.e., a cell with a function or equation) in an Excel model (e.g., cell G6 is selected in our example).
- Select **Risk Simulator | Tools | Tornado Analysis**.
- Review the precedents and rename them as appropriate (renaming the precedents to shorter names allows a more visually pleasing tornado and spider chart) and click OK. Alternatively, click on *Use Cell Address* to apply cell locations as the variable names.

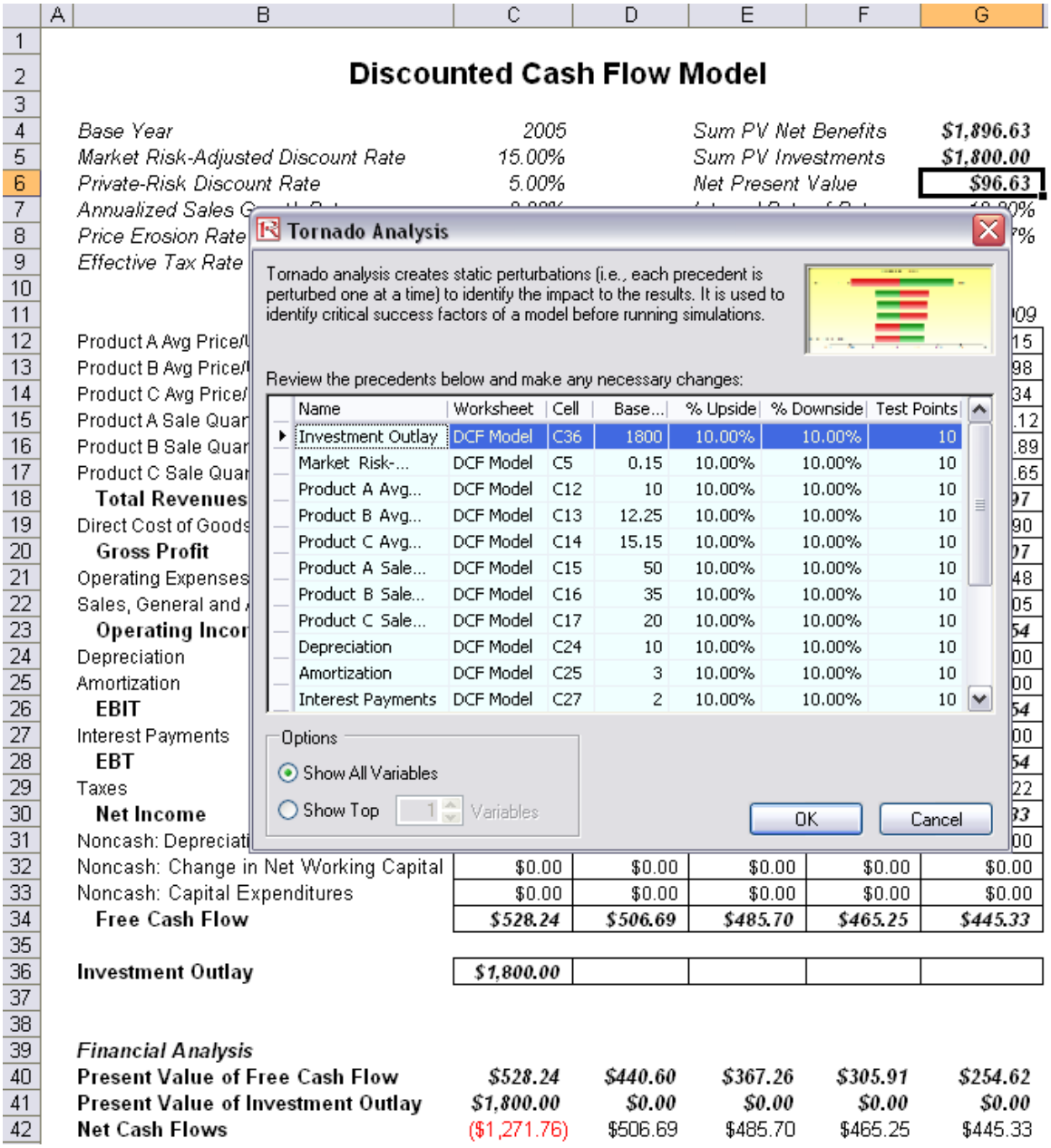


Figure 20: Running a tornado analysis

Results Interpretation:

Figure 21 shows the resulting tornado analysis report, which indicates that capital investment has the largest impact on net present value (NPV), followed by tax rate, average sale price and quantity demanded of the product lines, and so forth. The report contains four distinct elements:

- Statistical summary listing the procedure performed.
- Sensitivity table (Figure 22) shows the starting NPV base value of $96.63 and how each input is changed (e.g., Investment is changed from $1,800 to $1,980 on the upside with a +10% swing, and from $1,800 to $1,620 on the downside with a –10% swing). The resulting upside and downside values on NPV are –$83.37 and $276.63, with a total change of $360, making it the variable with the highest impact on NPV. The precedent variables are ranked from the highest impact to the lowest impact.
- The spider chart (Figure 23) illustrates these effects graphically. The y-axis is the NPV target value while the x-axis depicts the percentage change on each of the precedent value. The central point is the base case value at $96.63 at 0% change from the base value of each precedent. Positively sloped lines indicate a positive relationship or effect; negatively sloped lines indicate a negative relationship (e.g., investment is negatively sloped, which means that the higher the investment level, the lower the NPV). The absolute value of the slope indicates the magnitude of the effect computed as the percentage change in the result given a percentage change in the precedent. A steep line indicates a higher impact on the NPV y-axis given a change in the precedent x-axis.
- The tornado chart (Figure 24) illustrates the results in another graphical manner, where the highest impacting precedent is listed first. The x-axis is the NPV value with the center of the chart being the base case condition. Green (lighter) bars in the chart indicate a positive effect; red (darker) bars indicates a negative effect. Therefore, for investments, the red (darker) bar on the right side indicate a negative effect of investment on higher NPV—in other words, capital investment and NPV are negatively correlated. The opposite is true for price and quantity of products A to C (their green or lighter bars are on the right side of the chart).

Tornado and Spider Charts

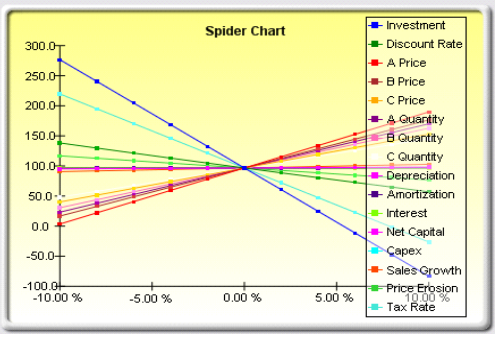
Statistical Summary

One of the powerful simulation tools is the tornado chart—it captures the static impacts of each variable on the outcome of the model. That is, the tool automatically perturbs each precedent variable in the model a user-specified preset amount, captures the fluctuation on the model's forecast or final result, and lists the resulting perturbations ranked from the most significant to the least. Precedents are all the input and intermediate variables that affect the outcome of the model. For instance, if the model consists of A = B + C, where C = D + E, then B, D, and E are the precedents for A (C is not a precedent as it is only an intermediate calculated value). The range and number of values perturbed is user-specified and can be set to test extreme values rather than smaller perturbations around the expected values. In certain circumstances, extreme values may have a larger, smaller, or unbalanced impact (e.g., nonlinearities may occur where increasing or decreasing economies of scale and scope creep occurs for larger or smaller values of a variable) and only a wider range will capture this nonlinear impact.

A tornado chart lists all the inputs that drive the model, starting from the input variable that has the most effect on the results. The chart is obtained by perturbing each precedent input at some consistent range (e.g., ±10% from the base case) one at a time, and comparing their results to the base case. A spider chart looks like a spider with a central body and its many legs protruding. The positively sloped lines indicate a positive relationship, while a negatively sloped line indicates a negative relationship. Further, spider charts can be used to visualize linear and nonlinear relationships. The tornado and spider charts help identify the critical success factors of an output cell in order to identify the inputs to simulate. The identified critical variables that are uncertain are the ones that should be simulated. Do not waste time simulating variables that are neither uncertain nor have little impact on the results.

Result

	Base Value: 96.6261638553219			Input Changes		
Precedent Cell	Output Downside	Output Upside	Effective Range	Input Downside	Input Upside	Base Case Value
Investment	$276.63	($83.37)	360.00	$1,620.00	$1,980.00	$1,800.00
Tax Rate	$219.73	($26.47)	246.20	36.00%	44.00%	40.00%
A Price	$3.43	$189.83	186.40	$9.00	$11.00	$10.00
B Price	$16.71	$176.55	159.84	$11.03	$13.48	$12.25
A Quantity	$23.18	$170.07	146.90	45.00	55.00	50.00
B Quantity	$30.53	$162.72	132.19	31.50	38.50	35.00
C Price	$40.15	$153.11	112.96	$13.64	$16.67	$15.15
C Quantity	$48.05	$145.20	97.16	18.00	22.00	20.00
Discount Rate	$138.24	$57.03	81.21	13.50%	16.50%	15.00%
Price Erosion	$116.80	$76.64	40.16	4.50%	5.50%	5.00%
Sales Growth	$90.59	$102.69	12.10	1.80%	2.20%	2.00%
Depreciation	$95.08	$98.17	3.08	$9.00	$11.00	$10.00
Interest	$97.09	$96.16	0.93	$1.80	$2.20	$2.00
Amortization	$96.16	$97.09	0.93	$2.70	$3.30	$3.00
Capex	$96.63	$96.63	0.00	$0.00	$0.00	$0.00
Net Capital	$96.63	$96.63	0.00	$0.00	$0.00	$0.00



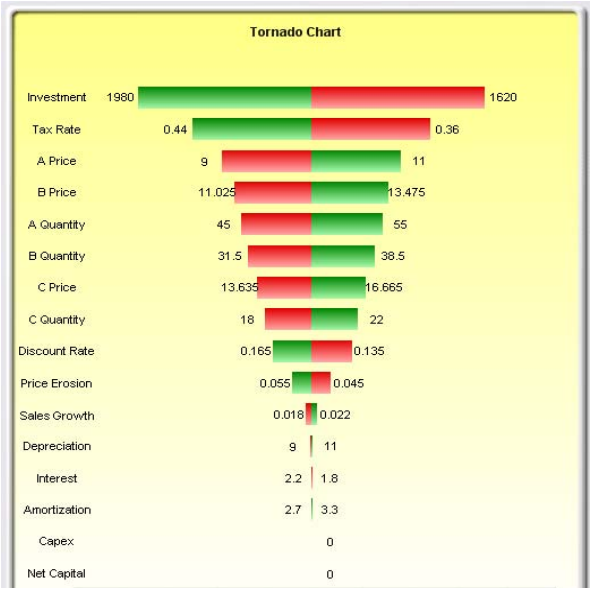


Figure 21: Tornado analysis report

Notes:

Remember that tornado analysis is a *static* sensitivity analysis applied on each input variable in the model—that is, each variable is perturbed individually and the resulting effects are tabulated. This makes tornado analysis a key component to execute before running a simulation. Caputiring and identifying the most important impact drivers in the modek is one of the very first steps in risk analysis. The next step is

to identify which of these important impact drivers are uncertain. These uncertain impact drivers are the critical success drivers of a project; the results of the model depend on these critical success drivers. These variables are the ones that should be simulated. Do not waste time simulating variables that are neither uncertain nor have little impact on the results. Tornado charts assist in identifying these critical success drivers quickly and easily. Following this example, it might be that price and quantity should be simulated, assuming if the required investment and effective tax rate are both known in advance and unchanging.

	Base Value: 96.6261638553219			Input Changes		
Precedent Cell	Output Downside	Output Upside	Effective Range	Input Downside	Input Upside	Base Case Value
Investment	$276.63	($83.37)	360.00	$1,620.00	$1,980.00	$1,800.00
Tax Rate	$219.73	($26.47)	246.20	36.00%	44.00%	40.00%
A Price	$3.43	$189.83	186.40	$9.00	$11.00	$10.00
B Price	$16.71	$176.55	159.84	$11.03	$13.48	$12.25
A Quantity	$23.18	$170.07	146.90	45.00	55.00	50.00
B Quantity	$30.53	$162.72	132.19	31.50	38.50	35.00
C Price	$40.15	$153.11	112.96	$13.64	$16.67	$15.15
C Quantity	$48.05	$145.20	97.16	18.00	22.00	20.00
Discount Rate	$138.24	$57.03	81.21	13.50%	16.50%	15.00%
Price Erosion	$116.80	$76.64	40.16	4.50%	5.50%	5.00%
Sales Growth	$90.59	$102.69	12.10	1.80%	2.20%	2.00%
Depreciation	$95.08	$98.17	3.08	$9.00	$11.00	$10.00
Interest	$97.09	$96.16	0.93	$1.80	$2.20	$2.00
Amortization	$96.16	$97.09	0.93	$2.70	$3.30	$3.00
Capex	$96.63	$96.63	0.00	$0.00	$0.00	$0.00
Net Capital	$96.63	$96.63	0.00	$0.00	$0.00	$0.00

Figure 22: Sensitivity table

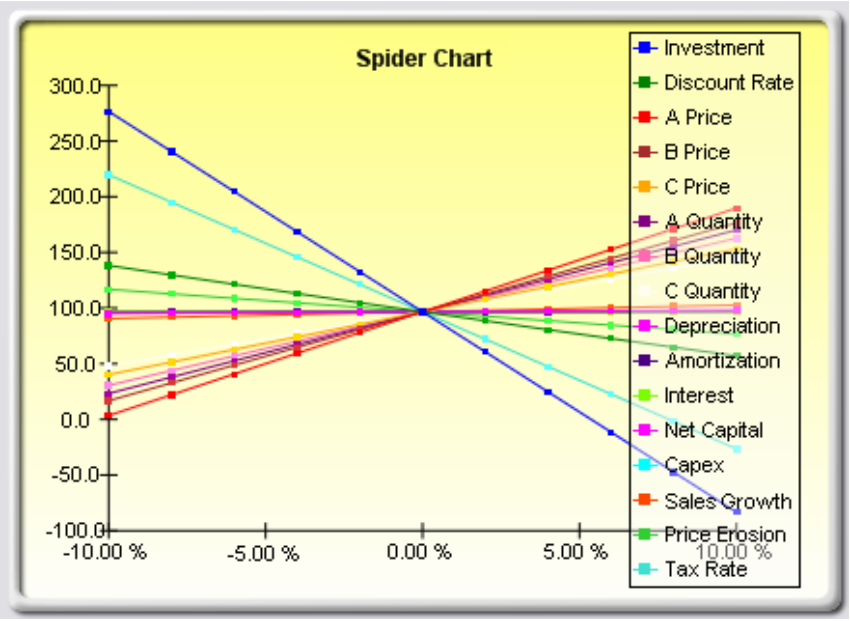


Figure 23: Spider chart

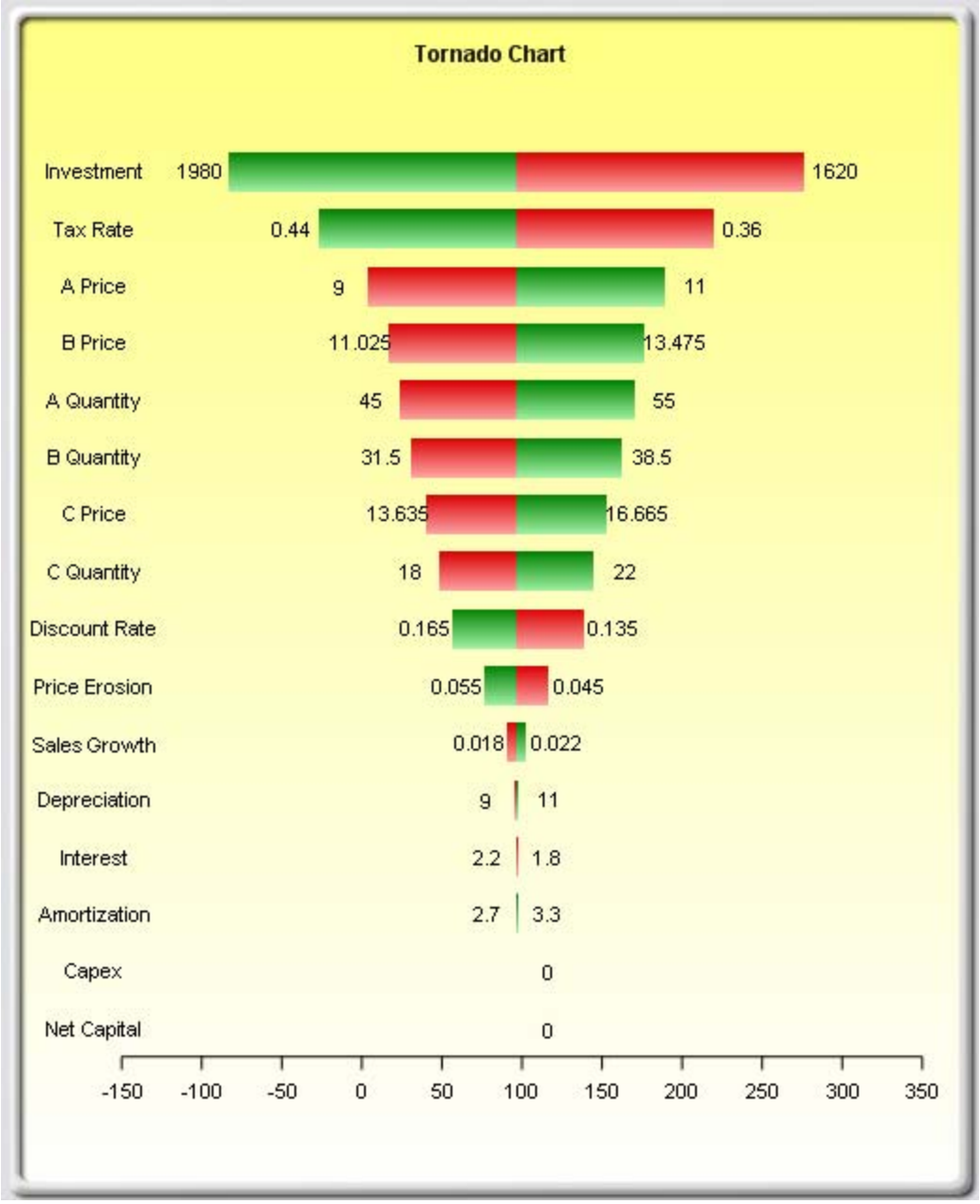


Figure 24: Tornado chart

Although the tornado chart is easier to read, the spider chart is important to determine if there are any nonlinearities in the model. For instance, Figure 25 shows another spider chart where nonlinearities are fairly evident (the lines on the graph are not straight but curved). The example model used is **Tornado and Sensitivity Charts (Nonlinear)**, which applies the Black-Scholes option pricing model. Such nonlinearities cannot be ascertained from a tornado chart and may be important information in the model or may provide decision makers important insight into the model's dynamics. For instance, in this Black-

Scholes model, the fact that stock price and strike price are nonlinearly related to the option value is important to know. This characteristic implies that option value will not increase or decrease proportionally to the changes in stock or strike price, and that there might be some interactions between these two prices as well as other variables. As another example, an engineering model depicting nonlinearities might indicate that a particular part or component, when subjected to a high enough force or tension, will break. Clearly, it is important to understand such nonlinearities.

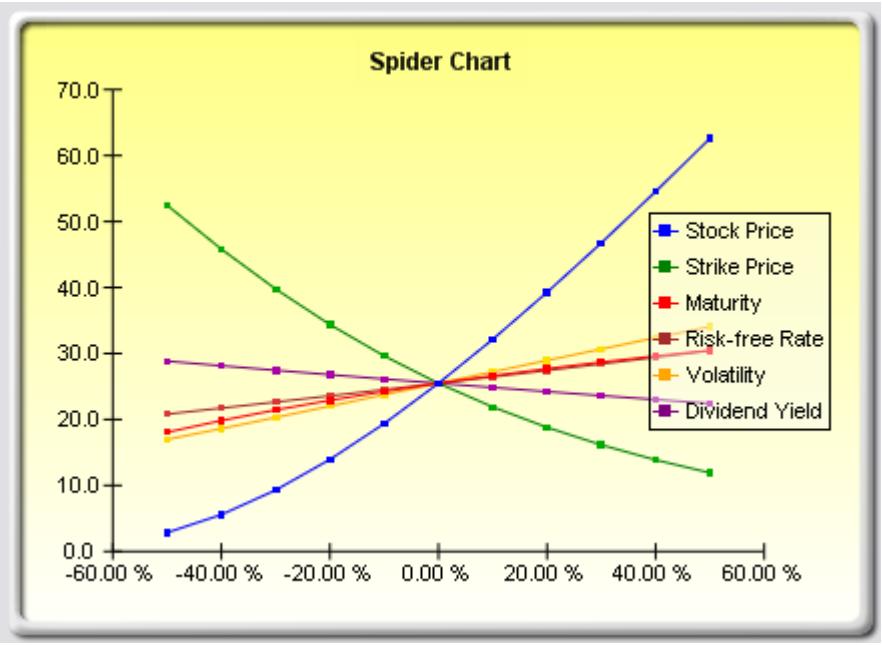


Figure 25: Nonlinear spider chart

Sensitivity Analysis

A related feature is sensitivity analysis. While tornado analysis (tornado charts and spider charts) applies static perturbations *before* a simulation run, sensitivity analysis applies dynamic perturbations created *after* the simulation run. Tornado and spider charts are the results of static perturbations, meaning that each precedent or assumption variable is perturbed a preset amount one at a time, and the fluctuations in the results are tabulated. In contrast, sensitivity charts are the results of dynamic perturbations in the sense that multiple assumptions are perturbed simultaneously and their interactions in the model and correlations among variables are captured in the fluctuations of the results. Tornado charts therefore identify which variables drive the results the most and hence are suitable for simulation; sensitivity charts identify the impact to the results when multiple interacting variables are simulated together in the model. This effect is clearly illustrated in Figure 26. Notice that the ranking of critical success drivers is similar to the tornado chart in the previous examples. However, if correlations are added between the assumptions, Figure 27 shows a very different picture. Notice for instance, price erosion had little impact on NPV but when some of the input assumptions are correlated, the interaction that exists between these correlated variables makes price erosion have more impact. Note that tornado analysis cannot capture these correlated dynamic relationships. Only after a simulation is run will such relationships become evident in a sensitivity analysis. A tornado chart's pre-simulation critical success factors therefore sometimes will be different from a sensitivity chart's post-simulation critical success factor. The post-simulation critical success factors should be the ones that are of interest as these more readily capture the interactions of the model precedents.

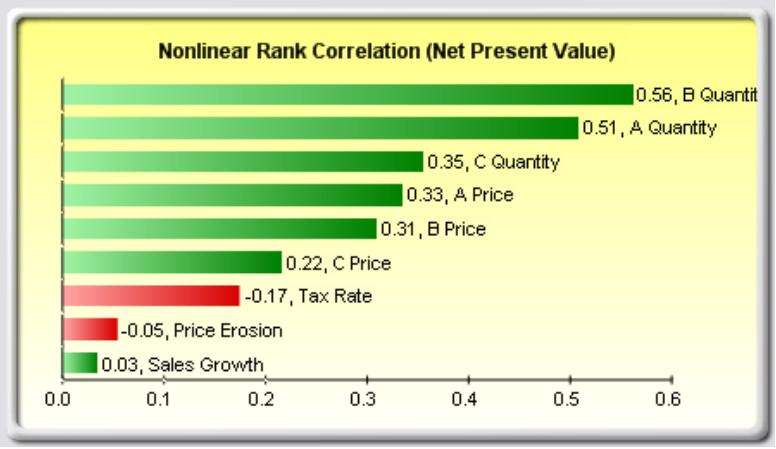


Figure 26: Sensitivity chart without correlations

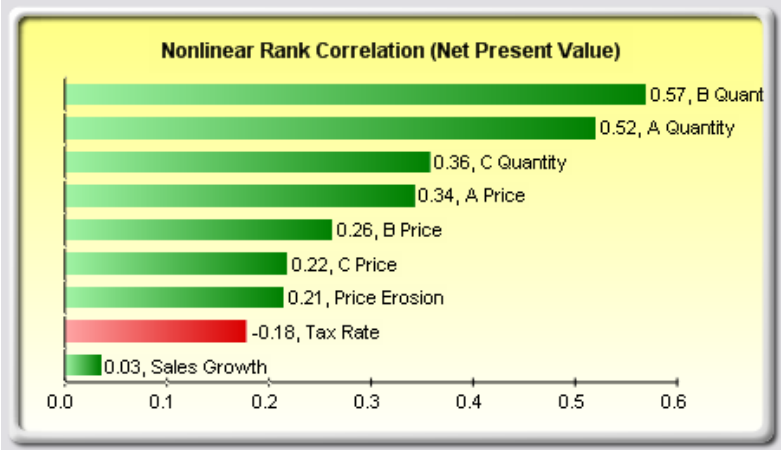


Figure 27: Sensitivity chart with correlations

Procedure:

Use these steps to create a sensitivity analysis:

- Open or create a model, define assumptions and forecasts, and run the simulation—the example here uses the Tornado and Sensitivity Charts (Linear) file
- Select **Risk Simulator | Tools | Sensitivity Analysis**
- Select the forecast of choice to analyze and click **OK** (Figure 28)

Note that sensitivity analysis cannot be run unless assumptions and forecasts have been defined and a simulation has been run.

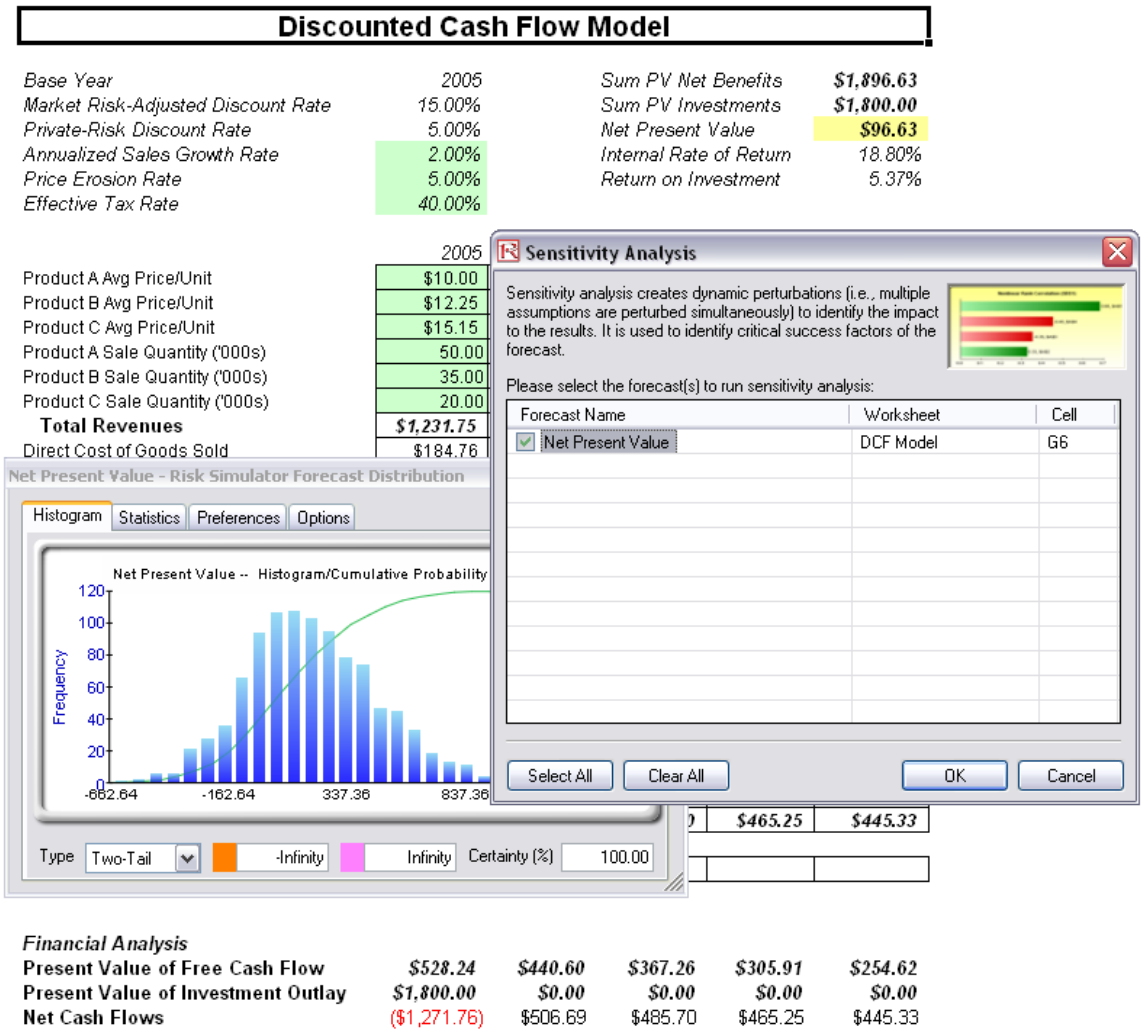


Figure 28: Running sensitivity analysis

Results Interpretation:

The results of the sensitivity analysis comprise a report and two key charts. The first is a nonlinear rank correlation chart (Figure 29) that ranks from highest to lowest the assumption-forecast correlation pairs. These correlations are nonlinear and nonparametric, making them free of any distributional requirements (i.e., an assumption with a Weibull distribution can be compared to another with a beta distribution). The results from this chart are fairly similar to that of the tornado analysis seen previously (of course without the capital investment value, which we decided was a known value and hence was not simulated), with one special exception. Tax rate was relegated to a much lower position in the sensitivity analysis chart (Figure 29) as compared to the tornado chart (Figure 24). This is because by itself, tax rate will have a significant impact. Once the other variables are interacting in the model, however, it appears that tax rate has less of a dominant effect. This is because tax rate has a smaller distribution as historical tax rates tend

not to fluctuate too much, and also because tax rate is a straight percentage value of the income before taxes, where other precedent variables have a larger effect on NPV. This example proves that it is important to perform sensitivity analysis after a simulation run to ascertain if there are any interactions in the model and if the effects of certain variables still hold. The second chart (Figure 30) illustrates the percent variation explained; that is, of the fluctuations in the forecast, how much of the variation can be explained by each of the assumptions after accounting for all the interactions among variables? Notice that the sum of all variations explained is usually close to 100% (sometimes other elements impact the model but they cannot be captured here directly), and if correlations exist, the sum may sometimes exceed 100% (due to the interaction effects that are cumulative).

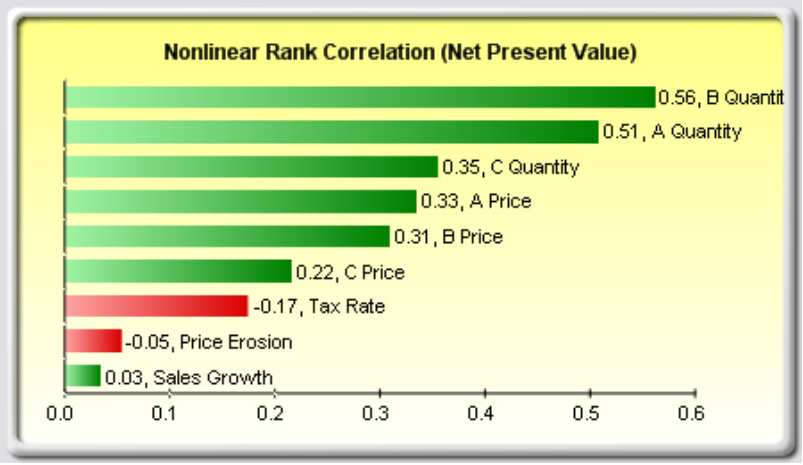


Figure 29: Rank correlation chart

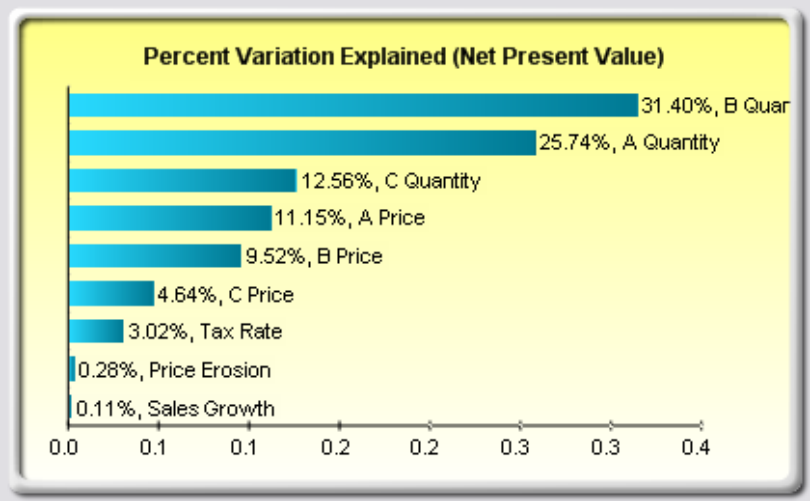


Figure 30: Contribution to variance chart

Notes:

Tornado analysis is performed before a simulation run while sensitivity analysis is performed after a simulation run. Spider charts in tornado analysis can consider nonlinearities while rank correlation charts in sensitivity analysis can account for nonlinear and distributional-free conditions.

Distributional Fitting: Single Variable and Multiple Variables

Another powerful simulation tool is distributional fitting; that is, which distribution does an analyst or engineer use for a particular input variable in a model? What are the relevant distributional parameters? If no historical data exist, then the analyst must make assumptions about the variables in question. One approach is to use the Delphi method, where a group of experts are tasked with estimating the behavior of each variable. For instance, a group of mechanical engineers can be tasked with evaluating the extreme possibilities of the diameter of a spring coil through rigorous experimentation or guesstimates. These values can be used as the variable's input parameters (e.g., uniform distribution with extreme values between 0.5 and 1.2). When testing is not possible (e.g., market share and revenue growth rate), management still can make estimates of potential outcomes and provide the best-case, most-likely case, and worst-case scenarios, whereupon a triangular or custom distribution can be created.

However, if reliable historical data are available, distributional fitting can be accomplished. Assuming that historical patterns hold and that history tends to repeat itself, historical data can be used to find the best-fitting distribution with their relevant parameters to better define the variables to be simulated. Figures 31 through 33 illustrate a distributional-fitting example. The next discussion uses the ***Data Fitting*** file in the examples folder.

Procedure:

Use these steps to perform a distributional fitting model:

- Open a spreadsheet with existing data for fitting (e.g., use the Data Fitting example file).
- Select the data you wish to fit not including the variable name. Data should be in a single column with multiple rows.
- Select **Risk Simulator | Tools | Distributional Fitting (Single-Variable)**.
- Select the specific distributions you wish to fit to or keep the default where all distributions are selected and click **OK** (Figure 31).
- Review the results of the fit, choose the relevant distribution you want and click **OK** (Figure 32).

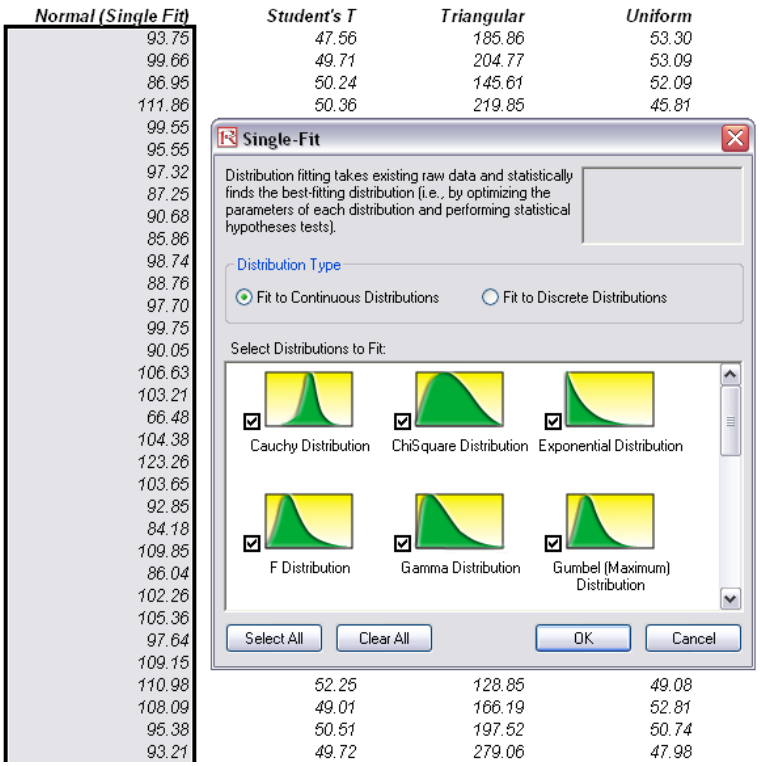


Figure 31: Single-variable distributional fitting

Results Interpretation:

The null hypothesis (H_o) being tested is such that the fitted distribution is the same distribution as the population from which the sample data to be fitted comes. Thus, if the computed p-value is lower than a critical alpha level (typically 0.10 or 0.05), then the distribution is the wrong distribution. Conversely, the *higher the p-value, the better the distribution fits the data*. Roughly, you can think of p-value as a *percentage explained*, that is, if the p-value is 1.00 (Figure 32), then setting a normal distribution with a mean of 100.67 and a standard deviation of 10.40 explains close to 100% of the variation in the data, indicating an especially good fit. The data was from a 1,000-trial simulation in Risk Simulator based on a normal distribution with a mean of 100 and a standard deviation of 10. Because only 1,000 trials were simulated, the resulting distribution is fairly close to the specified distributional parameters, and in this case, about a 100% precision.

Both the results (Figure 32) and the report (Figure 33) show the test statistic, p-value, theoretical statistics (based on the selected distribution), empirical statistics (based on the raw data), the original data (to maintain a record of the data used), and the assumption complete with the relevant distributional parameters (i.e., if you selected the option to automatically generate assumption and if a simulation profile already exists). The results also rank all the selected distributions and how well they fit the data.

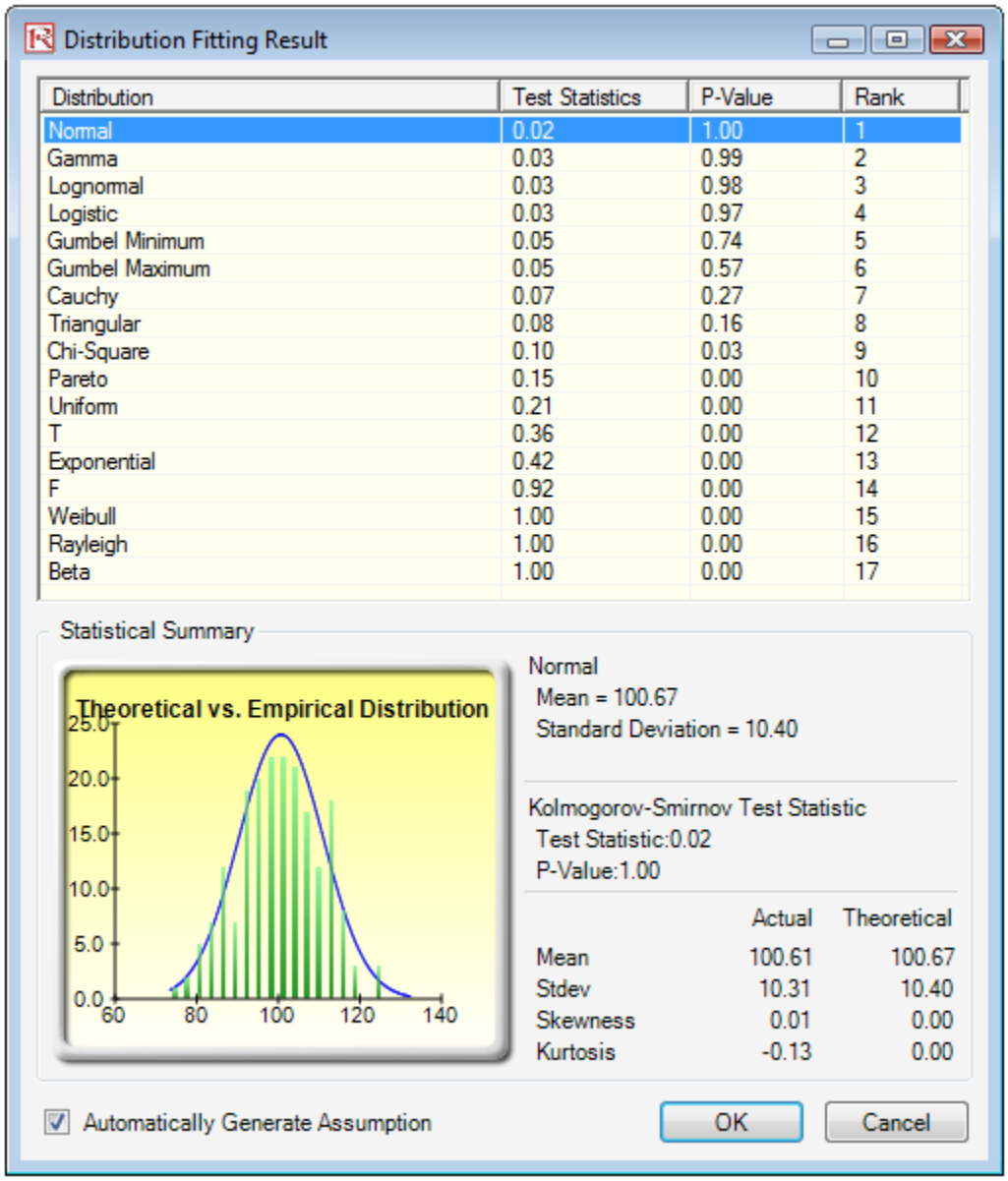
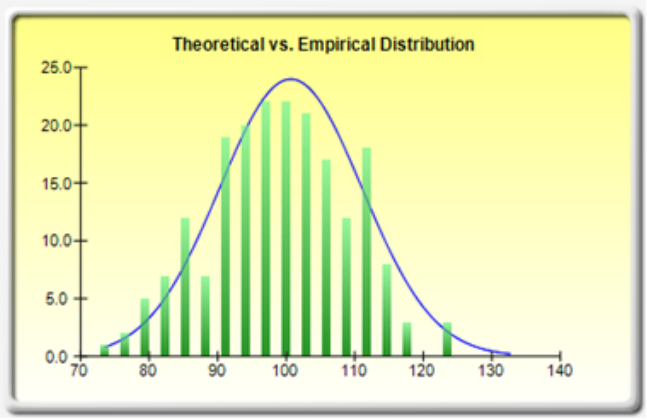


Figure 32: Distributional fitting result

Statistical Summary

Fitted Assumption	100.61
Fitted Distribution	**Normal**
Mean	100.67
Sigma	10.40
Kolmogorov-Smirnov Statistic	0.02
P-Value for Test Statistic	0.9996

	Actual	Theoretical
Mean	100.61	100.67
Standard Deviation	10.31	10.40
Skewness	0.01	0.00
Excess Kurtosis	-0.13	0.00



Original Fitted Data

73.53	78.21	78.52	79.50	79.72	79.74	81.56	82.08	82.68	82.75	83.34	83.64	84.09
84.66	85.00	85.35	85.51	86.04	86.79	86.82	86.91	87.02	87.03	87.45	87.53	87.66
88.05	88.45	88.51	89.95	90.19	90.54	90.68	90.96	91.25	91.49	91.56	91.94	92.06
92.36	92.41	92.45	92.70	92.80	92.84	93.21	93.26	93.48	93.73	93.75	93.77	93.82
94.00	94.15	94.51	94.57	94.64	94.69	94.95	95.57	95.62	95.71	95.78	95.83	95.97
96.20	96.24	96.40	96.43	96.47	96.81	96.88	97.00	97.07	97.21	97.23	97.48	97.70
97.77	97.85	98.15	98.17	98.24	98.28	98.32	98.33	98.35	98.65	99.03	99.27	99.46
99.47	99.55	99.73	99.96	100.08	100.24	100.36	100.42	100.44	100.48	100.49	100.83	101.17
101.28	101.34	101.45	101.46	101.55	101.73	101.74	101.81	102.29	102.55	102.58	102.60	102.70
103.17	103.21	103.22	103.32	103.34	103.45	103.65	103.66	103.72	103.81	103.90	103.99	104.46
104.57	104.76	105.20	105.44	105.50	105.52	105.58	105.66	105.87	105.90	105.90	106.29	106.35
106.59	107.01	107.68	107.70	107.93	108.17	108.20	108.34	108.42	108.43	108.49	108.70	109.15
109.22	109.35	109.52	109.75	110.04	110.16	110.25	110.54	111.05	111.06	111.44	111.76	111.90
111.95	112.07	112.19	112.29	112.32	112.42	112.48	112.85	112.92	113.50	113.59	113.63	113.70
114.13	114.14	114.21	114.91	114.95	115.40	115.58	115.66	116.58	116.98	117.60	118.67	119.24
119.52	124.14	124.16	124.39	132.30								

Figure 33: Distributional fitting report

Bootstrap Simulation

Bootstrap simulation is a simple technique that estimates the reliability or accuracy of forecast statistics or other sample raw data. Bootstrap simulation can be used to answer a lot of confidence and precision-based questions in simulation. For instance, suppose an identical model (with identical assumptions and forecasts but without any random seeds) is run by 100 different people, the results will clearly be slightly different. The question is, if we collected all the statistics from these 100 people, how will the mean be distributed, or the median, or the skewness, or excess kurtosis? Suppose one person has a mean value of, say, 1.50 while another has 1.52. Are these two values statistically significantly different from one another or are they statistically similar and the slight difference is due entirely to random chance? What about 1.53? So, how far is far enough to say that the values are statistically different? In addition, if a model's resulting skewness is –0.19, is this forecast distribution negatively skewed or is it statistically close enough to zero to state that this distribution is symmetrical and not skewed? Thus, if we bootstrapped this forecast 100 times (i.e., run a 1,000-trial simulation for 100 times and collect the 100 skewness coefficients), the skewness distribution would indicate how far zero is away from –0.19. If the

90% confidence on the bootstrapped skewness distribution contains the value zero, then we can state that on a 90% confidence level, this distribution is symmetrical and not skewed, and the value –0.19 is statistically close enough to zero. Otherwise, if zero falls outside of this 90% confidence area, then this distribution is negatively skewed. The same analysis can be applied to excess kurtosis and other statistics.

Essentially, bootstrap simulation is a hypothesis-testing tool. Classical methods used in the past relied on mathematical formulas to describe the accuracy of sample statistics. These methods assume that the distribution of a sample statistic approaches a normal distribution, making the calculation of the statistic's standard error or confidence interval relatively easy. However, when a statistic's sampling distribution is not normally distributed or easily found, these classical methods are difficult to use. In contrast, bootstrapping analyzes sample statistics empirically by sampling the data repeatedly and creating distributions of the different statistics from each sampling. The classical methods of hypothesis testing are available in Risk Simulator and are explained in the next section. Classical methods provide higher power in their tests but rely on normality assumptions and can be used only to test the mean and variance of a distribution, as compared to bootstrap simulation, which provides lower power but is nonparametric and distribution-free, and can be used to test any distributional statistic.

Procedure:

- Run a simulation with assumptions and forecasts
- Select **Risk Simulator | Tools | Nonparametric Bootstrap**
- Select only *one* forecast to bootstrap, select the statistic(s) to bootstrap, and enter the number of bootstrap trials and click **OK** (Figure 34)

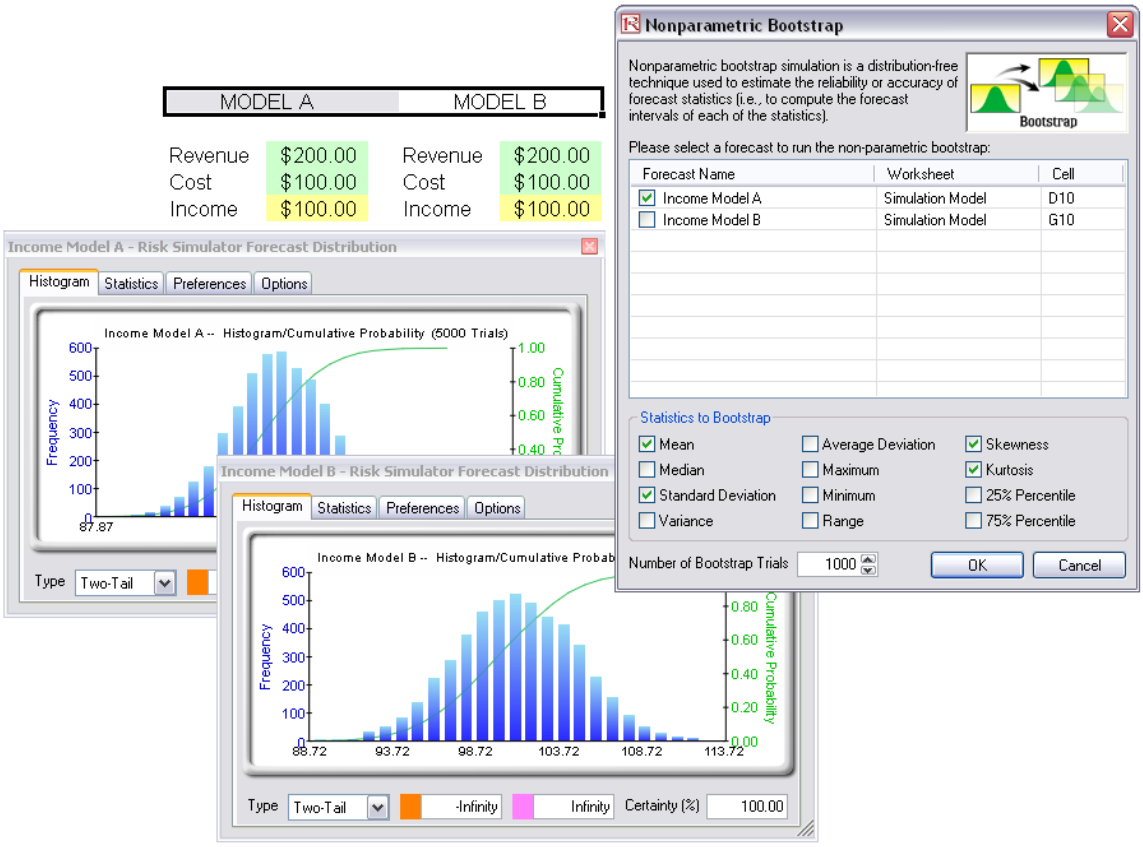


Figure 34: Nonparametric bootstrap simulation

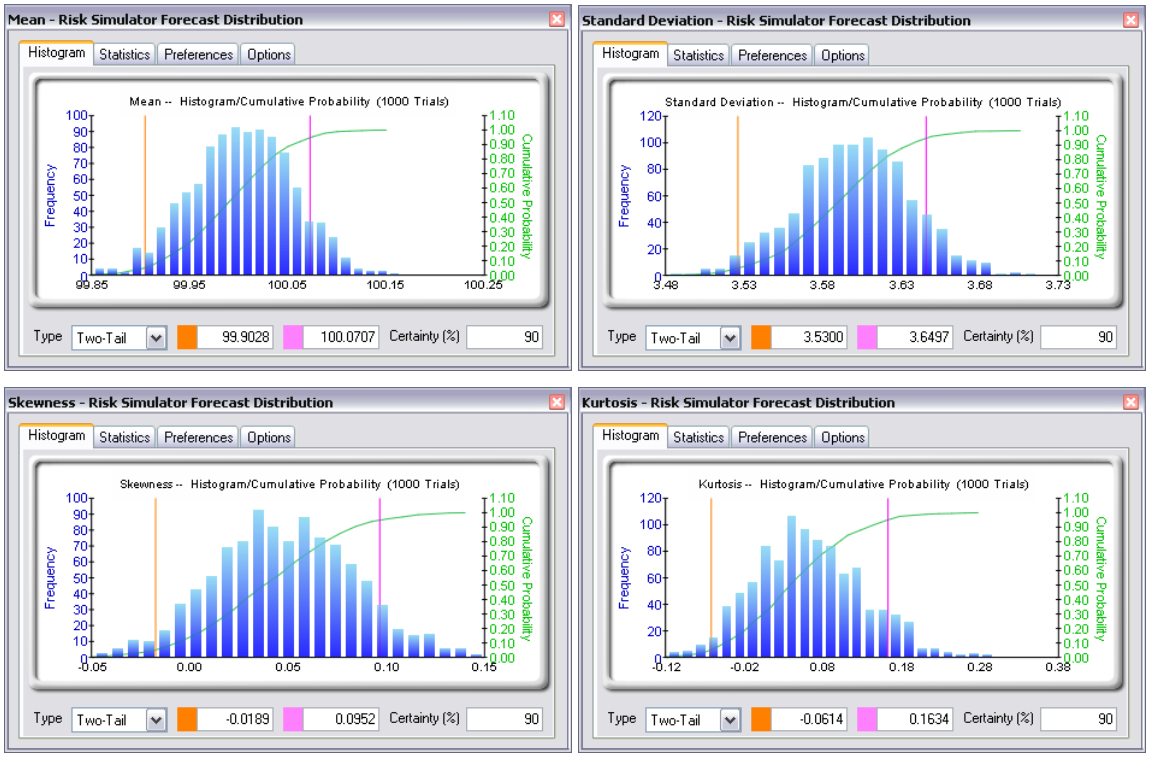


Figure 35: Bootstrap simulation results

Results Interpretation:

Figure 35 illustrates some sample bootstrap results. The example file used was *Hypothesis Testing and Bootstrap Simulation.* For instance, the 90% confidence for the skewness statistic is between –0.0189 and 0.0952, such that the value 0 falls within this confidence, indicating that on a 90% confidence, the skewness of this forecast is not statistically significantly different from zero, or that this distribution can be considered as symmetrical and not skewed. Conversely, if the value 0 falls outside of this confidence, then the opposite is true. The distribution is skewed (positively skewed if the forecast statistic is positive, and negatively skewed if the forecast statistic is negative).

Notes:

The term *bootstrap* comes from the saying, "to pull oneself up by one's own bootstraps," and is applicable because this method uses the distribution of statistics themselves to analyze the accuracy of the statistics. Nonparametric simulation is simply randomly picking golf balls from a large basket with replacement where each golf ball is based on a historical data point. Suppose there are 365 golf balls in the basket (representing 365 historical data points). Imagine if you will that the value of each golf ball picked at random is written on a large whiteboard. The results of the 365 balls picked with replacement are written in the first column of the board with 365 rows of numbers. Relevant statistics (e.g., mean,

median, mode, standard deviation, etc.) are calculated on these 365 rows. The process is then repeated, say, 5,000 times. The whiteboard will now be filled with 365 rows and 5,000 columns. Hence, 5,000 sets of statistics (that is, there will be 5,000 means, 5,000 medians, 5,000 modes, 5,000 standard deviations, and so forth) are tabulated and their distributions shown. The relevant *statistics of the statistics* are then tabulated, where from these results you can ascertain how confident the simulated statistics are. Finally, bootstrap results are important because according to the *Law of Large Numbers* and *Central Limit Theorem* in statistics, the mean of the sample means is an unbiased estimator and approaches the true population mean when the sample size increases.

Hypothesis Testing

A hypothesis test is performed when testing the means and variances of two distributions to determine if they are statistically identical or statistically different from one another (i.e., to see if the differences between the means and variances of two different forecasts that occur are based on random chance or if they are, in fact, statistically significantly different from one another).

This analysis is related to bootstrap simulation with several differences. Classical hypothesis testing uses mathematical models and is based on theoretical distributions. This means that the precision and power of the test is higher than bootstrap simulation's empirically based method of simulating a simulation and letting the data tell the story. However, the classical hypothesis test is applicable only for testing means and variances of two distributions (and by extension, standard deviations) to see if they are statistically identical or different. In contrast, nonparametric bootstrap simulation can be used to test for any distributional statistics, making it more useful; the drawback is its lower testing power. Risk Simulator provides both techniques from which to choose.

Procedure:

- Run a simulation.
- Select **Risk Simulator | Tools | Hypothesis Testing**.
- Select the two forecasts to test, select the type of hypothesis test you wish to run, and click **OK** (Figure 36).

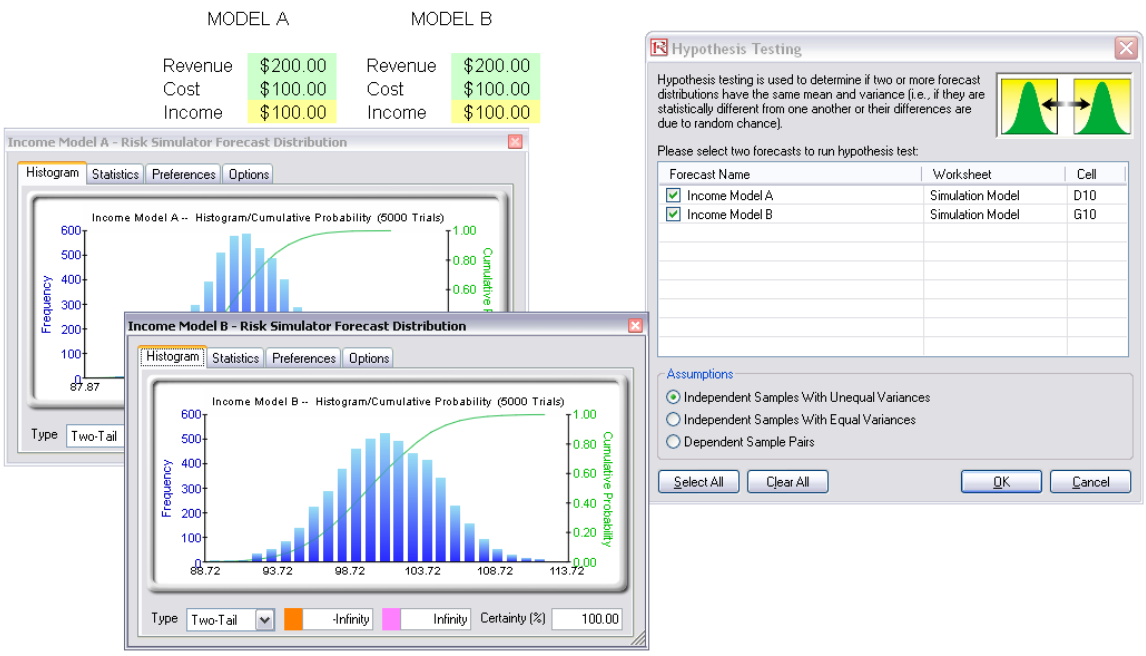


Figure 36: Hypothesis testing

Results Interpretation:

A two-tailed hypothesis test is performed on the null hypothesis (H_o) such that the population means of the two variables are statistically identical to one another. The alternative hypothesis (H_a) is such that the population means are statistically different from one another. If the calculated p-values are less than or equal to 0.01, 0.05, or 0.10 alpha test levels, it means that the null hypothesis is rejected, which implies that the forecast means are statistically significantly different at the 1%, 5% and 10% significance levels. If the null hypothesis is not rejected when the p-values are high, the means of the two forecast distributions are statistically similar to one another. The same analysis is performed on variances of two forecasts at a time using the pairwise F-test. If the p-values are small, then the variances (and standard deviations) are statistically different from one another; otherwise, for large p-values, the variances are statistically identical to one another. The example file used was *Hypothesis Testing and Bootstrap Simulation.*

Notes:

The two-variable t-test with unequal variances (the population variance of forecast 1 is expected to be different from the population variance of forecast 2) is appropriate when the forecast distributions are from different populations (e.g., data collected from two different geographical locations, or two different operating business units, and so forth). The two-variable t-test with equal variances (the population variance of forecast 1 is expected to be equal to the population variance of forecast 2) is appropriate when

the forecast distributions are from similar populations (e.g., data collected from two different engine designs with similar specifications, and so forth). The paired dependent two-variable t-test is appropriate when the forecast distributions are from exactly the same population and subjects (e.g., data collected from the same group of patients before an experimental drug was used and after the drug was applied, and so forth).

Data Extraction, Saving Simulation Results, and Generating Reports

Raw data of a simulation can be extracted very easily using Risk Simulator's *Data Extraction* routine. Both assumptions and forecasts can be extracted but a simulation must be run first. The extracted data can then be used for a variety of other analyses and the data can be extracted to different formats—for use in spreadsheets, databases, and other software products.

Procedure:

- Open or create a model, define assumptions and forecasts, and run the simulation.
- Select **Risk Simulator | Tools | Data Extraction**.
- Select the assumptions and/or forecasts you wish to extract the data from and click OK.

The simulated data can be extracted to an Excel worksheet, a flat text file (for easy import into other software applications), or as *.risksim files (which can be reopened as Risk Simulator forecast charts at a later date). Finally, you can create a simulation report of all the assumptions and forecasts in the model by going to **Risk Simulator | Create Report**. A sample report is shown in Figure 37.

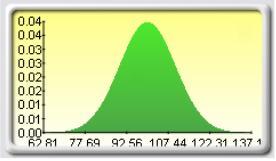
Simulation - Example Profile

General

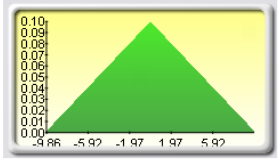
Number of Trials	1000
Stop Simulation on Error	No
Random Seed	123456
Enable Correlations	Yes

Assumptions

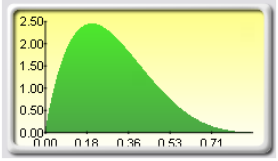
Name	ıple First Assumption
Enabled	Yes
Cell	E8
Dynamic Simulation	No
Range	
Minimum	-Infinity
Maximum	+Infinity
Distribution	Normal
Mean	100
Standard Deviation	10



Name	e Second Assumption
Enabled	Yes
Cell	E9
Dynamic Simulation	No
Range	
Minimum	-Infinity
Maximum	+Infinity
Distribution	Triangular
Minimum	-10
Most Likely	0
Maximum	10



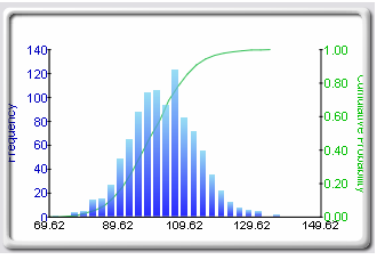
Name	ıple Third Assumption
Enabled	Yes
Cell	E10
Dynamic Simulation	No
Range	
Minimum	-Infinity
Maximum	+Infinity
Distribution	Beta
Alpha	2
Beta	5



Forecasts

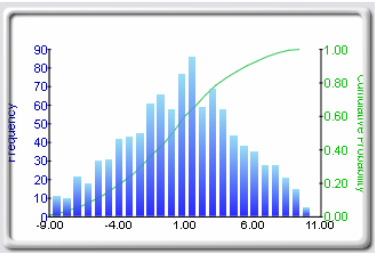
Name	Sample First Forecast
Enabled	Yes
Cell	E12
Forecast Precision	
Precision Level	—
Error Level	—

Number of Datapoints	1000
Mean	100.0400
Median	99.8427
Standard Deviation	9.8331
Variance	96.6903
Average Deviation	7.8397
Maximum	134.5452
Minimum	66.9132
Range	67.6320
Skewness	0.1121
Kurtosis	0.1401
25% Percentile	93.3563
75% Percentile	106.3153
Error Precision at 95%	0.0061



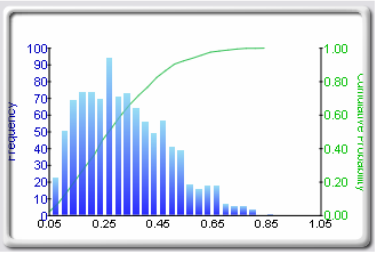
Name	nple Second Forecast
Enabled	Yes
Cell	E13
Forecast Precision	
Precision Level	—
Error Level	—

Number of Datapoints	1000
Mean	-0.0806
Median	0.0755
Standard Deviation	4.1171
Variance	16.9506
Average Deviation	3.3389
Maximum	9.3923
Minimum	-9.7671
Range	19.1594
Skewness	-0.0494
Kurtosis	-0.5394
25% Percentile	-2.8924
75% Percentile	2.8015
Error Precision at 95%	3.1644



Name	ample Third Forecast
Enabled	Yes
Cell	E14
Forecast Precision	
Precision Level	—
Error Level	—

Number of Datapoints	1000
Mean	0.2861
Median	0.2621
Standard Deviation	0.1593
Variance	0.0254
Average Deviation	0.1305
Maximum	0.8358
Minimum	0.0126
Range	0.8232
Skewness	0.5797
Kurtosis	-0.2064
25% Percentile	0.1590
75% Percentile	0.3935
Error Precision at 95%	0.0345



Correlation Matrix

	Sample First Assumption	ssumption	ssumption
Sample First Assumption	1.00		
Sample Second Assumption	0.00	1.00	
Sample Third Assumption	0.00	0.00	1.00

Figure 37 – Sample Simulation Report

Regression and Forecasting Diagnostic Tool

This advanced analytical tool in Risk Simulator is used to determine the econometric properties of your data. The diagnostics include checking the data for heteroskedasticity, nonlinearity, outliers, specification errors, micronumerosity, stationarity and stochastic properties, normality and sphericity of the errors, and multicollinearity. Each test is described in more detail in their respective reports in the model.

Procedure:

- Open the example model (**Risk Simulator | Examples | Regression Diagnostics**) and go to the *Time-Series Data* worksheet and select the data including the variable names (cells **C5:H55**).
- Click on **Risk Simulator** | **Tools** | **Diagnostic Tool**.
- Check the data and select the *Dependent Variable Y* from the drop down menu. Click **OK** when finished (Figure 38).

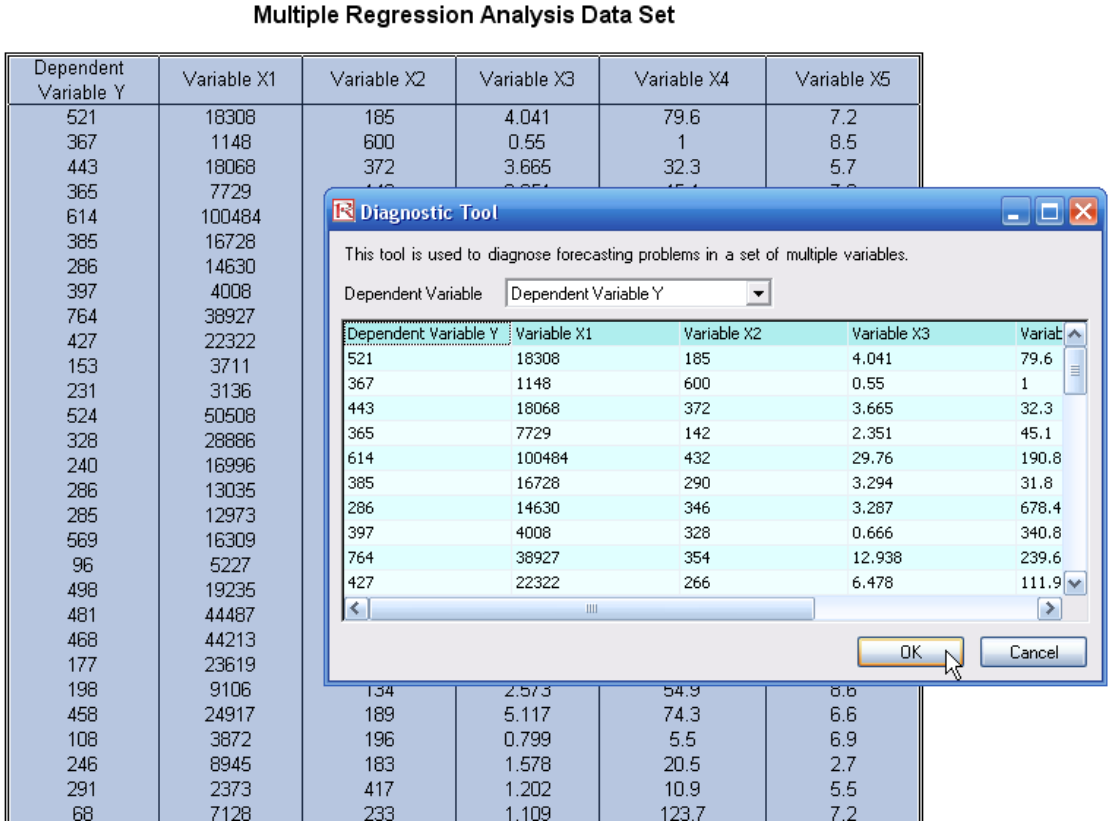


Figure 38 – Running the Data Diagnostic Tool

A common violation in forecasting and regression analysis is heteroskedasticity, that is, the variance of the errors increases over time (see Figure 39 for test results using the diagnostic tool). Visually, the width

of the vertical data fluctuations increases or fans out over time, and typically, the coefficient of determination (R-squared coefficient) drops significantly when heteroskedasticity exists. If the variance of the dependent variable is not constant, then the error's variance will not be constant. Unless the heteroskedasticity of the dependent variable is pronounced, its effect will not be severe: the least-squares estimates will still be unbiased, and the estimates of the slope and intercept will either be normally distributed if the errors are normally distributed, or at least normally distributed asymptotically (as the number of data points becomes large) if the errors are not normally distributed. The estimate for the variance of the slope and overall variance will be inaccurate, but the inaccuracy is not likely to be substantial if the independent-variable values are symmetric about their mean.

If the number of data points is small (micronumerosity), it may be difficult to detect assumption violations. With small sample sizes, assumption violations such as non-normality or heteroskedasticity of variances are difficult to detect even when they are present. With a small number of data points, linear regression offers less protection against violation of assumptions. With few data points, it may be hard to determine how well the fitted line matches the data, or whether a nonlinear function would be more appropriate. Even if none of the test assumptions are violated, a linear regression on a small number of data points may not have sufficient power to detect a significant difference between the slope and zero, even if the slope is nonzero. The power depends on the residual error, the observed variation in the independent variable, the selected significance alpha level of the test, and the number of data points. Power decreases as the residual variance increases, decreases as the significance level is decreased (i.e., as the test is made more stringent), increases as the variation in observed independent variable increases, and increases as the number of data points increases.

Values may not be identically distributed because of the presence of outliers. Outliers are anomalous values in the data. Outliers may have a strong influence over the fitted slope and intercept, giving a poor fit to the bulk of the data points. Outliers tend to increase the estimate of residual variance, lowering the chance of rejecting the null hypothesis, i.e., creating higher prediction errors. They may be due to recording errors, which may be correctable, or they may be due to the dependent-variable values not all being sampled from the same population. Apparent outliers may also be due to the dependent-variable values being from the same, but non-normal, population. However, a point may be an unusual value in either an independent or dependent variable without necessarily being an outlier in the scatter plot. In regression analysis, the fitted line can be highly sensitive to outliers. In other words, least squares regression is not resistant to outliers, thus, neither is the fitted-slope estimate. A point vertically removed from the other points can cause the fitted line to pass close to it, instead of following the general linear

trend of the rest of the data, especially if the point is relatively far horizontally from the center of the data.

However, great care should be taken when deciding if the outliers should be removed. Although in most cases when outliers are removed, the regression results look better, a priori justification must first exist. For instance, if one is regressing the performance of a particular firm's stock returns, outliers caused by downturns in the stock market should be included; these are not truly outliers as they are inevitabilities in the business cycle. Forgoing these outliers and using the regression equation to forecast one's retirement fund based on the firm's stocks will yield incorrect results at best. In contrast, suppose the outliers are caused by a single nonrecurring business condition (e.g., merger and acquisition) and such business structural changes are not forecast to recur, then these outliers should be removed and the data cleansed prior to running a regression analysis. The analysis here only identifies outliers and it is up to the user to determine if they should remain or be excluded.

Sometimes, a nonlinear relationship between the dependent and independent variables is more appropriate than a linear relationship. In such cases, running a linear regression will not be optimal. If the linear model is not the correct form, then the slope and intercept estimates and the fitted values from the linear regression will be biased, and the fitted slope and intercept estimates will not be meaningful. Over a restricted range of independent or dependent variables, nonlinear models may be well approximated by linear models (this is in fact the basis of linear interpolation), but for accurate prediction a model appropriate to the data should be selected. A nonlinear transformation should first be applied to the data before running a regression. One simple approach is to take the natural logarithm of the independent variable (other approaches include taking the square root or raising the independent variable to the second or third power) and run a regression or forecast using the nonlinearly-transformed data.

Diagnostic Results

	Heteroskedasticity		Micronumerosity	Outliers			Nonlinearity	
Variable	W-Test p-value	Hypothesis Test result	Approximation result	Natural Lower Bound	Natural Upper Bound	Number of Potential Outliers	Nonlinear Test p-value	Hypothesis Test result
Y			no problems	-7.86	671.70	**2**		
Variable X1	0.2543	Homoskedastic	no problems	-21377.95	64713.03	**3**	0.2458	linear
Variable X2	0.3371	Homoskedastic	no problems	77.47	445.93	**2**	**0.0335**	**nonlinear**
Variable X3	0.3649	Homoskedastic	no problems	-5.77	15.69	**3**	**0.0305**	**nonlinear**
Variable X4	0.3066	Homoskedastic	no problems	-295.96	628.21	**4**	0.9298	linear
Variable X5	0.2495	Homoskedastic	no problems	3.35	9.38	**3**	0.2727	linear

Figure 39 – Results from Tests of Outliers, Heteroskedasticity, Micronumerosity, and Nonlinearity

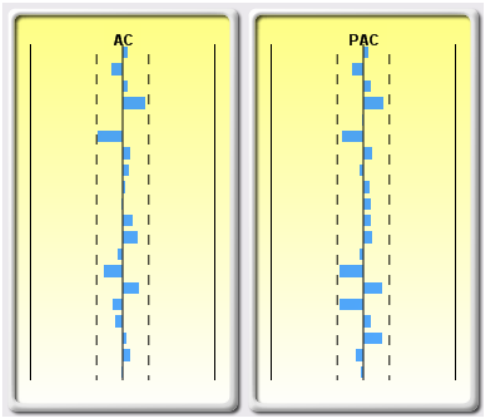
Another typical issue when forecasting time-series data is whether the independent-variable values are truly independent of each other or are they dependent. Dependent variable values collected over a time-series may be autocorrelated. For serially correlated dependent-variable values, the estimates of the slope and intercept will be unbiased, but the estimates of their forecast and variances will not be reliable and hence the validity of certain statistical goodness-of-fit tests will be flawed. For instance, interest rates, inflation rates, sales, revenues, and many other time-series data are typically autocorrelated, where the value in the current period is related to the value in a previous period, and so forth (clearly, the inflation rate in March is related to February's level, which in turn, is related to January's level, and so forth). Ignoring such blatant relationships will yield biased and less accurate forecasts. In such events, an autocorrelated regression model or an ARIMA model may be better suited (**Risk Simulator | Forecasting | ARIMA**). Finally, the autocorrelation functions of a series that is nonstationary tend to decay slowly (see Nonstationary report in the model).

If autocorrelation AC(*1*) is nonzero, it means that the series is first order serially correlated. If AC(*k*) dies off more or less geometrically with increasing lag, it implies that the series follows a low-order autoregressive process. If AC(*k*) drops to zero after a small number of lags, it implies that the series follows a low-order moving-average process. Partial correlation PAC(*k*) measures the correlation of values that are *k* periods apart after removing the correlation from the intervening lags. If the pattern of autocorrelation can be captured by an autoregression of order less than *k*, then the partial autocorrelation at lag k will be close to zero. Ljung-Box Q-statistics and their p-values at lag *k* has the null hypothesis that there is no autocorrelation up to order k. The dotted lines in the plots of the autocorrelations are the approximate two standard error bounds. If the autocorrelation is within these bounds, it is not significantly different from zero at the 5% significance level.

Autocorrelation measures the relationship to the past of the dependent *Y* variable to itself. Distributive Lags, in contrast, are time-lag relationships between the dependent *Y* variable and different independent *X* variables. For instance, the movement and direction of mortgage rates tend to follow the Federal Funds Rate but at a time lag (typically 1 to 3 months). Sometimes, time lags follow cycles and seasonality (e.g., ice cream sales tend to peak during the summer months and are hence related to last summer's sales, 12 months in the past). The distributive lag analysis (Figure 40) shows how the dependent variable is related to each of the independent variables at various time lags, when all lags are considered simultaneously, to determine which time lags are statistically significant and should be considered.

Autocorrelation

Time Lag	AC	PAC	Lower Bound	Upper Bound	Q-Stat	Prob
1	0.0580	0.0580	-0.2828	0.2828	0.1786	0.6726
2	-0.1213	-0.1251	-0.2828	0.2828	0.9754	0.6140
3	0.0590	0.0756	-0.2828	0.2828	1.1679	0.7607
4	0.2423	0.2232	-0.2828	0.2828	4.4865	0.3442
5	0.0067	-0.0078	-0.2828	0.2828	4.4890	0.4814
6	-0.2654	-0.2345	-0.2828	0.2828	8.6516	0.1941
7	0.0814	0.0939	-0.2828	0.2828	9.0524	0.2489
8	0.0634	-0.0442	-0.2828	0.2828	9.3012	0.3175
9	0.0204	0.0673	-0.2828	0.2828	9.3276	0.4076
10	-0.0190	0.0865	-0.2828	0.2828	9.3512	0.4991
11	0.1035	0.0790	-0.2828	0.2828	10.0648	0.5246
12	0.1658	0.0978	-0.2828	0.2828	11.9466	0.4500
13	-0.0524	-0.0430	-0.2828	0.2828	12.1394	0.5162
14	-0.2050	-0.2523	-0.2828	0.2828	15.1738	0.3664
15	0.1782	0.2089	-0.2828	0.2828	17.5315	0.2881
16	-0.1022	-0.2591	-0.2828	0.2828	18.3296	0.3050
17	-0.0861	0.0808	-0.2828	0.2828	18.9141	0.3335
18	0.0418	0.1987	-0.2828	0.2828	19.0559	0.3884
19	0.0869	-0.0821	-0.2828	0.2828	19.6894	0.4135
20	-0.0091	-0.0269	-0.2828	0.2828	19.6966	0.4770



Distributive Lags

P-Values of Distributive Lag Periods of Each Independent Variable

Variable	1	2	3	4	5	6	7	8	9	10	11	12
X1	0.8467	0.2045	0.3336	0.9105	0.9757	0.1020	0.9205	0.1267	0.5431	0.9110	0.7495	0.4016
X2	0.6077	0.9900	0.8422	0.2851	**0.0638**	**0.0032**	0.8007	0.1551	0.4823	0.1126	**0.0519**	0.4383
X3	0.7394	0.2396	0.2741	0.8372	0.9808	**0.0464**	0.8355	**0.0545**	0.6828	0.7354	0.5093	0.3500
X4	**0.0061**	0.6739	0.7932	0.7719	0.6748	0.8627	0.5586	0.9046	0.5726	0.6304	0.4812	0.5707
X5	0.1591	0.2032	0.4123	0.5599	0.6416	0.3447	0.9190	0.9740	0.5185	0.2856	0.1489	0.7794

Figure 40 – Autocorrelation and Distributive Lag Results

Another requirement in running a regression model is the assumption of normality and sphericity of the error term. If the assumption of normality is violated or outliers are present, then the linear regression goodness-of-fit test may not be the most powerful or informative test available, and this could mean the difference between detecting a linear fit or not. If the errors are not independent and not normally distributed, it may indicate that the data might be autocorrelated or suffer from nonlinearities or other more destructive errors. Independence of the errors can also be detected in the heteroskedasticity tests (Figure 41).

The Normality test on the errors performed is a nonparametric test, which makes no assumptions about the specific shape of the population from which the sample is drawn, allowing for smaller sample data sets to be analyzed. This test evaluates the null hypothesis of whether the sample errors were drawn from a normally distributed population, versus an alternate hypothesis that the data sample is not normally distributed. If the calculated D-Statistic is greater than or equal to the D-Critical values at various significance values then reject the null hypothesis and accept the alternate hypothesis (the errors are not normally distributed). Otherwise, if the D-Statistic is less than the D-Critical value, do not reject the null hypothesis (the errors are normally distributed). This test relies on two cumulative frequencies: one

derived from the sample data set, the second from a theoretical distribution based on the mean and standard deviation of the sample data.

Test Result

Regression Error Average	0.00
Standard Deviation of Errors	141.83
D Statistic	0.1036
D Critical at 1%	0.1138
D Critical at 5%	0.1225
D Critical at 10%	0.1458

Null Hypothesis: The errors are normally distributed.

Conclusion: The errors are normally distributed at the 1% alpha level.

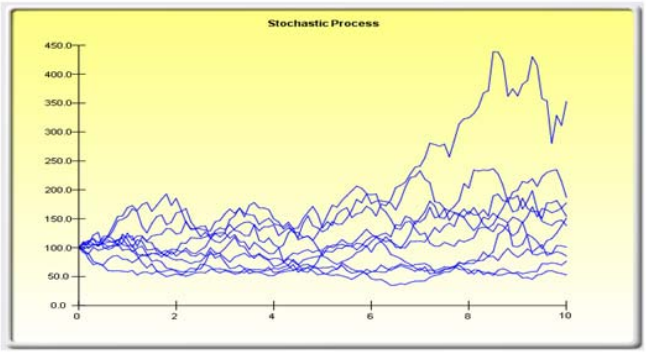
Errors	Relative Frequency	Observed	Expected	O-E
-219.04	0.02	0.02	0.0612	-0.0412
-202.53	0.02	0.04	0.0766	-0.0366
-186.04	0.02	0.06	0.0948	-0.0348
-174.17	0.02	0.08	0.1097	-0.0297
-162.13	0.02	0.10	0.1265	-0.0265
-161.62	0.02	0.12	0.1272	-0.0072
-160.39	0.02	0.14	0.1291	0.0109
-145.40	0.02	0.16	0.1526	0.0074
-138.92	0.02	0.18	0.1637	0.0163
-133.81	0.02	0.20	0.1727	0.0273
-120.76	0.02	0.22	0.1973	0.0227
-120.12	0.02	0.24	0.1985	0.0415

Figure 41 – Test for Normality of Errors

Sometimes, certain types of time-series data cannot be modeled using any other methods except for a stochastic process, because the underlying events are stochastic in nature. For instance, you cannot adequately model and forecast stock prices, interest rates, price of oil, and other commodity prices using a simple regression model, because these variables are highly uncertain and volatile, and does not follow a predefined static rule of behavior, in other words, the process is not stationary. Stationarity is checked here using the Runs Test while another visual clue is found in the Autocorrelation report (the ACF tends to decay slowly). A stochastic process is a sequence of events or paths generated by probabilistic laws. That is, random events can occur over time but are governed by specific statistical and probabilistic rules. The main stochastic processes include Random Walk or Brownian Motion, Mean-Reversion, and Jump-Diffusion. These processes can be used to forecast a multitude of variables that seemingly follow random trends but restricted by probabilistic laws. The process-generating equation is known in advance but the actual results generated is unknown (Figure 42).

The Random Walk Brownian Motion process can be used to forecast stock prices, prices of commodities, and other stochastic time-series data given a drift or growth rate and volatility around the drift path. The Mean-Reversion process can be used to reduce the fluctuations of the Random Walk process by allowing the path to target a long-term value, making it useful for forecasting time-series variables that have a long-term rate such as interest rates and inflation rates (these are long-term target rates by regulatory authorities or the market). The Jump-Diffusion process is useful for forecasting time-series data when the variable can occasionally exhibit random jumps, such as oil prices or price of electricity (discrete

exogenous event shocks can make prices jump up or down). These processes can also be mixed and matched as required.



Statistical Summary

The following are the estimated parameters for a stochastic process given the data provided. It is up to you to determine if the probability of fit (similar to a goodness-of-fit computation) is sufficient to warrant the use of a stochastic process forecast, and if so, whether it is a random walk, mean-reversion, or a jump-diffusion model, or combinations thereof. In choosing the right stochastic process model, you will have to rely on past experiences and *a priori* economic and financial expectations of what the underlying data set is best represented by. These parameters can be entered into a stochastic process forecast (**Simulation** | **Forecasting** | **Stochastic Processes**).

Periodic

Drift Rate	*-1.48%*	*Reversion Rate*	*283.89%*	*Jump Rate*	*20.41%*
Volatility	*88.84%*	*Long-Term Value*	*327.72*	*Jump Size*	*237.89*

Probability of stochastic model fit: 46.48%
A high fit means a stochastic model is better than conventional models.

Runs	*20*	*Standard Normal*	*-1.7321*
Positive	*25*	*P-Value (1-tail)*	*0.0416*
Negative	*25*	*P-Value (2-tail)*	*0.0833*
Expected Run	*26*		

A low p-value (below 0.10, 0.05, 0.01) means that the sequence is not random and hence suffers from stationarity problems, and an ARIMA model might be more appropriate. Conversely, higher p-values indicate randomness and stochastic process models might be appropriate.

Figure 42 – Stochastic Process Parameter Estimation

Multicollinearity exists when there is a linear relationship between the independent variables. When this occurs, the regression equation cannot be estimated at all. In near collinearity situations, the estimated regression equation will be biased and provide inaccurate results. This situation is especially true when a step-wise regression approach is used, where the statistically significant independent variables will be thrown out of the regression mix earlier than expected, resulting in a regression equation that is neither efficient nor accurate. One quick test of the presence of multicollinearity in a multiple regression equation is that the R-squared value is relatively high while the t-statistics are relatively low.

Another quick test is to create a correlation matrix between the independent variables. A high cross-correlation indicates a potential for autocorrelation. The rule of thumb is that a correlation with an absolute value greater than 0.75 is indicative of severe multicollinearity. Another test for multicollinearity is the use of the Variance Inflation Factor (VIF), obtained by regressing each independent variable to all the other independent variables, obtaining the R-squared value and calculating the VIF. A VIF exceeding

2.0 can be considered as severe multicollinearity. A VIF exceeding 10.0 indicates destructive multicollinearity (Figure 43).

Correlation Matrix

CORRELATION	X2	X3	X4	X5
X1	0.333	0.959	0.242	0.237
X2	1.000	0.349	0.319	0.120
X3		1.000	0.196	0.227
X4			1.000	0.290

Variance Inflation Factor

VIF	X2	X3	X4	X5
X1	1.12	**12.46**	1.06	1.06
X2	N/A	1.14	1.11	1.01
X3		N/A	1.04	1.05
X4			N/A	1.09

Figure 43 – Multicollinearity Errors

The Correlation Matrix lists the Pearson's Product Moment Correlations (commonly referred to as the Pearson's R) between variable pairs. The correlation coefficient ranges between –1.0 and + 1.0 inclusive. The sign indicates the direction of association between the variables while the coefficient indicates the magnitude or strength of association. The Pearson's R only measures a linear relationship, and is less effective in measuring non-linear relationships.

To test whether the correlations are significant, a two-tailed hypothesis test is performed and the resulting p-values are listed above. P-values less than 0.10, 0.05, and 0.01 are highlighted in blue to indicate statistical significance. In other words, a p-value for a correlation pair that is less than a given significance value is statistically significantly different from zero, indicating that there is significant a linear relationship between the two variables.

The Pearson's Product Moment Correlation Coefficient (R) between two variables (x and y) is related to the covariance (cov) measure where $R_{x,y} = \frac{COV_{x,y}}{s_x s_y}$. The benefit of dividing the covariance by the product of the two variables' standard deviation (s) is that the resulting correlation coefficient is bounded between –1.0 and +1.0 inclusive. This makes the correlation a good relative measure to compare among different variables (particularly with different units and magnitude). The Spearman rank-based nonparametric

correlation is also included below. The Spearman's R is related to the Pearson's R in that the data is first ranked and then correlated. The rank correlations provide a better estimate of the relationship between two variables when one or both of them is nonlinear.

It must be stressed that a significant correlation does not imply causation. Associations between variables in no way imply that the change of one variable causes another variable to change. When two variables that are moving independently of each other but in a related path, they may be correlated but their relationship might be spurious (e.g., a correlation between sunspots and the stock market might be strong but one can surmise that there is no causality and that this relationship is purely spurious).

Statistical Analysis Tool

Another very powerful tool in Risk Simulator is the Statistical Analysis Tool, which determines the statistical properties of the data. The diagnostics run include checking the data for various statistical properties, from basic descriptive statistics to testing for and calibrating the stochastic properties of the data.

Procedure:

- Open the example model (**Risk Simulator | Examples | Statistical Analysis**) and go to the *Data* worksheet and select the data including the variable names (cells **C5:E55**).
- Click on **Risk Simulator** | **Tools** | **Statistical Analysis** (Figure 44).
- Check the *data type*, whether the data selected is from a single variable or multiple variables arranged in rows. In our example, we assume that the data areas selected are from multiple variables. Click **OK** when finished.
- *Choose the statistical tests* you wish to perform. The suggestion (and by default) is to choose all the tests. Click **OK** when finished (Figure 45).

Spend some time going through the reports generated to get a better understanding of the statistical tests performed (sample reports are shown in Figures 46-49).

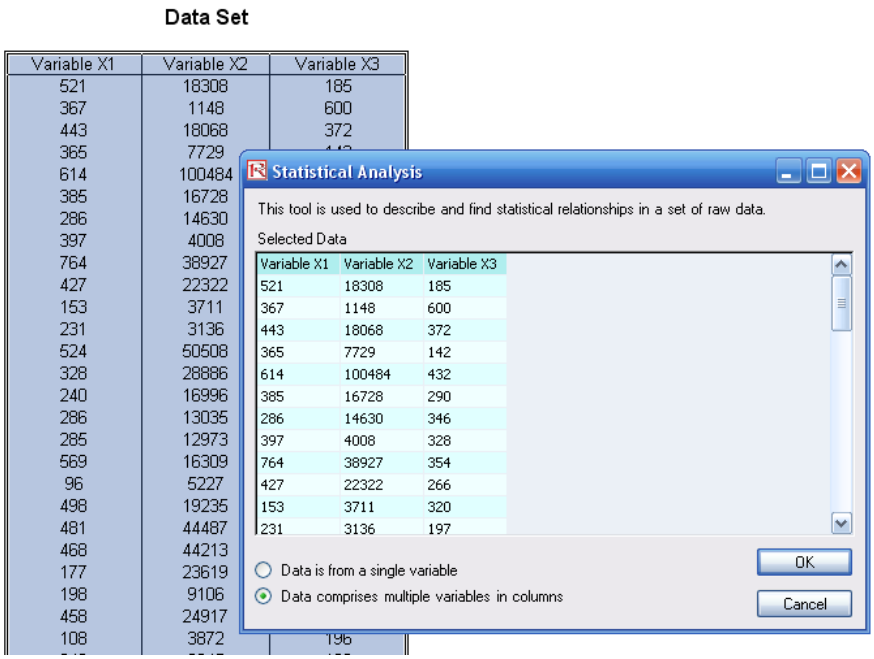


Figure 44 – Running the Statistical Analysis Tool

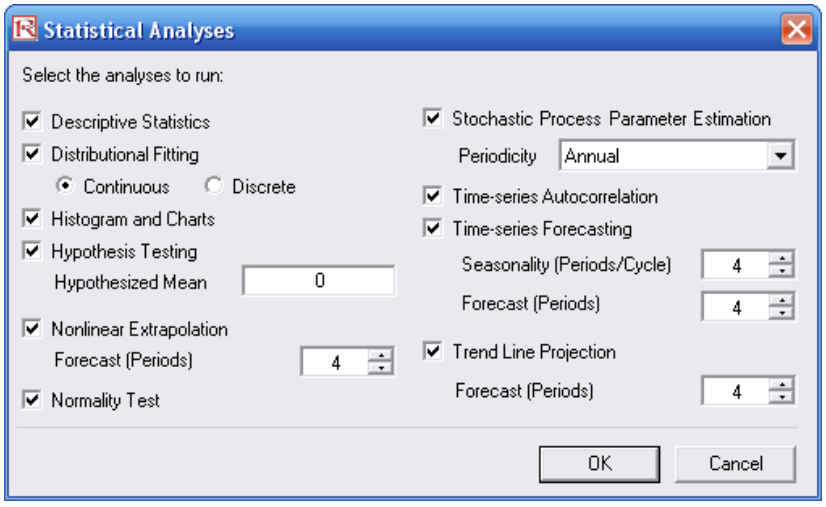


Figure 45 – Statistical Tests

Descriptive Statistics

Analysis of Statistics

Almost all distributions can be described within 4 moments (some distributions require one moment, while others require two moments, and so forth). Descriptive statistics quantitatively capture these moments. The first moment describes the location of a distribution (i.e., mean, median, and mode) and is interpreted as the expected value, expected returns, or the average value of occurrences.

The Arithmetic Mean calculates the average of all occurrences by summing up all of the data points and dividing them by the number of points. The Geometric Mean is calculated by taking the power root of the products of all the data points and requires them to all be positive. The Geometric Mean is more accurate for percentages or rates that fluctuate significantly. For example, you can use Geometric Mean to calculate average growth rate given compound interest with variable rates. The Trimmed Mean calculates the arithmetic average of the data set after the extreme outliers have been trimmed. As averages are prone to significant bias when outliers exist, the Trimmed Mean reduces such bias in skewed distributions.

The Standard Error of the Mean calculates the error surrounding the sample mean. The larger the sample size, the smaller the error such that for an infinitely large sample size, the error approaches zero, indicating that the population parameter has been estimated. Due to sampling errors, the 95% Confidence Interval for the Mean is provided. Based on an analysis of the sample data points, the actual population mean should fall between these Lower and Upper Intervals for the Mean.

Median is the data point where 50% of all data points fall above this value and 50% below this value. Among the three first moment statistics, the median is least susceptible to outliers. A symmetrical distribution has the Median equal to the Arithmetic Mean. A skewed distribution exists when the Median is far away from the Mean. The Mode measures the most frequently occurring data point.

Minimum is the smallest value in the data set while Maximum is the largest value. Range is the difference between the Maximum and Minimum values.

The second moment measures a distribution's spread or width, and is frequently described using measures such as Standard Deviations, Variances, Quartiles, and Inter-Quartile Ranges. Standard Deviation indicates the average deviation of all data points from their mean. It is a popular measure as is associated with risk (higher standard deviations mean a wider distribution, higher risk, or wider dispersion of data points around the mean) and its units are identical to original data set's. The Sample Standard Deviation differs from the Population Standard Deviation in that the former uses a degree of freedom correction to account for small sample sizes. Also, Lower and Upper Confidence Intervals are provided for the Standard Deviation and the true population standard deviation falls within this interval. If your data set covers every element of the population, use the Population Standard Deviation instead. The two Variance measures are simply the squared values of the standard deviations.

The Coefficient of Variability is the standard deviation of the sample divided by the sample mean, proving a unit-free measure of dispersion that can be compared across different distributions (you can now compare distributions of values denominated in millions of dollars with one in billions of dollars, or meters and kilograms, etc.). The First Quartile measures the 25th percentile of the data points when arranged from its smallest to largest value. The Third Quartile is the value of the 75th percentile data point. Sometimes quartiles are used as the upper and lower ranges of a distribution as it truncates the data set to ignore outliers. The Inter-Quartile Range is the difference between the third and first quartiles, and is often used to measure the width of the center of a distribution.

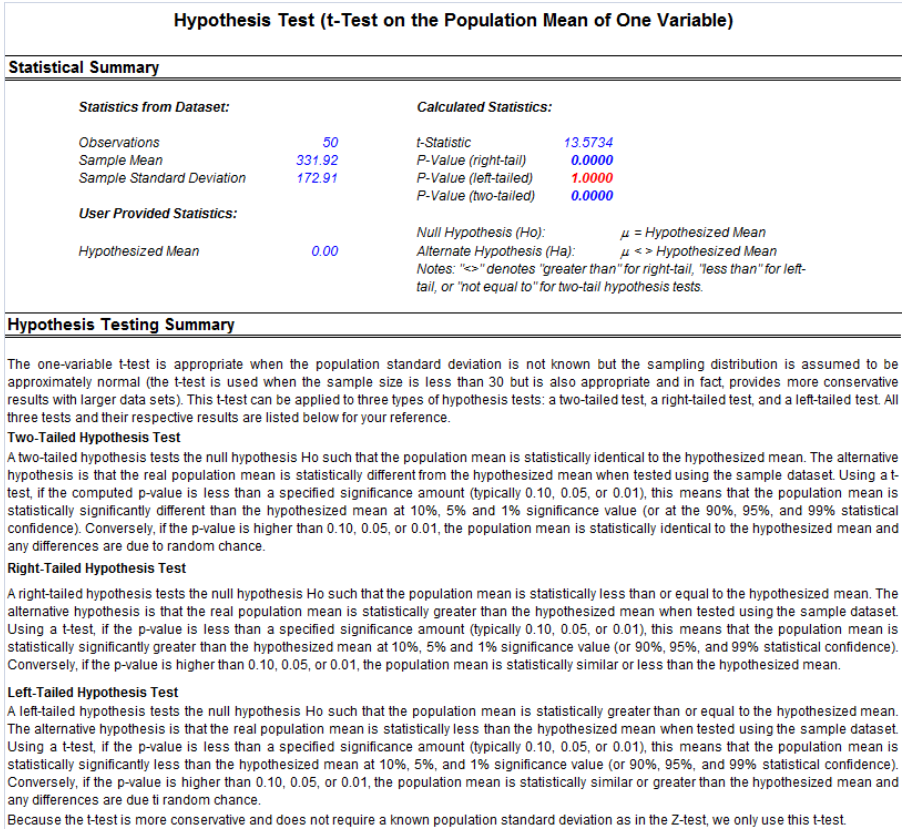
Skewness is the third moment in a distribution. Skewness characterizes the degree of asymmetry of a distribution around its mean. Positive skewness indicates a distribution with an asymmetric tail extending toward more positive values. Negative skewness indicates a distribution with an asymmetric tail extending toward more negative values.

Kurtosis characterizes the relative peakedness or flatness of a distribution compared to the normal distribution. It is the fourth moment in a distribution. A positive Kurtosis value indicates a relatively peaked distribution. A negative kurtosis indicates a relatively flat distribution. The Kurtosis measured here has been centered to zero (certain other kurtosis measures are centered around 3.0). While both are equally valid, centering across zero makes the interpretation simpler. A high positive Kurtosis indicates a peaked distribution around its center and leptokurtic or fat tails. This indicates a higher probability of extreme events (e.g., catastrophic events, terrorist attacks, stock market crashes) than is predicted in a normal distribution.

Summary Statistics

Statistics	*Variable X1*		
Observations	*50.0000*	*Standard Deviation (Sample)*	*172.9140*
Arithmetic Mean	*331.9200*	*Standard Deviation (Population)*	*171.1761*
Geometric Mean	*281.3247*	*Lower Confidence Interval for Standard Deviation*	*148.6090*
Trimmed Mean	*325.1739*	*Upper Confidence Interval for Standard Deviation*	*207.7947*
Standard Error of Arithmetic Mean	*24.4537*	*Variance (Sample)*	*29899.2588*
Lower Confidence Interval for Mean	*283.0125*	*Variance (Population)*	*29301.2736*
Upper Confidence Interval for Mean	*380.8275*	*Coefficient of Variability*	*0.5210*
Median	*307.0000*	*First Quartile (Q1)*	*204.0000*
Mode	*47.0000*	*Third Quartile (Q3)*	*441.0000*
Minimum	*764.0000*	*Inter-Quartile Range*	*237.0000*
Maximum	*717.0000*	*Skewness*	*0.4838*
Range		*Kurtosis*	*-0.0952*

Figure 46 – Sample Statistical Analysis Tool Report



Hypothesis Test (t-Test on the Population Mean of One Variable)

Statistical Summary

Statistics from Dataset:		*Calculated Statistics:*	
Observations	*50*	*t-Statistic*	*13.5734*
Sample Mean	*331.92*	*P-Value (right-tail)*	***0.0000***
Sample Standard Deviation	*172.91*	*P-Value (left-tailed)*	***1.0000***
		P-Value (two-tailed)	***0.0000***
User Provided Statistics:			
		Null Hypothesis (Ho):	*μ = Hypothesized Mean*
Hypothesized Mean	*0.00*	*Alternate Hypothesis (Ha):*	*μ < > Hypothesized Mean*

Notes: "<>" denotes "greater than" for right-tail, "less than" for left-tail, or "not equal to" for two-tail hypothesis tests.

Hypothesis Testing Summary

The one-variable t-test is appropriate when the population standard deviation is not known but the sampling distribution is assumed to be approximately normal (the t-test is used when the sample size is less than 30 but is also appropriate and in fact, provides more conservative results with larger data sets). This t-test can be applied to three types of hypothesis tests: a two-tailed test, a right-tailed test, and a left-tailed test. All three tests and their respective results are listed below for your reference.

Two-Tailed Hypothesis Test

A two-tailed hypothesis tests the null hypothesis Ho such that the population mean is statistically identical to the hypothesized mean. The alternative hypothesis is that the real population mean is statistically different from the hypothesized mean when tested using the sample dataset. Using a t-test, if the computed p-value is less than a specified significance amount (typically 0.10, 0.05, or 0.01), this means that the population mean is statistically significantly different than the hypothesized mean at 10%, 5% and 1% significance value (or at the 90%, 95%, and 99% statistical confidence). Conversely, if the p-value is higher than 0.10, 0.05, or 0.01, the population mean is statistically identical to the hypothesized mean and any differences are due to random chance.

Right-Tailed Hypothesis Test

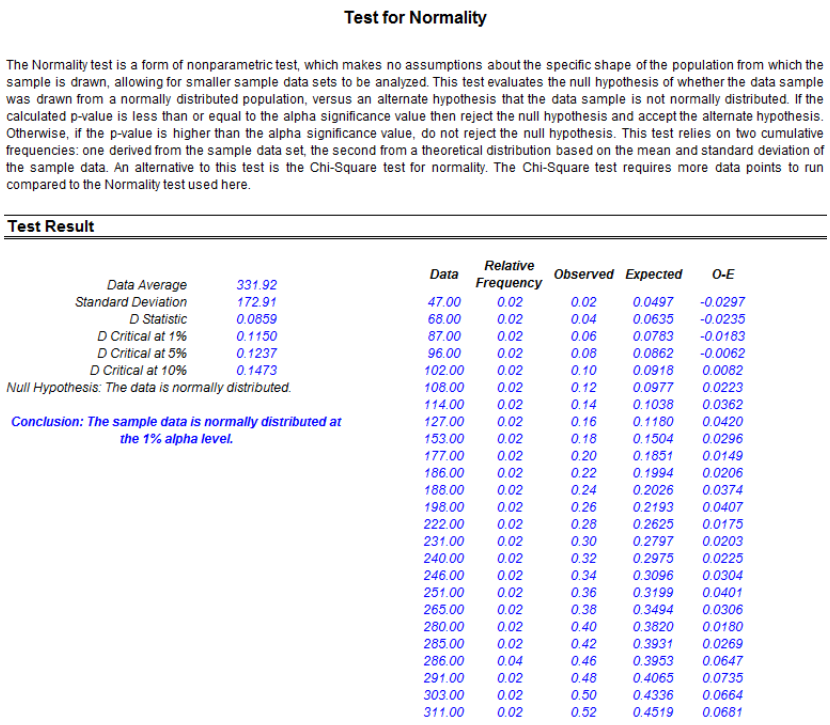
A right-tailed hypothesis tests the null hypothesis Ho such that the population mean is statistically less than or equal to the hypothesized mean. The alternative hypothesis is that the real population mean is statistically greater than the hypothesized mean when tested using the sample dataset. Using a t-test, if the p-value is less than a specified significance amount (typically 0.10, 0.05, or 0.01), this means that the population mean is statistically significantly greater than the hypothesized mean at 10%, 5% and 1% significance value (or 90%, 95%, and 99% statistical confidence). Conversely, if the p-value is higher than 0.10, 0.05, or 0.01, the population mean is statistically similar or less than the hypothesized mean.

Left-Tailed Hypothesis Test

A left-tailed hypothesis tests the null hypothesis Ho such that the population mean is statistically greater than or equal to the hypothesized mean. The alternative hypothesis is that the real population mean is statistically less than the hypothesized mean when tested using the sample dataset. Using a t-test, if the p-value is less than a specified significance amount (typically 0.10, 0.05, or 0.01), this means that the population mean is statistically significantly less than the hypothesized mean at 10%, 5%, and 1% significance value (or 90%, 95%, and 99% statistical confidence). Conversely, if the p-value is higher than 0.10, 0.05, or 0.01, the population mean is statistically similar or greater than the hypothesized mean and any differences are due ti random chance.

Because the t-test is more conservative and does not require a known population standard deviation as in the Z-test, we only use this t-test.

Figure 47 – Sample Statistical Analysis Tool Report (Hypothesis Testing of One Variable)



Test for Normality

The Normality test is a form of nonparametric test, which makes no assumptions about the specific shape of the population from which the sample is drawn, allowing for smaller sample data sets to be analyzed. This test evaluates the null hypothesis of whether the data sample was drawn from a normally distributed population, versus an alternate hypothesis that the data sample is not normally distributed. If the calculated p-value is less than or equal to the alpha significance value then reject the null hypothesis and accept the alternate hypothesis. Otherwise, if the p-value is higher than the alpha significance value, do not reject the null hypothesis. This test relies on two cumulative frequencies: one derived from the sample data set, the second from a theoretical distribution based on the mean and standard deviation of the sample data. An alternative to this test is the Chi-Square test for normality. The Chi-Square test requires more data points to run compared to the Normality test used here.

Test Result

Data Average	*331.92*
Standard Deviation	*172.91*
D Statistic	*0.0859*
D Critical at 1%	*0.1150*
D Critical at 5%	*0.1237*
D Critical at 10%	*0.1473*

Null Hypothesis: The data is normally distributed.

Conclusion: The sample data is normally distributed at the 1% alpha level.

Data	*Relative Frequency*	*Observed*	*Expected*	*O-E*
47.00	*0.02*	*0.02*	*0.0497*	*-0.0297*
68.00	*0.02*	*0.04*	*0.0635*	*-0.0235*
87.00	*0.02*	*0.06*	*0.0783*	*-0.0183*
96.00	*0.02*	*0.08*	*0.0862*	*-0.0062*
102.00	*0.02*	*0.10*	*0.0918*	*0.0082*
108.00	*0.02*	*0.12*	*0.0977*	*0.0223*
114.00	*0.02*	*0.14*	*0.1038*	*0.0362*
127.00	*0.02*	*0.16*	*0.1180*	*0.0420*
153.00	*0.02*	*0.18*	*0.1504*	*0.0296*
177.00	*0.02*	*0.20*	*0.1851*	*0.0149*
186.00	*0.02*	*0.22*	*0.1994*	*0.0206*
188.00	*0.02*	*0.24*	*0.2026*	*0.0374*
198.00	*0.02*	*0.26*	*0.2193*	*0.0407*
222.00	*0.02*	*0.28*	*0.2625*	*0.0175*
231.00	*0.02*	*0.30*	*0.2797*	*0.0203*
240.00	*0.02*	*0.32*	*0.2975*	*0.0225*
246.00	*0.02*	*0.34*	*0.3096*	*0.0304*
251.00	*0.02*	*0.36*	*0.3199*	*0.0401*
265.00	*0.02*	*0.38*	*0.3494*	*0.0306*
280.00	*0.02*	*0.40*	*0.3820*	*0.0180*
285.00	*0.02*	*0.42*	*0.3931*	*0.0269*
286.00	*0.04*	*0.46*	*0.3953*	*0.0647*
291.00	*0.02*	*0.48*	*0.4065*	*0.0735*
303.00	*0.02*	*0.50*	*0.4336*	*0.0664*
311.00	*0.02*	*0.52*	*0.4519*	*0.0681*

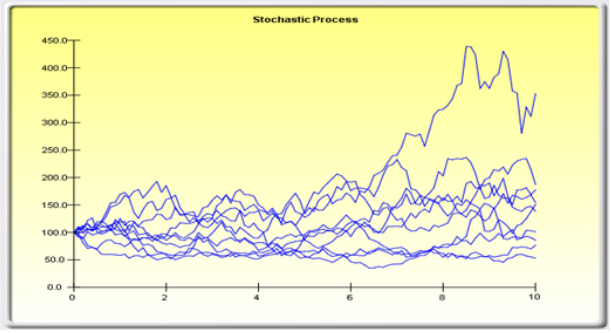
Figure 48 – Sample Statistical Analysis Tool Report (Normality Test)

Stochastic Process - Parameter Estimations

Statistical Summary

A stochastic process is a sequence of events or paths generated by probabilistic laws. That is, random events can occur over time but are governed by specific statistical and probabilistic rules. The main stochastic processes include Random Walk or Brownian Motion, Mean-Reversion, and Jump-Diffusion. These processes can be used to forecast a multitude of variables that seemingly follow random trends but yet are restricted by probabilistic laws. The process-generating equation is known in advance but the actual results generated is unknown.

The Random Walk Brownian Motion process can be used to forecast stock prices, prices of commodities, and other stochastic time-series data given a drift or growth rate and a volatility around the drift path. The Mean-Reversion process can be used to reduce the fluctuations of the Random Walk process by allowing the path to target a long-term value, making it useful for forecasting time-series variables that have a long-term rate such as interest rates and inflation rates (these are long-term target rates by regulatory authorities or the market). The Jump-Diffusion process is useful for forecasting time-series data when the variable can occasionally exhibit random jumps, such as oil prices or price of electricity (discrete exogenous event shocks can make prices jump up or down). Finally, these three stochastic processes can be mixed and matched as required.



Statistical Summary

The following are the estimated parameters for a stochastic process given the data provided. It is up to you to determine if the probability of fit (similar to a goodness-of-fit computation) is sufficient to warrant the use of a stochastic process forecast, and if so, whether it is a random walk, mean-reversion, or a jump-diffusion model, or combinations thereof. In choosing the right stochastic process model, you will have to rely on past experiences and a priori economic and financial expectations of what the underlying data set is best represented by. These parameters can be entered into a stochastic process forecast (**Simulation | Forecasting | Stochastic Processes**).

(Annualized)

Drift Rate	*5.86%*	*Reversion Rate*	*N/A*	*Jump Rate*	*16.33%*
Volatility	*7.04%*	*Long-Term Value*	*N/A*	*Jump Size*	*21.33*

Probability of stochastic model fit: *4.63%*

Figure 49 – Sample Statistical Analysis Tool Report (Stochastic Parameter Estimation)

Distributional Analysis Tool

This is a statistical probability tool in Risk Simulator that is rather useful in a variety of settings, and can be used to compute the probability density function (PDF), which is also called the probability mass function (PMF) for discrete distributions (we will use these terms interchangeably), where given some distribution and its parameters, we can determine the probability of occurrence given some outcome *x*. In addition, the cumulative distribution function (CDF) can also be computed, which is the sum of the PDF values up to this *x* value. Finally, the inverse cumulative distribution function (ICDF) is used to compute the value *x* given the probability of occurrence.

This tool is accessible via **Risk Simulator | Tools | Distributional Analysis**. As an example, Figure 50 shows the computation of a binomial distribution (i.e., a distribution with two outcomes, such as the tossing of a coin, where the outcome is either Heads or Tails, with some prescribed probability of heads

and tails). Suppose we toss a coin two times, and set the outcome Heads as a success, we use the binomial distribution with Trials = 2 (tossing the coin twice) and Probability = 0.50 (the probability of success, of getting Heads). Selecting the PDF and setting the range of values *x* as from 0 to 2 with a step size of 1 (this means we are requesting the values 0, 1, 2 for *x*), the resulting probabilities are provided in the table and graphically, as well as the theoretical four moments of the distribution. As the outcomes of the coin toss is Heads-Heads, Tails-Tails, Heads-Tails, and Tails-Heads, the probability of getting exactly no Heads is 25%, one Heads is 50%, and two Heads is 25%.

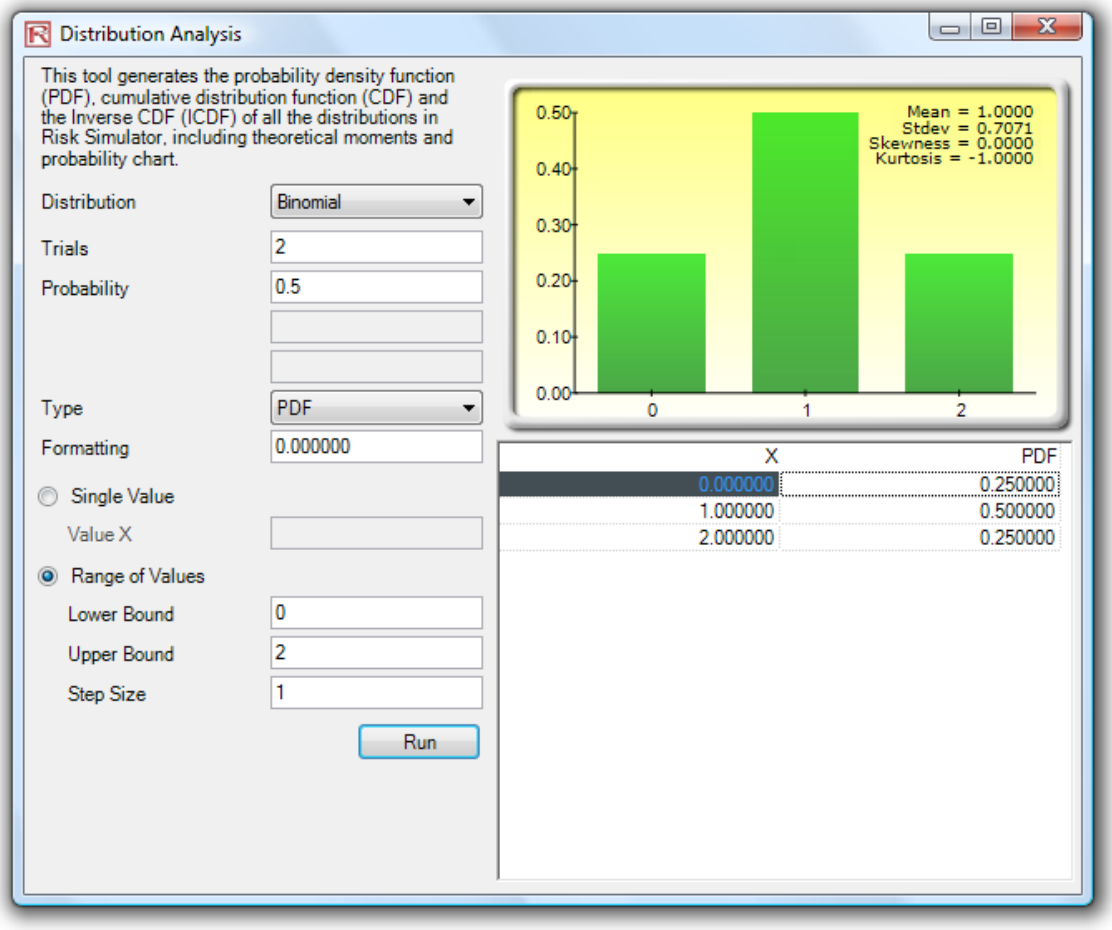


Figure 50 – Distributional Analysis Tool (Binomial Distribution with 2 Trials)

Similarly, we can obtain the exact probabilities of tossing the coin, say 20 times, as seen in Figure 51. The results are presented both in table and graphical formats.

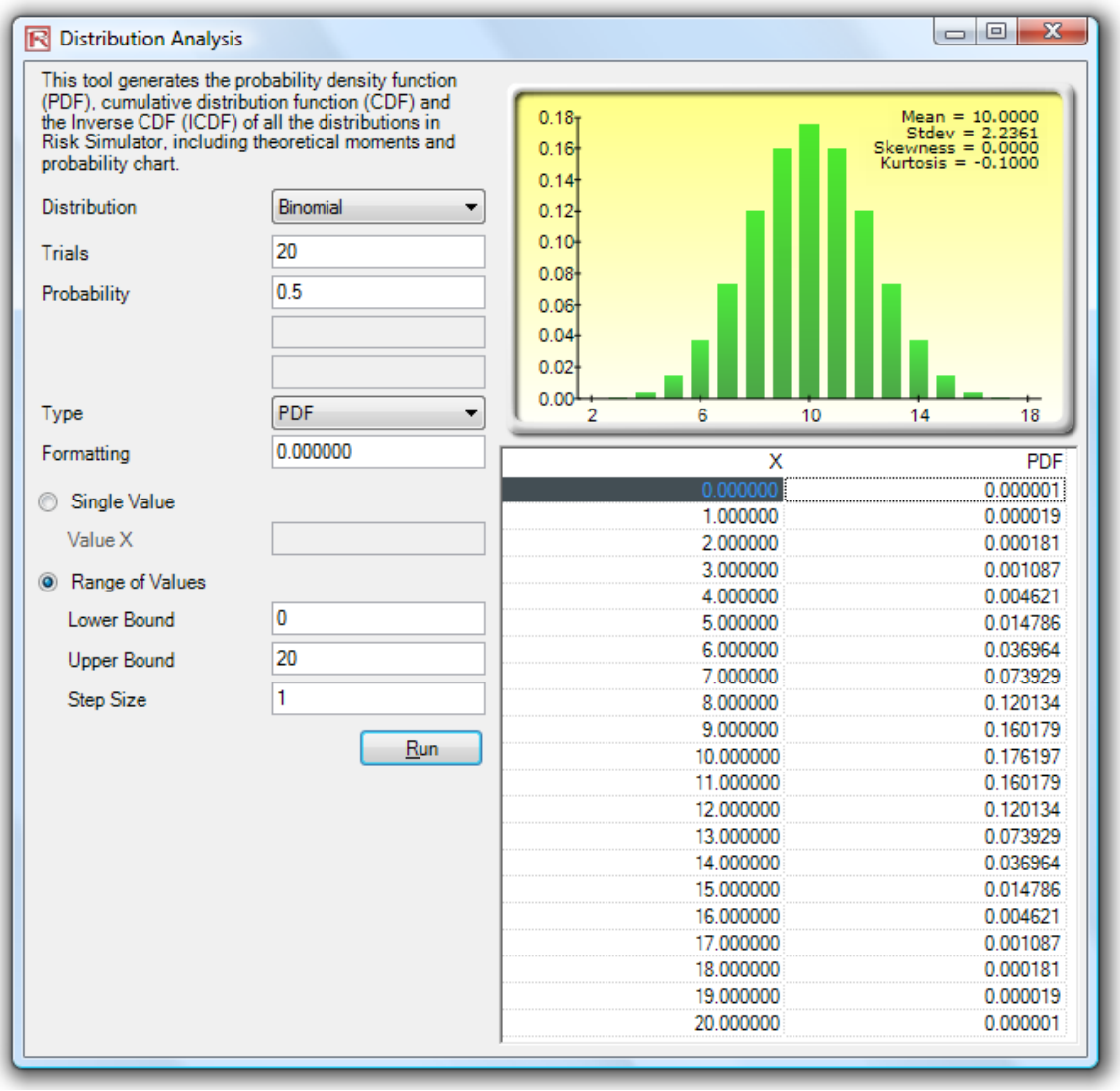


Figure 51 – Distributional Analysis Tool (Binomial Distribution with 20 Trials)

Figure 51 shows the same binomial distribution but now the CDF is computed. The CDF is simply the sum of the PDF values up to the point *x*. For instance, in Figure 51, we see that the probabilities of 0, 1, and 2 are 0.000001, 0.000019, and 0.000181, whose sum is 0.000201, which is the value of the CDF at *x* = 2 in Figure 52. Whereas the PDF computes the probabilities of getting 2 heads, the CDF computes the probability of getting no more than 2 heads (or probabilities of 0, 1, and 2 heads). Taking the complement (i.e., 1 – 0.00021 obtains 0.999799 or 99.9799%) provides the probability of getting at least 3 heads or more.

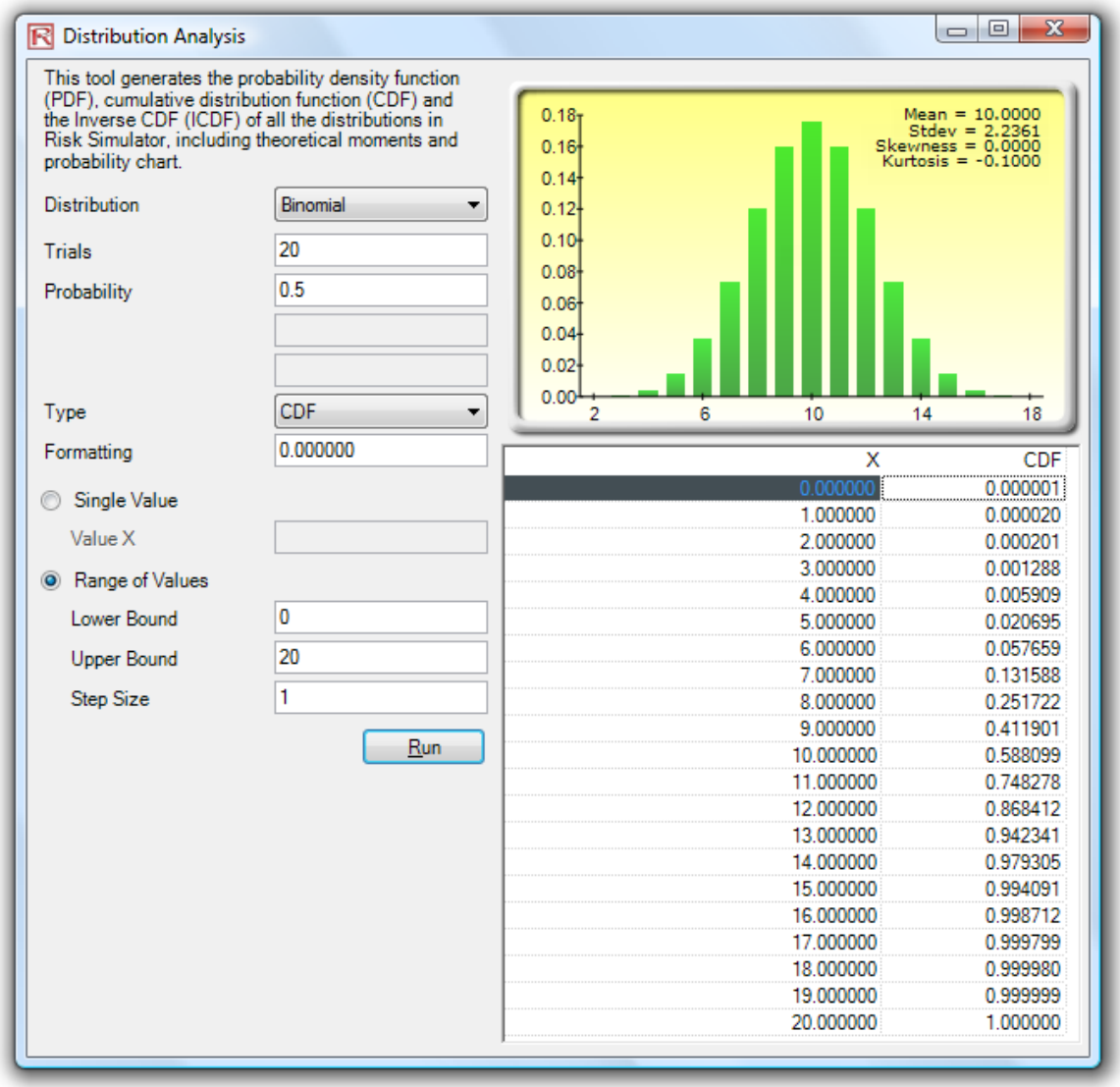


Figure 52 – Distributional Analysis Tool (Binomial Distribution's CDF with 20 Trials)

Using this Distributional Analysis tool, even more advanced distributions can be analyzed, such as the gamma, beta, negative binomial, and many others in Risk Simulator. As further example of the tool's use in a continuous distribution and the ICDF functionality, Figure 53 shows the standard normal distribution (normal distribution with a mean of zero and standard deviation of one), where we apply the ICDF to find the value of *x* that corresponds to the cumulative probability of 97.50% (CDF). That is, a one-tail CDF of 97.50% is equivalent to a two-tail 95% confidence interval (there is a 2.50% probability in the right tail and 2.50% in the left tail, leaving 95% in the center or confidence interval area, which is equivalent to a 97.50% area for one tail). The result is the familiar Z-Score of 1.96. Therefore, using this Distributional Analysis tool, the standardized scores for other distributions, the exact and cumulative probabilities of other distributions can all be obtained quickly and easily.

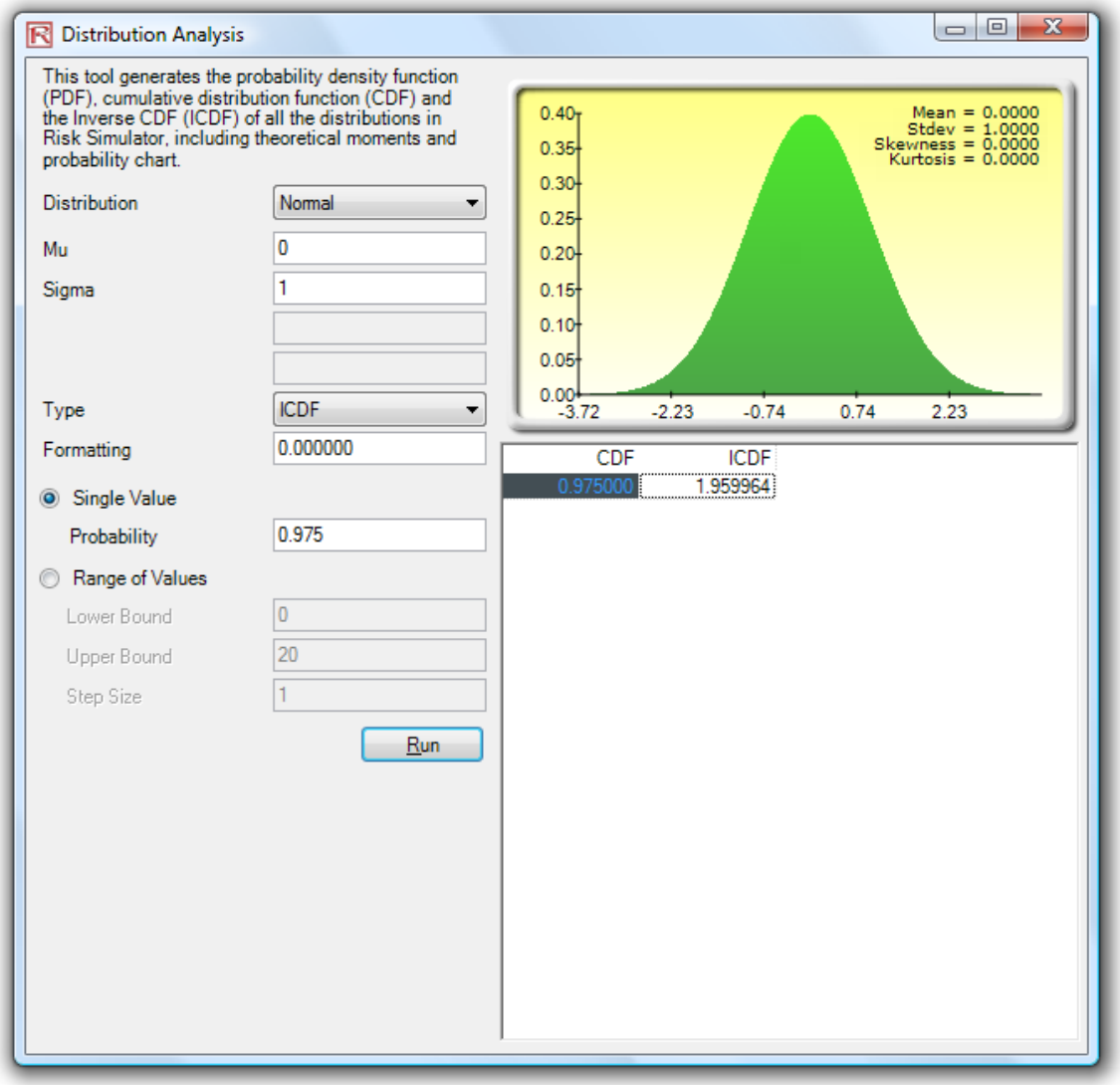


Figure 53 – Distributional Analysis Tool (Normal Distribution's ICDF and Z-Score)