

Simulation for Applied Risk Management

with an Introduction to *SIMETAR*®

James W. Richardson

Regents Professor & TAES Senior Faculty Fellow

Department of Agricultural Economics
Texas A&M University



SIMETAR®
©2008

Simetar Quickstart

Simulating Random Variables

- Basic Univariate Probability Distributions
 - =UNIFORM (min, max); USD=UNIFORM()
 - =NORM (mean, std dev); SND=NORM()
 - =EMPIRICAL (sorted data)
 - =BETAINV (USD, alpha, beta, min, max)
 - =GAMMINV (USD, alpha, beta)
 - =BERNOULLI (p)
- Multivariate Probability Distributions
 - =MVNORM (mean vector, covariance matrix)
 - =MVEMPIRICAL (matrix of variables)
 - =CSND (correlation matrix)
 - =CUSD (correlation matrix)

Toolbar Icons


























-  -- Simulation options for simulating an Excel workbook
-  -- Set all random variables to expected value
-  -- Calculate summary statistics
-  -- Multiple regression (OLS, GLS, Logit, Probit, Ridge, 2SLS, WLS)
-  -- Simple regression
-  -- Calculate correlation and covariance matrices
-  -- Data manipulation and matrix operation functions
-  -- Stochastic dominance with respect to a function
-  -- Stochastic efficiency with respect to a function
-  -- Develop a stoplight chart for comparing risky alternatives
-  -- Statistical tests for validating simulated random variables
-  -- AR and VAR time series model estimates
-  -- Forecasting with exponential smoothing and seasonal indices
-  -- Univariate parameter estimator for 16 probability distributions
-  -- Develop line chart with and without labels on data points
-  -- Fan graph of alternative stochastic scenarios
-  -- Histogram of alternative stochastic scenarios
-  -- Cumulative distribution function chart
-  -- Probability density function chart
-  -- Probability plot charts (NP, QQ and PP plots)
-  -- Box plot chart
-  -- Scatter matrix plot
-  -- Estimate parameters for an empirical probability distribution
-  -- Personal settings for Simetar
-  -- Help for 250+ user defined functions in Simetar

Table of Contents

Forward

Chapter 1 – Introduction to Risk

Review of Literature.....	2
References	3

Chapter 2 – Terminology for Simulation Modeling and Steps for Model Development

Terminology	1
Advantages and Disadvantages of Simulation.....	3
Complexity of a Simulation Model or What is in a Model?.....	5
Steps for Model Development.....	6
References	13

Chapter 3 – Model Validation

Verification.....	1
Validation	1
Steps for Model Verification and Validation.....	2
The 4 Ps for Simulation Model Validation.....	5
Validation Never Ends	5
Validation Statistics.....	6
Other Statistical Tests for Validation	11
Graphical Tools for Validation	12
Normality Test.....	15
References	16

Chapter 4 – Stochastic Simulation

Random Variables in a Simulation Model.....	2
Iterations.....	4
Number of Iterations.....	6
Monte Carlo vs. Latin Hypercube Sampling	7
Pseudo-Random Number Generator.....	7
Simulating Uncertainty.....	10
References	11

Chapter 5 – Distributions Frequently Used for Simulation

Continuous Distributions.....	1
Discrete Distributions.....	6
Conditional Probability Distributions.....	7
When are these Distributions Used?.....	9
Inverse Transform Method for Simulating Random Variables	11
Summary	18
References	19

Chapter 6 – Parameter Estimation for Univariate Probability Distributions

Random Variables Without Trend.....	2
Random Variables With a Trend	7
Random Variables as a Function of Other Variables	12
Random Variables a Function of an Autoregressive Structure.....	16
Estimating Parameters for Other Distributions.....	16
Summary	19
References	20

Chapter 7 – Parameter Estimation for Multivariate Probability Distributions

Ignoring Correlation	1
Multivariate Normal (MVN) Distribution	2
Multivariate Empirical (MVE) Distribution	8
Mixed Multivariate Probability Distributions	13
Simulating Large Multivariate Distributions	16
Appendix: Simulation of a MVN Distribution	22
References	24

Chapter 8 – Simulating Inter- and Intra- Temporal Multivariate Distributions

Review of Literature	1
Simulating Multivariate Non-Normally Distributed Random Variables	2
Parameter Estimation for a MVE Probability Distribution	3
Simulation of a MVE Probability Distribution	6
Numerical Application of the MVE Distribution	9
Summary	11
References	18

Chapter 9 – Coefficient of Variation Stationarity and Simulating Heteroskedastic Error Terms

CV Stationarity for the Normal Distribution	1
CV Stationarity and the Empirical Distribution	3
Controlling Heteroskedasticity for Simulation	5
References	8

Chapter 10 – Simulating Alternative Scenarios and Selecting the “Best” Scenario

Simulating Multiple Scenarios	1
Ranking Risky Alternatives	3
Mean Only	4
Standard Deviation	4
Mean Variance (MV)	5
Minimum and Maximum (or Worst and Best Case)	5
Relative Risk (CV)	6
Probabilities of Target Values	7
Complete Distribution - CDF Chart	8
Expected Utility (EU)	9
Break Even Risk Aversion Coefficients (BRACs)	15
Stochastic Efficiency Respect to a Function (SERF)	16
Summary of Risk Ranking Procedures	20
References	21

Chapter 11 – Bootstrap Simulation

Bootstrap Simulation	1
Bootstrap Simulation of Multivariate Distributions	3
Bootstrap Simulation and Regression Analysis	4
References	8

Chapter 12 – Optimization of a Simulation Model

Principles of Control Theory	1
Numerical Solution of Optimal Control Problems	3
Numerical Optimization in Excel	4
Optimizing a Deterministic Econometric Model	6
Optimizing a Stochastic Econometric Model	8
References	10

Chapter 13 – Special Problems in Modeling Risk

Simulating Risky Cash Flows.....	1
Income Taxes	3
Debt Amortization	4
Net Present Value (NPV) and Internal Rate of Return (IRR).....	4
Estimating Insurance Premiums	6
Machinery Replacement.....	7
Farm Level Simulation Model.....	8
Simulating an Econometric Model.....	10
References	14

Chapter 14 – Simulation Applications for Business Management

Financial Risk Management.....	1
Portfolio Analysis.....	3
Project Management.....	6
Bid Analysis	9
Project Feasibility	11
Inventory Management.....	15
References	18

Chapter 15 – Probabilistic Forecasting

Probabilistic Forecasting	1
Quantifying Forecast Error.....	2
Trend Regression Forecasts.....	3
Multiple Regression	6
Seasonal Forecasts Using Dummy Variables	8
Seasonal Forecasts Using Harmonic Regression.....	10
Seasonal Indices	12
Cycles	15
Moving Average Forecasts.....	16
Decomposition Forecasting	18
Exponential Smoothing	24
Time Series Analysis.....	30
References	38

Chapter 16 – Simetar© Simulation & Econometrics To Analyze Risk

1.0 What is Simetar?	1
1.1 Installing Simetar.....	1
2.0 Simulating Random Variables	3
2.1 Probability Distributions in Simetar.....	3
2.2 Simulation Engine in Simetar	5
2.3 Specifying Options in the Simulation Engine.....	6
2.4 User Defined Settings	7
3.0 Probability Distributions Simulated in Simetar	8
3.1 Uniform Probability Distribution.....	8
3.2 Normal Related Probability Distributions.....	8
3.3 Continuous Probability Distributions.....	11
3.4 Finite-Range Continuous Probability Distributions.....	14
3.5 Analogs to Finite Range Probability Distributions	15
3.6 Discrete Probability Distributions.....	16
3.7 Sample Based Probability Distributions	18
3.8 Time Series Probability Distributions.....	21
3.9 Multivariate Distributions.....	21
3.10 Iteration Counter	28

4.0 Parameter Estimation for Parametric Probability Distributions	29
4.1 Parametric Probability Distributions.....	29
4.2 Empirical Probability Distributions.....	30
4.3 Multivariate Probability Distributions.....	30
4.4 GRKS Probability Distributions.....	32
5.0 Statistical Tests for Model Validation	32
5.1 Univariate Distribution Tests.....	32
5.2 Multivariate Distribution Tests.....	33
5.3 Test Correlation.....	34
5.4 Test Mean and Standard Deviation.....	35
5.5 Univariate Tests for Normality.....	35
5.6 Multivariate Tests for Normality.....	36
5.7 Compare Means (ANOVA).....	36
5.8 Compare Two Cumulative Distribution Functions (CDFs).....	37
6.0 Graphical Tools for Analyzing Simulation Results	38
6.1 Line Graph.....	38
6.2 CDF Graph.....	39
6.3 PDF Graph.....	40
6.4 Histograms.....	41
6.5 Fan Graph.....	41
6.6 StopLight Graph.....	42
6.7 Probability Plots.....	43
6.8 Box Plots.....	43
6.9 Scatter Matrix Graph.....	44
7.0 Scenario Analysis	45
8.0 Sensitivity Analysis	46
9.0 Sensitivity Elasticity Analysis	47
10.0 Simulation and Optimization	48
11.0 Numerical Methods for Ranking Risky Alternatives	49
11.1 Stochastic Dominance (SD).....	49
11.2 Stochastic Efficiency with Respect to a Function (SERF).....	50
11.3 Risk Premiums.....	51
11.4 Target Probabilities for Ranking Risky Alternatives.....	51
11.5 Target Quantiles for Ranking Risky Alternatives.....	52
12.0 Tools for Data Analysis and Manipulation	52
12.1 Matrix Operations.....	52
12.2 Data Manipulation.....	55
12.3 Box-Cox Transformation.....	56
12.4 Workbook Documentation.....	56
13.0 Regression Analysis	57
13.1 Simple Regression.....	58
13.2 Multiple Regression.....	58
13.3 Bivariate Response Regression.....	61
14.0 Cyclical Analysis and Exponential Forecasting	62
14.1 Seasonal Index.....	62
14.2 Seasonal Decomposition Forecasting.....	63
14.3 Moving Average Forecast.....	64
14.4 Exponential Smoothing Forecast.....	64
14.5 Measuring Forecast Errors.....	66

15.0 Time Series Analysis and Forecasting	67
15.1 Tests for Stationarity.....	67
15.2 Number of Lags.....	68
15.3 Sample Autocorrelation Coefficients.....	69
15.4 Maximum Likelihood Ratio Test.....	70
15.5 Estimating and Forecasting Autoregressive (AR) Models.....	70
15.6 Estimating and Forecasting Vector Autoregressive (VAR) Models.....	72
16.0 Other Statistical and Data Analysis Functions	74
16.1 Summary Statistics.....	74
16.2 Jackknife Estimator.....	74
17.0 Function Evaluation	75
17.1 Optimize a Function.....	75
17.2 Value of a Function.....	76
17.3 Integral of a Function.....	76
18.0 Getting Help with Simetar	77
19.0 Solutions to Problems in Simetar Application	78
20.0 List of All Simetar Functions	82
21.0 Cross Reference of Functions and Demonstration Programs	86

Chapter 1 Introduction to Risk

“Reality is only one realization.” -- Chip Conley

Decision making in business implies that management has a choice among alternative actions. The alternative actions could be different combinations of crops to produce, alternative production systems for crops or livestock, or different marketing or financial strategies for an agribusiness. If the decisions are to be made in a risk free setting, the manager can easily determine which strategy is best ... the one with the greatest economic return. When decisions are made in a risky environment the manager cannot use such a simple rule because the economic return for each alternative is a distribution of returns rather than a single value. One approach to decision making under risk is to simulate the alternative strategies to estimate the distribution for each alternatives' return and then base the decision on these simulated distributions.

The purpose of simulation in risk analysis is to estimate distributions of economic returns for alternative strategies so the decision maker can make better management decisions.

Without risk there is very little to be gained from simulation beyond what you can do with a calculator. Simulation models without risk (deterministic models) can be used to learn how an economic system works and how it responds to managerial changes. Deterministic models can answer the first round of “What if?” questions about the system. However, without risk in the model the outcomes for the alternative strategies will not be robust enough for actual decision making in a risky economic environment.

At this point it is worthwhile to define risk. In an economic system there is risk associated with yields (production), input prices, output prices, interest rates, annual rates of change in input prices, market share, etc. For farmers, ranchers, and agribusiness managers there is a common thread among these variables, namely: the manager does not have control over these variables. **Risk is the part of a business decision, the manager cannot control.**

The Texas Agricultural Extension Service surveyed producers in three regions of Texas and three regions of Kansas as to their perceptions of sources of risk. Farmers in both states ranked price risk as the most important, followed by yield risk and then changes in input costs. The producers were also asked about their risk management techniques.

Review of Literature

The history of risk analysis is outlined by Bernstein and is highly recommended reading for those wanting a deeper explanation of risk. Hardaker, et al., Fleisher and Robison and Barry provide descriptions of risk in agriculture and ways that firm managers can manage risk.

A comprehensive, although not complete, review of literature in the area of simulation analysis by agricultural economists from the 40s through the 70s is provided by Johnson and Rausser. Their work is highly recommended for the serious student of simulation. A second review of literature on simulation worthy of note is by Anderson in 1974.

The application of simulation techniques to project assessment and investment analysis is described best by Pouliquen and by Reutlinger. They were innovators in the use of risk analysis to evaluate risky investments using probability distributions of benefit cost ratios. These early works have stood the test of time and remain relevant to project assessment under risk today.

An early introduction to risk analysis for business decisions is provided by Jones and by House and more recently by Winston. Richardson and Mapp's firm level modeling effort in 1976 lead to Richardson and Nixon's development of the Farm Level and Income Simulator (FLIPSIM) and more than 250 firm level simulation analyses. (See the Economic Models bibliography at www.afpc.tamu.edu for an extensive list of publications that have used FLIPSIM to analyze a range of topics in agricultural economics.) This line of firm level simulation modeling continues with the development of the Farm Assistance model by Klose and Gray's agribusiness model named FRAN.

References

- Anderson, J.R. "Simulation: Methodology and Applications in Agricultural Economics." Review of Marketing and Agricultural Economics, 43(1974): 3-55.
- Bernstein, P.L. Against the Gods: The Remarkable Story of Risk. New York: John Wiley & Sons, Inc. 1996.
- Fleisher, B. Agricultural Risk Management. Boulder, CO: Lynne Rienner Publishers, Inc., 1990.
- Gray, A.W. "Agribusiness Strategic Planning Under Risk." Texas A&M University, Department of Agricultural Economics, Ph.D. Dissertation, 1998.
- Hardaker, J.B., Huirne, R.B.M., and J.R. Anderson. Coping with Risk in Agriculture. Wallingford, UK: CAB International, 1997.
- House, W.C., editor, Business Simulation for Decision Making. New York: PBI, 1977.
- Johnson, S.R. and G.C. Rausser. "Systems Analysis and Simulation: A Survey of Applications in Agricultural and Resource Economics." Ed. G.C. Judge, R.H. Day, S.R. Johnson, G.C. Rausser, and L.R. Martin. A Survey of Agricultural Economics Literature, (Volume 2), Quantitative Methods in Agricultural Economics, 1940s to 1970s. Minneapolis: U. Minnesota Press, 1977.
- Jones, G.T. Simulation and Business Decisions. Middlesex, England: Penguin Books Ltd., 1972.
- Klose, S.L. "A Decision Support for Agricultural Producers." Texas A&M University, Department of Agricultural Economics, Ph.D. Dissertation, 2001.
- Law, A.M. and W.D. Kelton. Simulation Modeling and Analysis, Third Edition. New York: McGraw-Hill, Inc., 2000.
- Pouliquen, L.Y. Risk Analysis in Project Analysis. Baltimore: The Johns Hopkins Press, 1970.
- Reutlinger, S. Techniques for Project Appraisal Under Uncertainty. Baltimore: The Johns Hopkins Press, 1970.
- Richardson, J.W. and H.P. Mapp, Jr. "Use of Probabilistic Cash Flows in Analyzing Investment Under Conditions of Risk and Uncertainty." Southern Journal Agric. Econ. Dec. 1976.
- Richardson, J.W. and C.J. Nixon. "A Description of FLIPSIM V: A General Firm Level Policy Simulation Model." Bulletin 1528, Texas Agricultural Experiment Station, July 1986.

Robison, L.J. and P.J. Barry. The Competitive Firm's Response to Risk. New York: MacMillan Publishing Co., 1987.

Winston, W.L. Simulation Modeling Using @Risk©. New York: Duxbury Press 1996.

Chapter 2

Terminology for Simulation Modeling and Steps for Model Development

Webster's New Collegiate Dictionary defines simulation as “the act or process of simulating,” or “the imitative representation of the functioning of one system or process by means of the functioning of another,” and “examination of a problem often not subject to direct experimentation by means of a simulating device.” All three of these definitions are useful to describe the material presented in this book. Economists and business analysts construct mathematical representations of actual systems that can not be experimented on directly for the purpose of simulating “What if ...?” questions. It is the goal of modelers to construct models that imitate how the real systems would respond to exogenous changes in management and policy.

One could compare simulation models developed by economists and business analysts to experiments run by bench scientists. Bench scientists conduct repeated experiments and record results on plants, animals, or cells under controlled situations for specific treatments. This process is a simulation, just the same as an economist who experiments on a business model to forecast sales and profits for the next year. For social scientists there is no laboratory to experiment on human subjects or “feedlot” where they can do feed or price trials on unsuspecting consumers.

For this text the working definition of a simulation model is: a mathematical representation of a business or economic system that reflects sufficient detail of the system to address the questions at hand. An economic model of a system can thus be very simple or very detailed depending upon the question being addressed. If the question is economic viability in 10 years one does not need to include inventory management rules and algorithms for purchasing Xerox paper, where as if the question is monthly cash flow management this may be a relevant part of the model.

Simulation is defined as the process of solving a mathematical simulation model representing an economic system for a set of exogenous variables. Alternative management strategies and policy scenarios constitute the exogenous variables and are the numerical representation of a “What if ...?” question. A simulation model is “solved” a large number of times to statistically represent all possible combinations of the random variables in the system. The results of a simulation process is a large number of simulated values for key output variables (KOVs) of interest to the decision makers. The simulated values for a KOV represent an empirical estimate of the probability distribution for the variable and quantifies the risk associated with the variable. This type of answer is analogous to performing a large number of lab trials on mice using the same dosage of product X to determine the mean and variance of a lethal dose.

Terminology

– Types of Simulation Models

Simulation models used for economic and business analysis are digital as opposed to analog models used by most bench scientists. Economists express the functions of a system as a series

of inter-related mathematical equations to solve for the variables of interest to decision makers. Analog models use real time examples such as test plots for crops or animal feed trials to show the response to “What if...?” questions. Digital models that produce results in seconds are cheap to build and run while analog models can take weeks or months to run and costs can be quite high.

The old standby for describing the difference between analog and digital models is a clock. The clock with two hands and numerals on the face is an analog model that simulates the passage of time by moving the hands around the center point. A digital clock shows the time with two digits representing the hour and two digits representing the minutes, and time is simulated by changing the numbers.

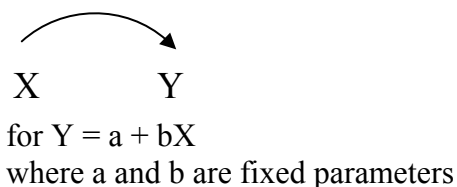
– Linear Programming vs. Simulation

Mathematical programming models are simulation models in one sense. They can represent the mathematical relationships necessary to describe an economic system or business. However, they seldom incorporate risk and they solve for the optimal answer and give the normative answer of what ought to be. Simulation models, on the other hand, incorporate risk and answer the positive question of what is the likely outcome.

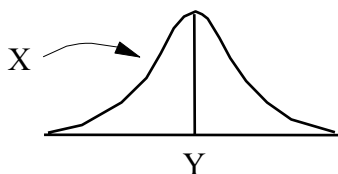
Both quantitative analysis techniques can be used to answer What if ...? questions. Experts in linear programming can configure their models to incorporate risk and expert simulators can optimize their simulation models. In general, programming models are used to answer the normative questions of what is the most profitable enterprise mix or the least cost method of meeting a goal. Simulation models are generally used to answer questions of what is the profit risk for specific enterprise mixes or what is the range of costs for alternative means of meeting a particular goal.

– Deterministic vs. Stochastic Simulation Models

Simulation models can be solved both deterministically and stochastically. Deterministic models are simulation models without risk and are solved using simple calculator arithmetic. Each X variable is mapped to a single outcome Y or



Stochastic simulation models are solved a large number of times using one value for X to generate a sample of outcomes for the dependent variable Y, recognizing that X has risk



$$Y = a + bX + \tilde{\epsilon}$$

where $\tilde{\epsilon}$ represents a probability distribution of the risk about the deterministic component of Y defined by $a + bX$

Because there is risk in the forecast for Y, it must be forecasted using a probability distribution rather than using a point estimate. The simulated distribution for Y informs the decision maker of the riskiness of the forecast for the KOV, the skewness of the outcome, and the chances of a favorable outcome, all answers not available from a deterministic or linear programming forecast.

– **Discrete vs. Continuous Event Simulation Models**

Another label used to describe simulation models is discrete and continuous. Discrete models refer to simulation models that have distinct increments for time, such as years, months, weeks or days. Continuous models have no time steps or there are so many periods in the time horizon that periods are not differentiated. Most economic models are discrete in that time is important and the models must account for changes in value, stock levels, cash flows, and debt repayment. Production models are particularly time sensitive because of the pre-plant activities, the growing season, and the harvest periods. Optimization models are generally continuous models in that the optimal activity combination is not sensitive to time periods. More complex optimization models, called poly-period programming models, optimize an object function over multiple periods.

– **Kinds of Simulation Models Economists Build**

Economists generally build and use digital, discrete-event, non-linear, stochastic models of business systems. Examples of economic simulation models are:

- Single enterprise model
- Whole farm model like FLIPSIM (Richardson and Nixon)
- Agribusiness model like FRAN (Gray or Gill)
- Commodity model such as cotton, or wheat, or corn, or dairy, etc.
- Supply/demand sector model with multiple commodities for policy analysis such as cotton – wheat – feedgrains – soybeans – rice and/or cattle – hogs – dairy – poultry (FAPRI Model) (Brown, Adams, Ray and Richardson)
- National economic model (Hughes)
- International model encompassing many countries and commodities (CARD at Iowa State)

Advantages and Disadvantages of Simulation

– **Advantages**

Simulate facilitates experimentation on real world systems that can not be tested in real life because they are too complex, involve human subjects, have too long of planning horizons, and/or the cost of experimentation is prohibitive. For example, it is not feasible to experiment on the U.S. agricultural economy with alternative loan rates, because of all the reasons listed above. So

simulation models of U.S. agriculture are developed and What if ...? questions are posed to project the probable impacts of alternative loan rates and policy settings. Once a model is developed and validated it can be used for a range of different analyses thus spreading the cost of model development over more periods and clients.

In the case of a business model, one can develop a model as part of a project feasibility study; test the feasibility of the project under a range of situations; and use the model to test alternative management options if the project is undertaken. This type of simulation model allows the analyst to experiment on a proposed system that does not yet exist and to test management plans without endangering the project once it is in place.

The instantaneous nature of simulation allows analysts to experiment on systems and projects that have very long planning horizons. Managers are hesitant to take a chance on a project which takes 5 to 10 years to produce a risky payoff. However, simulation can produce an estimate of the risky payoff for an investment in minutes so managers can make better decisions.

Once a simulation model has been developed and accepted by management as a realistic representation of the business it can be used for training. In this mode a simulation model can be used to train newly hired managers as well as mid-level managers. Using the What if ...? paradigm managers can learn how the business responds to a wide range of policy, economic, environmental, and management situations. Learning on a virtual reality version of the business does not endanger the real business as trainees learn how to and how not to manage the firm.

A final advantage of simulation is that much is learned about a system during the model development and validation phase. This human capital developed through building a simulation model is an asset to the business because now a team of developers have a comprehensive overview of how the business operates, rather than just one or two people at the top knowing how it all works.

– **Disadvantages**

A major disadvantage to simulation models is the cost of developing and maintaining a model. Without a doubt the business must be committed to funding the human capital required to develop a comprehensive simulation model. However, once a model is developed it can be used for years with the proper updating and modification (e.g., FLIPSIM was written in 1979, released in 1980 and has been used for 100's of analyses).

The biggest disadvantage is that people tend to believe projections developed by a simulation model just "because the model said so!" This should never be the case. Models are constructed by humans and provide projections based on input from more humans, so how do they produce perfect forecasts? They don't! Most models produce so many output values that people get lost in the output, so they bestow undeserved confidence on the model's output. The fastest way to kill the credibility of a model is to start believing the output without critically checking all projections before releasing the results. As simulation models get larger and more detailed they become harder to critically check the output. Given today's modern computers it is easier to build and use large stochastic models, we have to include risk in business decisions, people expect too much out of simulation models, models generate such large volumes of output it

is difficult to check the results, and in today's economy one must produce a result instantly, thus problems will arise.

The final disadvantage of simulation model's is that they only provide an estimate of the true probability distribution for key output variables. As indicated in Figure 2.1 the simulated pdf may be a reasonable estimate of the true distribution, but it will never be perfect.

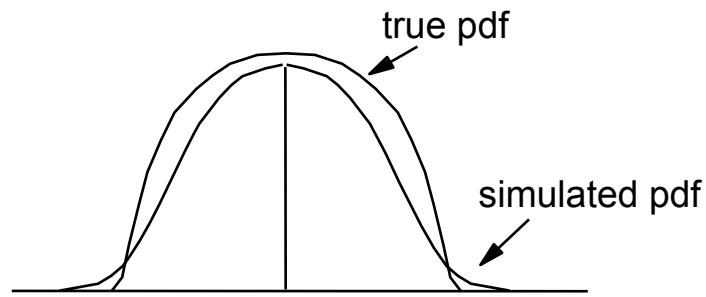


Figure 2.1. Comparison of the True PDF and the Simulated PDF.

– **Final Word of Warning:**

“Simulation is to teach us about a system and to facilitate better decisions, not to predict point estimates and make decisions.”

Complexity of a Simulation Model or What is in a Model?

A simulation model need only be as sophisticated as the client requires to provide a good answer to a relevant problem in a timely manner. If your client is a group of risk averse investors being asked to put up \$50 million you will need a very detailed model of the investment with full accounting for risk in all components of the system. If on the other hand, you are modeling the toss of the coin at the start of a football game, the model need not be very sophisticated.

In general all models have the following parts

1. An initial environment stating values for all exogenous variables and their assumptions over the planning horizon.
2. Control variables the user can manage for alternative scenarios.
3. A specific sequence of equations that determines the computational flow in the system.
4. A probability distribution simulation component for simulating the stochastic variables in the system.
5. A formatted report containing the intermediate and final results of the equations in a logical format using tables and charts useful to management.

When developing simulation models in Excel the use of an organizational template is recommended. An example of a simulation model template is included in Figure 2.2. Sheet 1 of a workbook model should be organized so the input data, the control variables and a summary of the key output variables appears in the first screen (upper left quadrant in Figure 2.2). Below this first screen the model should contain a logical flow of equations to calculate the intermediate and final or key output variables for the system (bottom left quadrant of Sheet 1 in Figure 2.2). In Sheet 2 of the workbook, the developer should start with all of the historical data for the stochastic variables in the model. Parameter estimation for the distributions assumed for the stochastic variables should follow. Actual simulation of the stochastic variables should take place in Sheet 2 (bottom third of Sheet 2 in Figure 2.2). Statistical summaries of the key output variables and the stochastic variables are placed in separate sheets by the simulation procedure (Simetar or @Risk©), as well as the graphs for key output variables, as depicted in Figure 2.2.

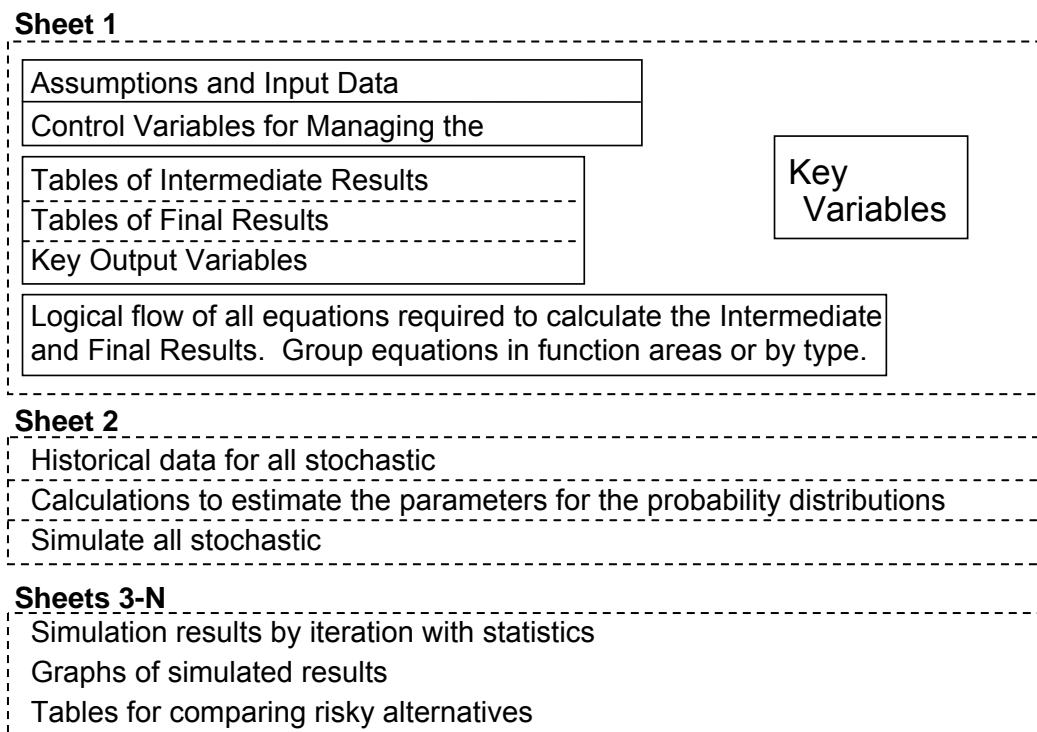


Figure 2.2. Recommended Organization of Simulation Models in Excel.

Steps for Model Development

There are numerous ways to build a simulation model. The least successful is to use the trial and error approach to develop each equation. The best approach is to investigate the system thoroughly, sketch a diagram of the system and then build it from the top down. Using the top down approach means that you first determine the output variables (KOVs in Figure 2.3) and then work backwards to determine the equations and parts of the model needed to properly calculate the output variables (levels 2, 3, 4, 5 in Figure 2.3). The steps to this approach are described in detail in this section.

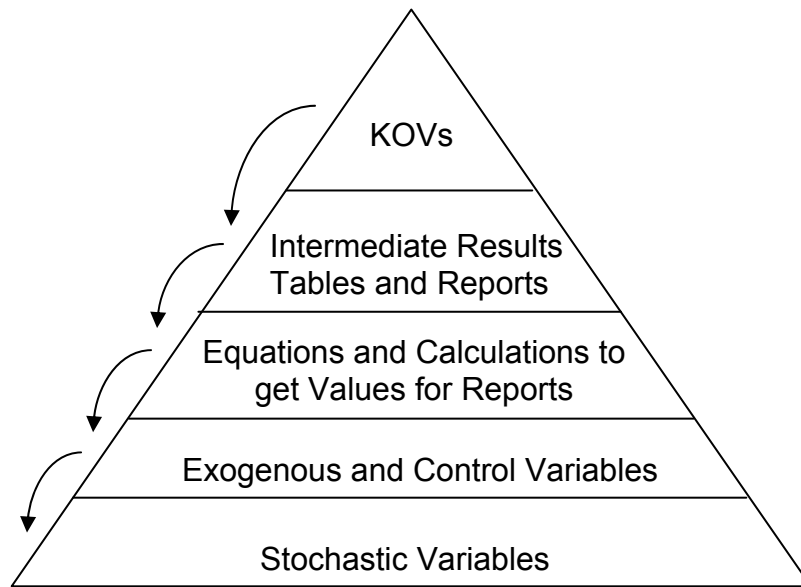


Figure 2.3. Simulation Pyramid.

– **Determine the Model’s Use and Sketch a Diagram of the System**

Determine the actual use the model will be put to and review the literature in the area to see what types of models have been built to do similar things. Develop a simple diagram of the concept behind your model. The diagram should be refined after each step in the process. Model purposes are many, here are a few possibilities:

- Investment in a risky venture or an analysis of alternative portfolios
- Managing a risky business or enterprise (a system)
- Policy analysis
- Pricing a product
- Assessing technology
- What if?

– **Define the Key Output Variables**

The key output variables (KOVs) are any variables the decision maker (the client) thinks are important to the decision at hand. The KOVs actually determine the type of model to develop. If the primary KOV is an internal rate of return (IROR), the model has to include all of the variables necessary to calculate IROR. On the other hand, if the KOV is the level of ending stocks of wheat for the U.S. then the model must include supply and demand equations for the U.S. wheat sector.

The model flow chart can start taking on some detail once the KOV’s are determined (Figure 2.4). Possible KOV’s for three different types of simulation models are listed here to illustrate the range of KOV’s that can be included in a model. Specific KOVs could be any variable the decision maker wants for basing his/her decisions on, such as:

- For a firm level model KOVs include financial ratios, such as:
 - debt asset ratio
 - internal rate of return (IRR)
 - net present value (NPV)
 - percentage change in real net worth ($\% \Delta \text{RNW}$)
 - probability of cash flow deficits
 - probability of refinancing
 - probability of insolvency
 - probability of economic success

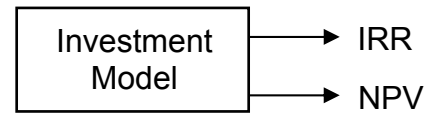


Figure 2.4. Model Diagram for Step 2.

- For a sector level model (econometric crop model) KOVs may include:
 - supply variables: yield, acres, production, supply
 - utilization variables: domestic use, exports, industrial use
 - price
 - government costs
 - stock levels
 - probability (low stocks)
 - probability government costs exceed a target level
- For an investment decision model KOVs may include:
 - profit or internal rate of return (IROR)
 - probability of a positive net present value (NPV)
 - probability of IRR exceeding the investor's minimum rate of return
 - probability of cash flow deficits
 - comparison of IRR or NPV between investment alternatives
 - benefit cost ratio

– Determine the Intermediate Outputs

Intermediate output variables are those variables in the model necessary to calculate the KOVs and to fill the output tables with information useful to the decision maker. These are the variables in the second level of the simulation pyramid in Figure 2.3.

- For a firm level model intermediate output variables are the values in the:
 - income statement
 - cash flow
 - balance sheet
 - enterprise summary of production, receipts, and costs

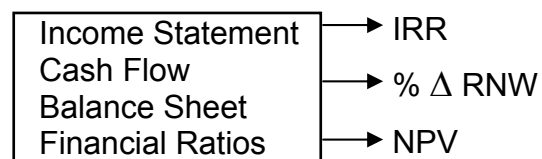


Figure 2.5. Model Design for Step 3.

- For a sector level model
 - acres planted by region, yield by region, total supply
 - equilibrium price
 - utilization (feed, seed, industrial, exports, stocks, CCC purchases)
 - ending stocks

- For an investment model
 - Profit each period
 - capital requirements each year
 - debt and assets each year

– **Write the Equations**

Write out all of the equations that will be required to calculate each of the intermediate and final variables in the results tables identified in the previous steps. The order of calculation for each of the variables should be established at this point based on the logical order that variables are used in the results tables. Update the model schematic diagram to show the logical flow of calculations and equations (Figure 2.6).

- For a firm level model, largely accounting identities define the relationships. Acres, yields, production, prices, receipts, costs, and government payments for each crop are used to calculate values for the income statement on a crop farm. Number of cows, calves born, calves sold, heifers retained, sale weights, prices, receipts, and costs for cattle are used for a cattle ranch model. For example, for a crop model the equations would include:

$$\begin{aligned} \text{Prod}_t &= \text{Acres}_t * \text{Yield}_t \\ \text{Receipts}_t &= \text{Prod}_t * \text{Price}_t \\ \text{Cost}_t &= \text{FixCost}_t + \text{Seed Cost}_t * \text{Acres}_t + \dots + \text{Fuel Cost}_t * \text{Acres}_t \\ \text{Net Returns}_t &= \text{Receipts}_t - \text{Costs}_t \end{aligned}$$

- For a sector model the equations can be a mix of identity and econometric variables and equations. Equations to calculate the endogenous variables identified in the previous should be hypothesized at this point. In writing out the hypothesized equations list all of the exogenous variables for each variable/equation. By writing out the equations we are developing a list of exogenous variables required by the model.

$$\begin{aligned} \text{Acres}_t &= \hat{a} + \hat{b} \text{ Price}_{t-1} + \hat{c} \text{ Policy Var}_t \\ \text{Yield}_t &= \hat{a} + \hat{b} \text{ Price}_{t-1} + \hat{c} \text{ Yield}_{t-1} \\ \text{Prod}_t &= \text{Acres}_t * \text{Yield}_t \\ \text{Supply}_t &= \text{Prod}_t + \text{Carryin}_t + \text{Imports}_t \\ \text{Price}_t &= \hat{a} - \hat{b} \text{ Supply}_t \\ \text{Feed } D_t &= \hat{a} - \hat{b} \text{ Price}_t + \hat{c} \text{ Price other crop}_t \\ &\vdots \\ &\vdots \\ &\vdots \\ \text{Food } D_t &= \hat{a} - \hat{b} \text{ Price}_t + \hat{c} \text{ Price other crop}_t + \hat{d} \text{ Income}_t \\ \text{Export } D_t &= \hat{a} - \hat{b} \text{ Price}_t + \hat{c} \text{ Price other crop}_t \\ \text{Total Use}_t &= \text{Feed } D_t + \text{Food } D_t + \text{Export } D_t \\ \text{Carryover}_t &= \text{Supply}_t - \text{Total Use}_t \\ \text{Government Cost}_t &= \text{Rate} * \text{Yield}_t * \text{Base Acres}_t \end{aligned}$$

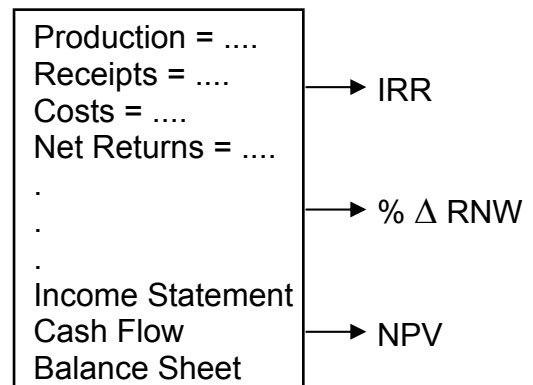


Figure 2.6. Model Design for Step 4.

– Define Input and Calculated Variables

Once the equations are specified it is easy then to define the variables you must specify as input (exogenous) and variables calculated in the model (endogenous). In Step 5, begin the process of estimating parameters for the endogenous equations.

- Exogenous variables are determined external to the model and constitute the initial environment or values for the model, i.e., you must input these values. Sources for projected values of the exogenous variables are government and university modelers, such as: USDA, WEFA, FAPRI, CARD, and Project LINK at the UN. Once the list of exogenous variables is completed, update the model diagram as in Figure 2.7. An example of exogenous variables for economic variables is:

- annual interest rates and rates of inflation for costs and assets
- initial costs of production
- initial endowment of assets and debts
- production functions and input/output coefficients
- government policy values for each policy tool
- carryin stocks

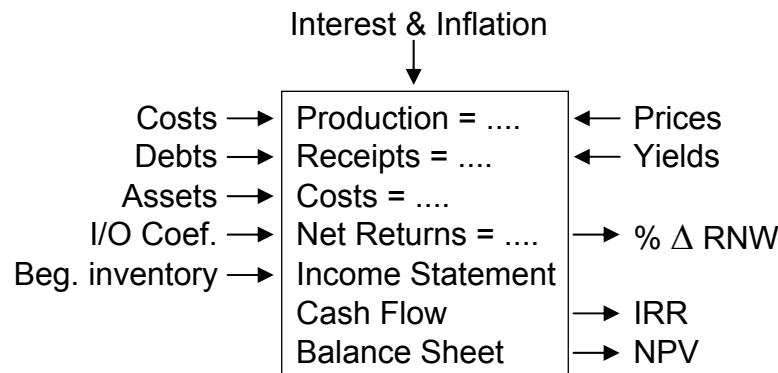


Figure 2.7. Model Design for Step 5.

Endogenous variables are determined internal to the model, i.e., they are calculated. Regression and time series techniques should be used where appropriate at this point to estimate the parameters for these equations. The rule is to develop the best model or equation to project each endogenous variable. Make sure that the econometrically estimated equations are compatible, in other words they simulate appropriately as a system of equations. A list of endogenous variables for economic models would include:

- production
- cash receipts
- costs
- price
- quantities demanded
- government costs
- net returns, cash flows, and net worth
- financial ratios

– **Identify Stochastic Variables**

Once the equations are specified and the parameters are estimated, define the variables which will be stochastic. The stochastic variables are: (1) variables we are still uncertain about, even after the best forecast we can devise, and (2) those variables the decision makers cannot control or predict. Parameters necessary for simulating the distribution for each stochastic variable must be calculated at this point. The subject of parameter estimation for stochastic variables is covered in detail in Chapters 6 and 7.

- Stochastic variables in a firm level model are yields (or production level) and prices. Yield is usually trend projected and there is an unexplained error term about the trend forecast that we sample from during simulation. The unexplained variability in Figure 2.8 is indicated by the residuals about the trend line or $Y = a + bT + e$. We use the \hat{e} as a measure of the risk about the projected \hat{Y} values. Similar specifications can be used to estimate parameters for the distributions of stochastic prices, and other stochastic variables.

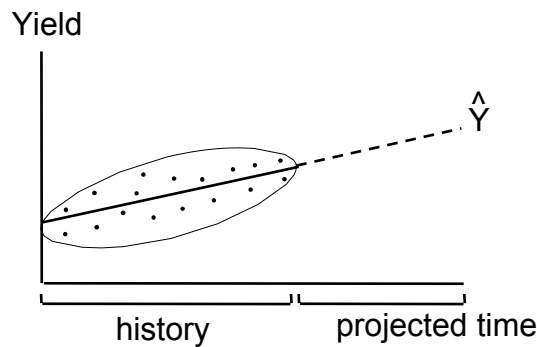


Figure 2.8. Uncertainty about Endogenous Forecasts

- For a commodity or sector level model, the unexplained variability (\hat{e}) for each econometric equation defines the probability distributions about the projected values, as indicated in Figure 2.8. Additionally, these models use the stochastic prices for previous and current periods to more realistically simulate risk in the system.

$$\text{Acres}_t = \hat{a} + \hat{b}\text{Price}_{t-1} + \hat{e}_1$$

$$\text{Yield}_t = \hat{a} + \hat{b}\text{Price}_{t-1} + \hat{e}_2$$

$$\text{Price}_t = \hat{a} + \hat{b} \text{Supply}_t + \hat{e}_3$$

$$\text{Feed demand}_t = \hat{a} + \hat{b}\text{Price}_t + \hat{e}_4$$

⋮
⋮

$$\text{Export demand}_t = \hat{a} + \hat{b}\text{Price}_t + \hat{e}_n$$

These error terms (\hat{e} 's) are the stochastic portions of the model and are sampled during simulation to incorporate risk.

– **Validation**

Validation of the model is a never ending process that should begin when the first diagram is drawn. All equations and relationships must be verified individually and in total to insure a logical and accurate series of calculations. Validation is treated with the weight it deserves in Chapter 3.

– **Documentation of the Model**

Document the model's capabilities and data requirements prior to saying the model is complete. Time fades memory and without documentation a model falls into a non-usable state as we forget how it was developed, verified, and validated.

Two demonstration spreadsheet simulation models are provided to show the differences in simulation models. Printouts for the demo programs are at the end of this chapter and the actual Excel workbooks are included in the CD. To view the programs on your computer use the Windows Explorer to open the CD and then double click on the name of the workbook to review.

- Deterministic Demo.XLS is a simple deterministic spreadsheet for a farm and contains a very rudimentary set of financial statements. Review the program to see if you can identify the different components of the "steps to actual model development."
- Cotton Model Demo.XLS is a deterministic model of the US cotton sector. The model uses a supply and utilization baseline from FAPRI in Table 1 and a set of assumed elasticities in Table 2 to simulate the US cotton sector for six years. The output variables are in Table 3 and constitute the supply and the demand components for cotton. All equations are solved using either identities or percentage changes from the baseline and the appropriate elasticities. The model follows the POLYSIM format developed by Ray and Richardson. Key output variables for the model are summarized as graphs on the second page of the printout. The exogenous variables the user can change are in the first two lines of Table 3 along with the elasticities in Table 2. The scenario analyzed is for a 5 percent sustained increase in cotton yield, beginning in 2000.

References

- Adams, Gary M. "Impact Multipliers of the U.S. Crops Sector: A Focus on the Effects of Commodity Interaction." Vol. I, II, University of Missouri-Columbia, Department of Agricultural Economics, Ph.D. Dissertation, 1994.
- Agrawal, R.C. and E.O. Heady. Operations Research Methods for Agricultural Decisions. Ames, IA: The Iowa State University Press, 1972.
- Backus, G.B.C., V.R. Eidman, and A.A. Dijkhuizen. "Farm Decision Making Under Risk and Uncertainty." Netherlands Journal of Agricultural Science, 45-2(July 1997): 307-328.
- Brown, D. Scott. "A Structural Approach to Modeling the U.S. Livestock and Dairy Industries: With Emphasis on the Dynamics Within the System." University of Missouri-Columbia, Department of Agricultural Economics, Ph.D. Dissertation, 1994.
- Fleisher, B. Agricultural Risk Management. Boulder, CO: Lynne Rienner Publishers, Inc., 1990.
- Hardaker, J.B., Huirne, R.B.M., and J.R. Anderson. Coping with Risk in Agriculture. Wallingford, UK: CAB International, 1997.
- Jones, G.T. Simulation and Business Decisions. Middlesex, England: Penguin Books Ltd., 1972.
- Law, A.M. and W.D. Kelton. Simulation Modeling and Analysis. Third Edition. New York: McGraw-Hill, Inc., 2000.
- Vose, D. Risk Analysis: A Quantitative Guide. New York: John Wiley & Sons, Ltd., Second Edition, 2000.
- Winston, W.L. Simulation Modeling Using @Risk©. New York: Duxbury Press, 1996.

Chapter 3

Model Validation

Model validation is the process modelers use to check simulation models for completeness, accuracy, and forecasting ability. The process consists of two parts: verification and validation. Verification is the mechanical process of testing every equation in the model to insure that it calculates correctly and checking the logic of the model to insure all equations are properly specified. Validation is the process of testing the accuracy of random variables and forecasts generated by the model.

Verification

Verification is the process of verifying that all of the equations in the model appropriately calculate what they are supposed to calculate. In other words, verification checks that if $Y = 10.8 + 13.4 X_1 - 3.8 X_2$ then the answer 260.8 is observed when $X_1 = 30$ and $X_2 = 40$. Verification involves the following:

- All equations must be checked by hand to insure arithmetic accuracy.
- The equations must be checked to insure that the correct variables are included and multiplied by the appropriate coefficients and added or subtracted correctly. (Excel provides tools for checking the dependence of equations; see Appendix B for help in using the trace precedents and dependence features in Excel.)
- All equations must be checked to insure that the variables are theoretically correct and have the right signs.
- The linkage between equations must be checked to insure they are in the right order; insure that answers from one equation become input in subsequent equations. Newer versions of Excel are more particular that the equations must be in a logical order from the top to the bottom.
- The time step incrementor needs to be checked, particularly if the model simulates across periods, i.e., weekly equations in an annual model or multiple year models. In a spreadsheet model time is implicit in the columns or rows, care should be used to insure that values for year 2 are not used as input in year 1, and so on.

Validation

Validation is the process used to insure that the random variables are simulated correctly and demonstrate the appropriate properties of the parent distribution. Once this first phase of validation is complete the overall forecast ability of the model must be checked. Most times the KOVs are not easily observed so the analysts must also validate the forecasts of variables used to calculate the KOVs. Validation involves answering the following questions:

- Do the random variables have the appropriate means, variances, and correlation?
- Does the model accurately forecast the system being analyzed?
- Do the results conform to theoretical expectations?
- Do the results conform to expectations of experts?

Van Horn suggested the use of a Touring Test – where the simulation results of a model are given to prospective clients, scientists, and other knowledgeable individuals to allow for evaluation of theoretical soundness and agreement with independent parallel research in the area. The Touring Test is an informal validation test but is useful because experts and prospective model users provide their own tests as to the validity of the forecast. This test has been used to test business simulation models such as: POLYSIM (Ray and Richardson), FLIPSIM (Richardson and Nixon), Ice Plant model (Richardson and Mapp), Farm Assistance (Klose), and Financial and Risk Analyzer (Gray).

Steps for Model Verification and Validation

– Validate the Stochastic Component

The stochastic component of a simulation model should be verified and validated first. This phase of validation is the easiest as it can make use of statistical tests to insure that the simulated variables are from the appropriate probability distribution, i.e., have the correct means, variance and correlation. The first step is to simulate the model and collect the simulated stochastic variables for the 100 plus iterations. The statistical tests and graphs used for validation are listed here and then described in detail in later sections of this chapter.

- Means -- test that the mean of each simulated variable equals its respective assumed or projected mean.
 - Univariate means test uses a Student's t test to determine if the simulated mean equals the assumed mean used for the random variable.
 - Multivariate probability distribution means test uses Hotelling's T-Squared test to simultaneously test whether the simulated vector means for the multivariate distribution are statistically equal to the vector of means for the original distribution.
- Variance -- test that the variance of each simulated variable equals the assumed variance used for the simulated variable.
 - Univariate variance test uses an F test to determine if the simulated variance equals the assumed variance used for the random variable.
 - Multivariate variance test uses Box's M test of homogeneity for covariances to simultaneously test whether the covariance of the simulated multivariate distribution equals the covariance of the original multivariate distribution.
- Correlation -- for a multivariate (MV) probability distribution, the historical correlation matrix used to simulate the MV distribution can be tested against the simulated variables to determine if they are appropriately correlated. A Student's t test for each of the coefficients in a correlation matrix is used for this test.
- Coefficient of variation -- the coefficient of variation should either be equal to its historical value or changing over time as specified in the model. Visual inspection is all

we have at this time to test this statistic. If expansion factors are used to alter the coefficient of variation over time, this fractional change from year-to-year must be checked.

- Minimum and Maximum -- the simulated min and max values should be equal to the historical data or their respective assumed input values. Visual inspection is all we can do to validate these statistics. This is critical for truncated distributions and normal distributions that are capable of simulating values outside the realm of our priors.
- Charts of the simulated stochastic variables vs. the historical data can be developed to help verify the simulated variables. Charts useful for this purpose are:
 - CDF or cumulative distribution functions can be developed of the simulated values and the historical data to insure that they have the same shape throughout the range of the data.
 - PDF or probability distribution functions can be developed for the simulated values and the historical data to check for similar shapes.
 - Fan graph to show how the relative variability of a stochastic variable changes over time. This is particularly useful for a variable that is simulated for more than two years.
 - Probability plots of the historical vs. the simulated data. Both QQ and PP plots can be used for this purpose.
 - Box plot charts of the simulated and historical data for each variable can be compared to see if the simulation process skewed the variable's results.

– Verification of Equations

The equations in the model must be verified one at a time. Start the process by setting the model to expected value mode so all stochastic variables equal their means. (In Simetar this can be done by clicking the Expected Value button on the Simetar tool bar.) In expected value mode all of the equations in the model are supposed to equal their means. If the equations do not equal their means based on your calculations with a hand-held calculator find out why.

A note on expected value is worth while at this point. If the model uses only normally distributed random variables, then all stochastic variables will equal their means in expected value mode. However, if the stochastic variables are distributed empirical these variables will not necessarily equal their means but may be slightly larger or smaller than the mean. The reason for this result is that in expected value mode, the mean USD of 0.5 does not have to correspond to zero on the inverse function for the empirical distributions. See Chapter 5 for a discussion of the empirical distribution and inverse transform method for simulating random variables.

In the verification process begin by checking the most basic equation and then working systematically through the model to the last KOV. In a business model the order for verifying

the equations would start with production and then move to local price, total receipts, then on to variable costs, fixed costs, total costs, and net cash income. The final equations to verify are the KOVs such as net worth, present value of ending net worth, net present value, and rate of return to equity.

Excel provides tools for verifying that the appropriate cells are linked via the trace precedence tools. See Appendix B for adding these icons to the tool bar. Once in place these icons can be used to verify where the inputs to an equation come from and where the results are used in subsequent equations in the model. The trace precedence tools allow you to trace a cell's inputs for 1, 2, 3, or more generations back and forward.

– Sensitivity Tests for Validation

Sensitivity tests show how sensitive the results for the KOV are to key inputs variables. Simulating the model with a key input variable at alternative levels and comparing the statistics for the KOVs shows how critical the variable is to the model. If the KOVs change greatly as the input variable changes, then more attention needs to be paid to forecasting that input variable or specifying its probability distribution. Sensitivity analysis is a method for determining which variables need the most attention when econometrically estimating the parameters for the probability distribution. If the means for the KOVs jump dramatically, or do not change as expected from one assumed value of the input variable to another, it can indicate an error in the model.

It is recommended that you conduct sensitivity tests on the means and standard deviations to determine the range of values under which the model produces reliable results, particularly if the stochastic variables are distributed normal. Alternative means for the stochastic variables should be tested to insure that the model is capable of simulating the full range of means available. For each mean tested, one should re-check the means tests described above.

With the Simetar Simulation Engine set to stochastic simulation mode the model can be simulated for a range of values on one input variable at a time. The Simulation Engine menu contains an option named Conduct Sensitivity Analysis. When the sensitivity analysis option is selected the user may select one input cell to manipulate over three ranges.

The selected input cell can be changed by, say, $\pm 3\%$, $\pm 6\%$ and, $\pm 9\%$ or any other fractions. The model is simulated for each of the settings specified for the variable with all other non-stochastic values in the model held constant. (See Chapters 10 and 16 for a complete description of sensitivity analysis with Simetar.) The user should collect statistics on the model's KOVs to see how they change as the specified test cell changes.

– Validation Using External Reviews

As a formal part of validation it is highly recommended that the model and its results be reviewed externally. Start by showing the results to experts who are used to reviewing or using the types of values the model forecasts, or a Touring Test. Seek experts in the field to look at particular parts of the forecast initially. Show them more and more of the results as they become familiar with the model. Most experts will want to know more about the model that generated

the forecast, once they have decided that it is worthwhile. At that time the model can be presented, first in terms of the theoretical assumptions and finally the details on how the model is designed and how it works.

To be meaningful a Touring Test should be designed as follows:

- Show the experts the assumed input data and the results for the KOVs.
- Ask their opinions as to the reasonableness of the forecast in terms of direction and magnitude for the KOVs, given the input assumptions.
- Change the input assumptions and repeat the process a second or third time.
- In the second and subsequent trials show more of the results used to calculate the KOVs. If the KOVs are standard financial ratios then show the pro forma financial tables.

The 4 Ps for Simulation Model Validation

By way of summarizing, model validation is a personal activity and is solely the responsibility of the model developer. The following simple nomenclature is offered to help modelers remember it is their responsibility to verify and validate their models.

- **Planning:** Developer(s)
 - Initial preparation for building a model must include plans for validation
- **Personal:** Developer(s) must verify every equation and their linkages
 - Logical sequence and interaction of components must be checked
 - Econometric and accounting equations tested for logic and goodness of fit
 - Do simulated PDFs reproduce their parameters?
 - Are the results consistent with theory?
- **Peers:** Other experts should be utilized for reviewing the model results
 - Tour the output results among experts
 - Model testing and use by practitioners trained in modeling
 - Apply model to different situations, use sensitivity tests
 - Review in professional journals
- **Producers:** Farmers, ranchers, commodity experts, and agribusiness managers are the ultimate judges as to whether a model is correct
 - Compare model results to observations by experts
 - Do projections of pro forma financial statements agree with the expert's expectations?
 - Biophysical growth model, do the results agree with field tests
 - Similarly, sector level models must perform relative to the expectations of commodity experts and policy analysts, consult an array of experts

Validation Never Ends

Each time the model is changed the modeler must repeat the verification and validation process. How extensive the re-validation process depends on how extensive were the model

changes. A simple rule is to repeat all of the verification steps and at least part of the validation steps.

Based on my observations, it takes a lifetime to build a good reputation and one bad study to lose it. Do not underestimate the importance of model validation. It is a continual process that successful modelers work at all the time.

Validation Statistics

Often times we have a historical data series with 8 to 20 observations that defines the probability distribution for random variable. After simulating the variable it is necessary to determine whether the simulated series is statistically equal to the parent or historical distribution. Put another way, if the simulation model does not reproduce the statistics for the random variables, the model is not validated and can not be used for decision making.

Statistical tests can be used to validate the stochastic variables in a simulation model. If the random variables are simulated as univariate probability distributions then Student-t tests and F tests are appropriate for testing the means and variances, respectively. When the random variables are simulated as a multivariate distribution then the Hotelling T-Squared test and Box's M test for homogeneity of covariances are used to simultaneously test the means and variances, respectively. The complete homogeneity test is a nonparametric test of the means and covariance for the simulated distribution versus the input distribution. Another test for multivariate probability distributions is to use a Student's-t test for each coefficient in the correlation matrix of the simulated data.


In addition to quantitative tests to validate the stochastic variables, it is useful to check the skewness and kurtosis of the simulated variables relative to their historical values. Graphical comparisons of the historical and simulated distributions are also useful for validating the process that generated the random variables.

– Univariate Distribution Validation Tests

- **Student-t test of Univariate Means**

Test that the simulated mean (\bar{X}_s) is not statistically different from the original mean (\bar{X}_h) using a Student's-t test. The null and alternative hypotheses are:

$$\begin{aligned} H_O: \bar{X}_s &= \bar{X}_h \\ H_A: \bar{X}_s &\neq \bar{X}_h \end{aligned}$$

The Simetar icon  (or menu item) for hypothesis tests can be used to perform a Student's-t test on the significance of the difference between the means of two distributions. Select the Compare Two Series tab in the Hypothesis Testing for Data dialog box. Three different Student's-t tests of the means are demonstrated in Validation Tests Demo.XLS and are summarized in Figure 3.1:

- Test 2 compares means of two series simulated with different distributions.
- Test 3 compares means for the historical data and the simulated random variable.
- Test 5 tests the mean of a simulated distribution against an assumed mean value.

In Simetar hypothesis tests, an interpretation of the test is provided (Figure 3.1). For example, Test 2 results in Validation Tests Demo.XLS are reported as: “Fail to Reject the H_0 that the Means are Equal.” The test statistic and the critical p-value statistic are provided for the tests. If the calculated t-statistic exceeds the critical value then one fails to reject the alternative hypothesis that the means are different.

Test 2.
Distribution Comparison of Two Data Series
 Confidence Level 95%

	Test Value	Critical Value	P-Value
2 Sample t-	-0.01	2.24	0.99 Fail to Reject the Ho that the Means are Equal
F-Test	1.13	1.16	0.09 Fail to Reject the Ho that the Variances are Equal

Test 3.
Compare the original data to the simulated data.
Distribution Comparison of Two Data Series
 Confidence Level 95%

	Test Value	Critical Value	P-Value
2 Sample t-	0.00	2.69	1.00 Fail to Reject the Ho that the Means are Equal
F-Test	1.13	1.90	0.34 Fail to Reject the Ho that the Variances are Equal

Test 5.
Test the Parameters for a Simulated Series Against Assumed Mean and Standard Deviation
 Assumed Mean 9.8
 Assumed Std Dev 6.7
Test of Hypothesis for Parameters
 Confidence Level 95%

	Given Value	Test Value	Critical Value	P-Value
t-Test	9.8	-0.01	2.25	0.99 Fail to Reject the Ho that the Mean is Equal to 9.8
Chi-Square	6.7	505.87	439.00	0.81 Fail to Reject the Ho that the Standard Deviation is Equal to 6.7


Figure 3.1. Summary of Univariate Tests for Means, Variances, and Standard Deviations.

• **F-test of Univariate Variances**

Test whether the simulated variance ($\hat{\sigma}^2$) is not statistically different from the original variance (σ^2) using an F-test. The null and alternative hypotheses are:

$$H_O: \hat{\sigma}^2 = \sigma^2$$

$$H_A: \hat{\sigma}^2 \neq \sigma^2$$

The Simetar icon  (or menu item) for tests can be used to perform an F-test of two variances. Select the Compare Two Series tab in the Hypothesis Testing for Data dialog box. In this case the F-test is testing whether the two variances are equal, or that the F ratio equals 1. Recall the F-test calculates an F-statistic as:

$$F\text{-statistic} = \hat{\sigma}^2 / \sigma^2$$

with $\hat{n} - 1$ and $n - 1$ degrees of freedom, where
 $\hat{\sigma}^2$ is the simulated variance and
 \hat{n} is the number of iterations, while
 σ^2 is the historical variance and
 n is the number of historical observations.

This is a two-sided F-test so if the calculated F-statistic exceeds the tabled F-value, we reject the null hypothesis and conclude that the two variances are different.

Test 2 in the Validation Test Demo.XLS workbook and Figure 3.1 provides an example of using the F-test to validate the variance for two distributions. In Test 3 of the Validation Test Demo.XLS workbook, the first distribution was specified as the original or historical data, while the second series was specified as the results of stochastically simulating the distribution as an Empirical distribution. In the demo, the F test failed to reject the null hypothesis that the original and simulated variances are equal. This combined with the 2 Sample Student's-t test that failed to reject that the means are equal, indicate that the simulation package accurately simulated the mean and variance for the random variable.

- **Chi-Squared Test of Univariate Standard Deviation**

Simetar includes a univariate Chi-Squared test for validating the standard deviation. The Test Parameters tab in the Hypothesis Testing for Data dialog box provides a Chi-Squared test of the standard deviation for a random variable against a specified value. Test 5 in Validation Test Demo.XLS demonstrates how to test the simulated data against a specified mean and standard deviation (Figure 3.1). The mean is tested using a Student's-t test.

The Chi-Squared test is used when the analyst does not have a historical data series to validate simulated values against. In this case the analyst must assume a mean and standard deviation, for example, $N(9.8, 6.7)$ in Test 5 for Figure 3.1. The Chi-Squared and Student's-t tests are used to validate that the simulated random variables, statistically, have the assumed mean and standard deviation. Test 5 in Figure 3.1 indicates that the simulated values have a mean and standard deviation that are statistically equal to the assumed values, at the alpha equal 5% level.

- **Multivariate (MV) Distribution Validation Tests**

- **Hotelling T-Squared MV Means Test**

The 2-Sample Hotelling T-Squared MV means test is used to statistically test if the collective means of a multivariate random sample comes from the same distribution as the historical distribution. This test uses the Mahalanobis distance or:

$$M = \sum_{i=1}^N (\bar{X}_{si} - \bar{X}_{hi})^2$$

where: \bar{X}_{si} represents the means of the N sampled (simulated) variables, and \bar{X}_{hi} represents the means of the N variables in the historical distribution.

The null and alternative hypotheses are:

$$H_0: \bar{X}_s = \bar{X}_h$$

$$H_A: \bar{X}_s \neq \bar{X}_h$$

Simetar calculates the Hotelling T-Squared test when the user specifies an MxN matrix for the 1st series and a PxN matrix for the 2nd series, for the Compare Two Series tab in the Hypothesis Testing for Data dialog box. In this case N represents the number of variables in both multivariate distributions, M represents number of historical observations, and P represents the number of iterations simulated. See Figure 3.2 and Test 7 in Validation Test Demo.XLS for an example of a multivariate means test. If the results of the test returns, “# Value” check for the following: singular covariance matrix, excessively long worksheet name, or excessively long workbook name.

Distribution Comparison of Pr & Sim Pr			
Confidence Level	95%		
	Test Value	Critical Val	P-Value
2 Sample Hotellin	0.00	13.76	1.000 <i>Fail to Reject the Ho that the Mean Vectors are Equal</i>
Box's M Test	1.07	32.67	1.000 <i>Fail to Reject the Ho that the Covariance Matrices are Equivalent*</i>
Complete Homog	1.08	40.11	1.000 <i>Fail to Reject the Ho that the Mean Vectors and Covariance Matrices are Equivalent, Respectively*</i>
<i>*Based on asymptotic distribution</i>			
Summary Statistics			
Pr			
Means	Covariance		
2.343	0.13209	0.272777	0.253247
5.887	0.272777	0.712801	0.594487
2.196	0.253247	0.594487	1.486227
137.4	7.440889	14.58022	25.36511
47.4	1.529778	3.991333	6.979556
119.9	1.657667	3.245444	-11.6142
			7.4408889
			1.529778
			3.991333
			3.245444
			-11.61422
			154
			202.3778
			112.9333
			-16.37778
			419.5667
Sim Pr			
Means	Covariance		
2.343068672	0.101692	0.199688	0.186812
5.887748123	0.199688	0.525034	0.427508
2.195439382	0.186812	0.427508	1.466344
137.3952958	5.270698	9.530387	22.06672
47.36071982	0.812737	2.361326	6.388249
119.9111196	1.21564	2.34509	-14.0917
			5.2706977
			0.812737
			2.361326
			6.388249
			166.36989
			107.2541
			-31.89461
			388.2429

Figure 3.2. Summary of Multivariate Tests for Means and Variances.

- **Box’s M Test for Homogeneity of Covariances**

Box’s M test is used to statistically test if the covariance of a multivariate random sample comes from the same distribution as the historical distribution. The test is a maximum likelihood ratio test of the covariance matrix. The null and alternative hypotheses are:

$$H_0: \Sigma_s = \Sigma_h$$

$$H_A: \Sigma_s \neq \Sigma_h$$

where: Σ_s represents the covariance for the multivariate distribution sample, and Σ_h represents the covariance for the historical multivariate distribution.

Simetar calculates the Box's M test when the user specifies an MxN matrix for the 1st series and a PxN matrix for the 2nd series, for the Compare Two Series tab in the Hypothesis Testing for Data dialog box. In this case N denotes the number of variables in both multivariate distribution and P represents the number of iterations simulated. See Figure 3.2 and Test 7 in Validation Test Demo.XLS for an example of a multivariate variance test. If the results of the test returns "# Value" check the following: singular covariance matrix, excessively long worksheet name, or excessively long workbook name.

- **Complete Homogeneity Test**

A multivariate distribution can also be tested using the nonparametric, complete homogeneity test. The mean vector and the covariance matrix for the simulated values is compared simultaneously to their corresponding values in the historical data series. An example of the complete homogeneity test is presented in Figure 3.2. As indicated by the name this is the most comprehensive of the multivariate distribution tests for validating the stochastic variable generation process in a simulation model.

The null and alternative hypotheses are:

$$H_0: \bar{X}_s = \bar{X}_h$$

$$H_A: \bar{X}_s \neq \bar{X}_h$$


$$H_0: \Sigma_s = \Sigma_h$$

$$H_A: \Sigma_s \neq \Sigma_h$$

where: \bar{X}_{si} represents the means of the N sampled (simulated) variables,
 \bar{X}_{hi} represents the means of the N variables in the historical distribution,
 Σ_s represents the covariance for the multivariate distribution sample, and
 Σ_h represents the covariance for the historical multivariate distribution.

- **Correlation Matrix Test**

When simulating a multivariate probability distribution you should also validate the correlation coefficients between the simulated variables and the assumed correlation matrix. The correlation coefficients among the simulated variables should be statistically

equal to the original or historical correlation coefficients for the random variables. A Student's-t test to determine if the correlation coefficients for two matrices are statistically equal is provided in Simetar under the hypothesis testing icon  , Hypothesis Test for Data in the Check Correlation tab. The null and alternative hypotheses tested using a Student's-t test are:

$$H_0: \hat{\rho}_{ij} = \rho_{ij}$$

$$H_A: \hat{\rho}_{ij} \neq \rho_{ij}$$

where: $\hat{\rho}_{ij}$ is the individual correlation coefficient between the simulated variables i and j, ρ_{ij} is the assumed correlation coefficient between variables i and j used to simulate the multivariate distribution.

Individual Student's -t tests are performed for all $j>i$ correlation coefficients in the upper right triangle matrix. Thus a distribution with 5 variables has 10 t-tests and a distribution with 6 variables requires 15 t-tests.

See Test 6 in the Validation Test Demo.XLS workbook for an example of testing an input correlation matrix against the implicit correlation among the variables in the simulated data (Figure 3.3). Test statistics (Student-t) less than the critical value indicate that the correlation coefficient for the simulated data is statistically not different from the original value at the indicated confidence level. For example, the simulated correlation coefficient of 0.54 was statistically equal to 0.50 based on a test statistic of 0.99 and a critical value of 1.96 at the alpha equal 5% level (Figure 3.3).

Test Correlation Coefficients				
Confidence Level	95%			
Critical Value	1.96			
	Variable 2	Variable 3	Variable 4	
Variable 1	0.39	0.68	1.36	
Variable 2		0.99	1.10	
Variable 3			0.29	

Figure 3.3. Example of Testing a 4x4 Correlation Matrix.

Other Statistical Tests for Validation

The second and fourth moments of the simulated random variables can be tested against their historical distributions. The third moment about the mean is skewness and the fourth moment is kurtosis. These statistical values can be calculated and compared to the historical distribution.

– Skewness Test

The skewness or left-right orientation of the distribution can be easily compared using Excel's =SKEW() function. The skewness of the raw data which was used to estimate the

parameters for a distribution can be estimated and compared to the skewness of the simulated values. For example, if the distribution for X is defined as:

$X = (1, 2, 4, 6, 8, 10, 13, 16, 18, 20)$ which yields a \bar{X} of 9.8 and $\hat{\sigma}$ equals 6.74.

(See Validation Test Demo.XLS for the distribution and the tests summarized here.) Skewness for X is equal to 0.2134 and is calculated using the Excel command: =SKEW(Array of Xs). The skewness for 500 simulated values of X that are generated assuming X is distributed normal, is -0.0209 and was calculated by the Excel command: =SKEW(Array of 500 \tilde{X} s). The hypothesis testing option in Simetar calculated skewness for the original distribution and the simulated distribution for direct comparison (lines 26-40) in Validation Test Demo.XLS.

Skewness is calibrated relative to a normal distribution which has no (or zero) skewness. A negative skewness indicates the distribution leans to the left. A positive skewness indicates a distribution that leans to the right. Using skewness for validation is limited, as there are no statistical tests to determine if the skewness for the simulated data equals the historical value. Visual inspection is the only way to check the skewness. Be forewarned that this statistic is particularly sensitive to the sample size. At small sample sizes of 100 or less this statistic is not reliable. As an experiment simulate $X \sim \text{Normal}(10,3)$ for 100 iterations and calculate the skewness, it is -0.3; considerably different from 0.0.

– Kurtosis Test

Kurtosis quantifies the peakedness of a distribution, relative to a normal distribution. Kurtosis is zero for a normal distribution and positive for one which is taller than the normal distribution. A negative kurtosis is associated with a distribution which is shorter than the normal and its tails are thicker than the normal. Excel provides a function for calculating kurtosis =KURT(Array of Xs).

Using the X distribution defined in Validation Test Demo.XLS, the original data has a kurtosis of -1.391. The simulated sample of Xs that is assumed normal has a kurtosis of -0.0208 and the empirical sample has a kurtosis of -1.345. There is no test statistic to compare the original and simulated kurtosis values, but -1.34 is closer than -0.02 to the original -1.39. The distribution comparison statistics, kurtosis and skewness, can be calculated by Simetar through the Compare Two Series tab in the Hypothesis Testing for Data icon (see Chapter 16). At small sample sizes of 100 or less the kurtosis statistic is not reliable. As an experiment simulate $X \sim \text{Normal}(10,3)$ for 100 iterations and calculate the skewness, it is 0.84; considerably greater than 0.0.

Graphical Tools for Validation

Simple graphs of the simulated random variables results are an excellent means of validating the random generation procedure used in the simulation model. Cumulative distribution, probability density, box plots, fans graphs and normality charts are useful tools for visually comparing the simulated random variables. Graphical tools should not be used in place of statistical validation tests, but they can be used to supplement the statistical tests described in the previous sections.

– CDF Graphs

The original data and its corresponding simulated values can be graphed as a cumulative distribution function (CDF) graph on a common axis. (See Chapter 16 to learn how this graph is developed in Simetar.) The graph (line) for the original data will lie very close to the simulated data's CDF line, if the data were simulated with the "correct" distribution. As an example, the distribution defined by $X = [2, 5, 8, 12, 14, 18, 22, 35]$ was simulated empirical and normal and the results are presented as a CDF graph in Figure 3.4. Assuming the data are normally distributed would be an error based on how far the normal distribution misses the true distribution. The simulated means are the same as the true distribution and the standard deviations are statistically the same as the true data, but the CDF chart suggests the empirical is a better assumed distribution than the normal.

A loss function to measure the difference between the CDF for the true distribution and an assumed distribution is available in Simetar. The `=CDFDEV(array of historical Xs, array of simulated Xs)` returns a scalar which measures the difference between the two distributions, even if they are of different lengths. If the simulated data exactly reproduce the actual distribution the CDFDEV value will be zero. If several different distributions are tested, use the one with the smallest CDFDEV value.

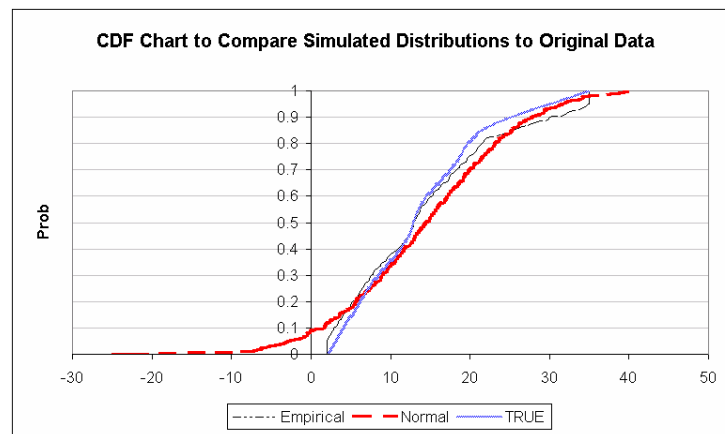


Figure 3.4. Example of Comparing the True Distribution's CDF to Simulated Distributions for Validation.

– PDF Graphs

The true distribution and the simulated data can be compared using a probability density function (PDF) graph (Figure 3.5). As demonstrated here the means appear quite similar, however, the lower tails on the normal distribution is too long to adequately represent the true series for a simulation model. See Chapter 16 for a description of how to develop PDF charts with Simetar.

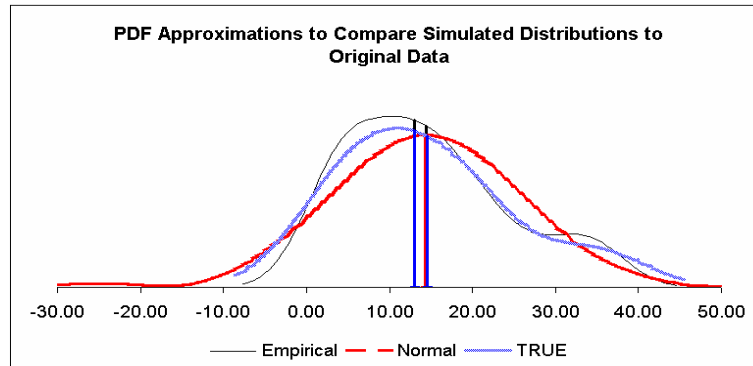


Figure 3.5. Example of Comparing Distributions Using a PDF Chart.

– Box Plots

The dispersion and quartile relationships for the true data and simulated values can be compared using Box plots. See Chapter 16 to learn how to develop Box plots in Simetar. The example in Figure 3.6 suggests that the normal distribution greatly overstates the downside risk and slightly overstates the upside risk, relative to the true distribution.

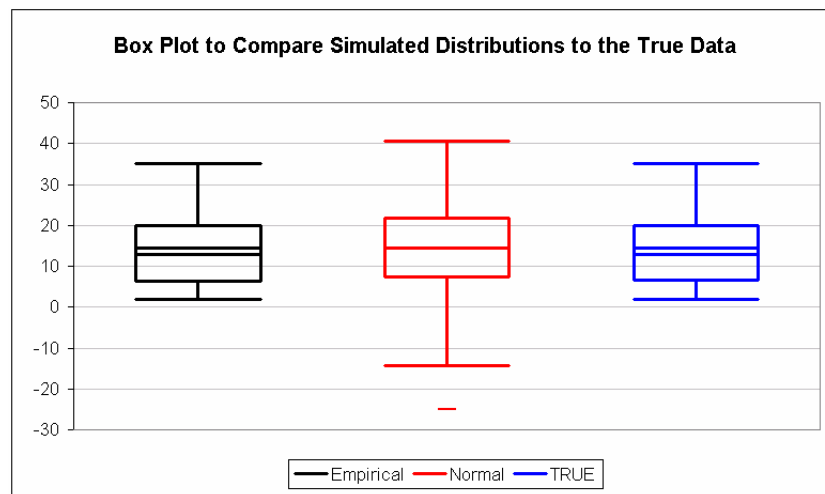


Figure 3.6. Box Plots for Comparing Two Simulated Distributions Against the True Distribution.

– Normality Plots

If a series is simulated assuming a normal distribution, the simulated random values should be tested for normality. An easy way to do this is to draw a normality plot using Simetar (see Chapter 16). If the simulated data conform to Normality the observations will be on a straight

line through the chart area (Figure 3.7). In contrast, the Normality plot for data that are not Normally distributed will resemble Figure 3.8.

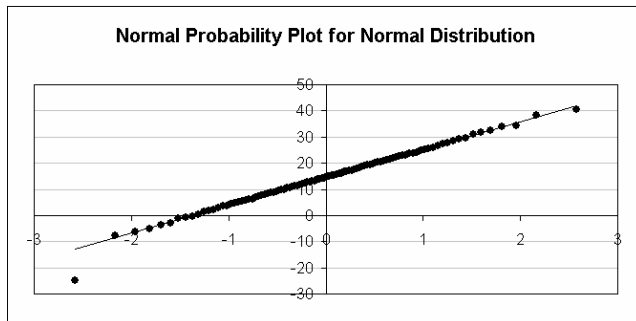


Figure 3.7. Normality Plot for Normally Distributed Data Series.

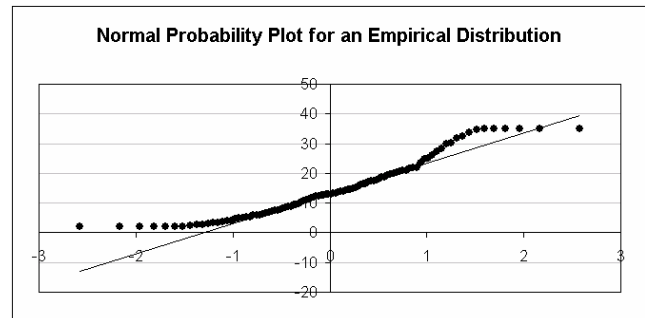


Figure 3.8. Normality Plot for Data Series Distributed Empirical.

Normality Test

Oftentimes analysts assume normality for the distribution of random variables. Recent research by Feldman, Richardson and Schumann suggest that when the analysts have no prior knowledge about the distribution, normality is the best default distribution to use. Five different normality distribution tests are provided in Simetar. Chapter 16 contains instructions on how to do a normality test in Simetar. The five normality tests in Simetar are:

- Chi Squared,
- K-S – Kolmogorov-Smirnoff,
- S-W – Shapiro-Wilks
- A-D – Anderson, Darling, and
- CvM – Cramer-von Mises

Examples of the normality tests are provided in Test 1 of Validation Test Demo.XLS. If sample size prevents the test from being reliable the letters “NA” will appear for the p-value. Tests for normality seldom reject the null hypothesis that the data are normally distributed, so be cautious in using the tests. See the Simetar Help file, Hypothesis Testing for Data, for a description of the normality tests.

References

- Fleisher, B. Agricultural Risk Management. Boulder, CO: Lynne Rienner Publishers, Inc., 1990.
- Feldman, P., J. W. Richardson, K. Schumann. "Distribution Choice under Null Priors and Small Sample Size." Department of Agricultural Economics, Texas A&M University, (mimeo) 2004.
- Gray, A.W. "Agribusiness Strategic Planning Under Risk." Texas A&M University, Department of Agricultural Economics, Ph.D. Dissertation, August 1998.
- Hardaker, J.B., Huirne, R.B.M., and J.R. Anderson. Coping with Risk in Agriculture. Wallingford, UK: CAB International, 2004.
- Klose, S.L. "A Decision Support for Agricultural Producers." Texas A&M University, Department of Agricultural Economics, Ph.D. Dissertation, 2001.
- Law, A.M. and W.D. Kelton. Simulation Modeling and Analysis. Third Edition. New York: McGraw-Hill, Inc., 2000.
- Ray, D.E. and J.W. Richardson. "Detailed Description of POLYSIM." Agricultural Experiment Station, Oklahoma State University and USDA, Technical Bulletin, T-151, December 1978.
- Richardson, J.W. and C.J. Nixon. "A Description of FLIPSIM V: A General Firm Level Policy Simulation Model." Bulletin 1528, Texas Agricultural Experiment Station, July 1986.
- Richardson, J.W. and H.P. Mapp, Jr. "Use of Probabilistic Cash Flows in Analyzing Investment Under Conditions of Risk and Uncertainty." Southern Journal Agric. Econ. Dec. 1976.
- Vose, D. Risk Analysis: A Quantitative Guide. New York: John Wiley & Sons, Ltd., Second Edition, 2000.

Chapter 4 Stochastic Simulation

Simulation models that do not include risk produce a deterministic (or predetermined) result given the input values. Stochastic models are deterministic simulation models that include variables which are not known with certainty but have a known probability distribution. A stochastic model is simulated a large number of times using randomly selected values for the risky variables to estimate the probable outcomes for key output variables (KOVs). The simulated sample of values for each KOV constitutes an estimate of the variable's probability distribution which can be used to make decisions in a risky environment.

Stochastic simulation involves simulating uncertain economic systems that are a function of risky variables, for the express purpose of making better decisions. In stochastic models future risk is assumed to mimic historical risk, so past variability is used to estimate parameters for the probability distributions of risky variables in a model. Probability distributions are simulated a large number of times to formulate probabilistic projections for the risky variables. The interaction of the risky variables with other variables in the system allows the modeler to project how risky a decision would likely perform under alternative management strategies. In this way models can provide decision makers useful information about the likely outcomes of alternative management decisions under risk.

Because a stochastic simulation model is a deterministic model with one or more variables that have been made stochastic. Development of a stochastic model usually starts with developing a deterministic model and then converting it to be stochastic by making some of the exogenous variables stochastic. Even when a model is stochastic it can be used to generate deterministic results by making the risk components zero or setting the simulation engine to its "Expected Value" mode.

A simple stochastic simulation model has one output variable, one or more input variables subject to risk, and one or more fixed input variables. The simple model's flowchart would be:

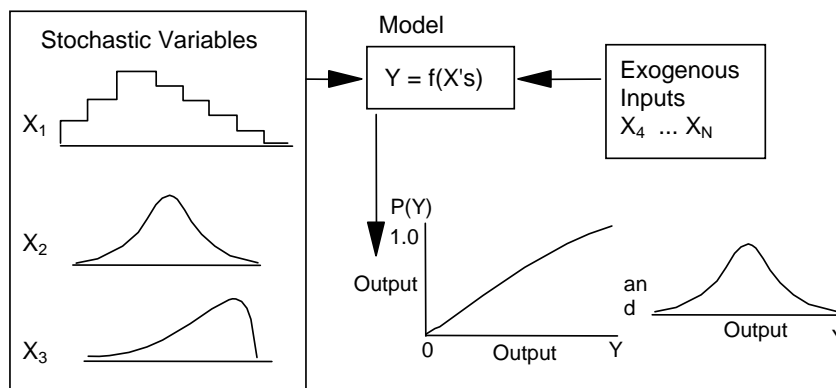


Figure 4.1. Schematic for a Stochastic Simulation Model

By sampling from the X_1 , X_2 , X_3 probability distributions a large number of times (or iterations), say, 500, the model can calculate enough values of Y to estimate the probability distribution for Y . The CDF and PDF graphs of the simulated Y output value are “estimates” of the true probability distribution for the output variable and are developed by graphing the simulated Y values over many iterations. This in itself is one of the **primary goals of simulation, the estimation of unobservable probability distributions through statistical sampling.**

For many people, the mystery of stochastic simulation is how do you get the 500 random values for the risky (stochastic) variables? That part is quite simple once you have done it. How the random values are generated depends on the computer language being used:

- Excel
The RAND function returns a single random value between 0 and 1. Each time F9 is pressed the value changes with another sample of the random number. To generate a sample of 100 random values from RAND, type or copy the RAND command into 100 cells. An example of using this procedure to generate random numbers is provided in Uniform Random Number Generator Demo.XLS.
- SIMETAR for Excel
The UNIFORM function returns a random number between zero and one. A sample of 100 values is generated by setting the number of iterations to 100, specifying the cell with =UNIFORM () as an output variable, and simulating the worksheet with SIMETAR.
- @Risk for Excel
The RISKUNIFORM function generates a random value between zero and one. A sample of 100 values in the cell is developed (simulated) by setting the number of iterations to 100, specifying the cell with =RiskUniform () as an output variable, and simulating the worksheet with @Risk©.

Random Variables in a Simulation Model

Any variable the manager or decision maker cannot control, can be made stochastic. A random variable may be partly determined by internal forces but still have a random component. Another type of random variable is a forecasted variable which has risk because the forecast is less than perfect. Variables affected by weather are also stochastic, e.g., yields for crops or livestock response to feed. Production can be stochastic due to risk caused by inconsistent quality of inputs and response to the inputs. Market share is often a stochastic variable because managers do not know how competitors will react to their promotional campaigns or pricing strategies. The quantity sold each period is a stochastic variable depending on many factors outside the manager’s control.

In all cases, stochastic variables are: crucial to the success of a business decision, out of the control of the decision maker, and can be specified by a probability distribution. If the variable is so risky that the decision maker has no idea what the probability distributions is, then that is uncertainty. (Uncertainty is treated at the end of this chapter.) Examples for five stochastic variables in a business model are listed below. Each variable is described in terms of

hypothetical parameters that could describe the probability distributions.

- Sales -- distributed normal with mean 1000 and standard deviation of 100
- Production -- distributed normal with mean $Y = a + b_1X_1 + b_2X_2$ and standard deviation of 31.0.
- Price -- distributed empirical over the range of 3.20, 3.40, 3.45, 3.50, 3.55, and 3.70.
- Interest rate -- distributed GRK with minimum of 0.05, mean equal 0.07 and maximum of 0.095.
- Market share -- distributed uniform with minimum of 0.40 and a maximum of 0.75.

The distributions assumed for sales and production contain a deterministic component and a stochastic component. The deterministic component for sales is the mean of 1000 and for production the deterministic component is the $a + b_1X_1 + b_2X_2$ mean. The stochastic components for sales and production is the standard deviation of 100 and 31, respectively. In a more general notation we would write production (Pd) as:

$$\tilde{P}d = a + b_1X_1 + b_2X_2 + \tilde{\epsilon}$$

where $\tilde{\epsilon}$ indicates the stochastic component.

Simulation engines such as SIMETAR provide an assortment of probability distributions or functions for simulating random variables. The purpose of providing different functions is that random variables can take on an assortment of distributions. For example, random variables can be distributed: uniform, normal, empirical, bernoulli, truncated normal, or triangular, to name a few. Because these functions are available the problem of generating random numbers in a simulation model is largely eliminated.

The Simetar functions for simulating (or generating) random numbers for more than 20 distributions are described in detail in Chapter 16. The Simetar functions that would be used to simulate the five sample stochastic variables above are presented next.

Sales simulated by = Norm (1000, 100)

Production simulated by = Norm ($a + b_1X_1 + b_2X_2$, 31.0)

Price simulated by = DEMPIRICAL (3.2, 3.4, 3.45, 3.5, 3.55, 3.7)

Interest rate simulated by = GRK (0.05, 0.07, 0.095)

Market share simulated by = Uniform (0.40, 0.75)

These random number functions are typed into cells in an Excel worksheet and every time the F9 key is pressed, Excel draws a new random value given the parameters for the variable. When Simetar simulates the workbook it generates 100 or more random values for the stochastic functions. Each set of random values is called an iteration (or state of nature). When the value for a random number changes, Excel automatically updates all equations that are dependent on the random variable cells. So when Simetar draws random values, Excel calculates dependent values for variables which use the stochastic variables, Simetar records values for the KOVs, and

then repeats the process for the next iteration. After the last iteration, Simetar calculates the statistics for each of the output variables.

It is recommended that stochastic variables be simulated (or generated) in a three step process so the experienced modeler can control the process and the student can learn how the process of generating random numbers really works. The three step process is described as follows for a variable that has a normally distributed error term:

- calculate the deterministic component of the random variable in one cell, or

$$\hat{Y} = a + bP$$
- simulate the stochastic component of the random variable in another cell, or

$$SND = NORM(0,1)$$
 where SND is a standard normal deviate,
- combine the deterministic and stochastic components to calculate the random number in a third cell, as

$$\tilde{Y} = \hat{Y} + (\hat{\sigma} * SND)$$
 where $\hat{\sigma}$ is the standard deviation about the deterministic forecast.

The multiple step process can be used to simulate any distribution, for an example of simulating various distributions see Uniform Random Number Generator Demo.XLS. A distinct advantage to using this multiple step process is that each part of the process can be verified independently during the verification/validation process.

A stochastic simulation model is simply a set of equations that define an economic model with at least one random variable. The random variable(s) can be simulated by a uniform distribution, a normal distribution or any other distribution that emulates their historical variability. The process of simulating random numbers in the model is determined by the computer language or engine being used for the model. The random variables can be generated in Excel as demonstrated in Uniform Random Number Generator Demo.XLS or in an Excel spreadsheet using Simetar or @Risk© functions. In either case we should focus more on what is in the model, how it works and the parameters that define the probability distributions and less on how to generate the random values.

Iterations

An iteration is one realization of a stochastic model, one state of nature, one roll of the dice, or one deal of the cards. Another way to think of an iteration is that it represents one solution for all equations in a model using one draw (or sample) of random values for all random variables. Drawing another sample of random values and re-calculating all of the equations in the model constitutes a second iteration. Note that the model must be designed and simulated so the only change from one iteration to another is the new random values for each of the random variables. All parameters must remain constant across iterations and none of the results of one iteration can be used as input into the next iteration.

The flowchart in Figure 4.2 demonstrates how a simulation engine simulates a model for 500 iterations. The model in Figure 4.2 uses 500 random prices and sales quantities to produce 500 possible P_r or profit values. The 500 P_r values constitute an estimate of the unobservable probability distribution for P_r . The results from the 500 simulated values for P_r can be summarized by way of sample statistics, a histogram, or a CDF, as indicated in Figure 4.3.

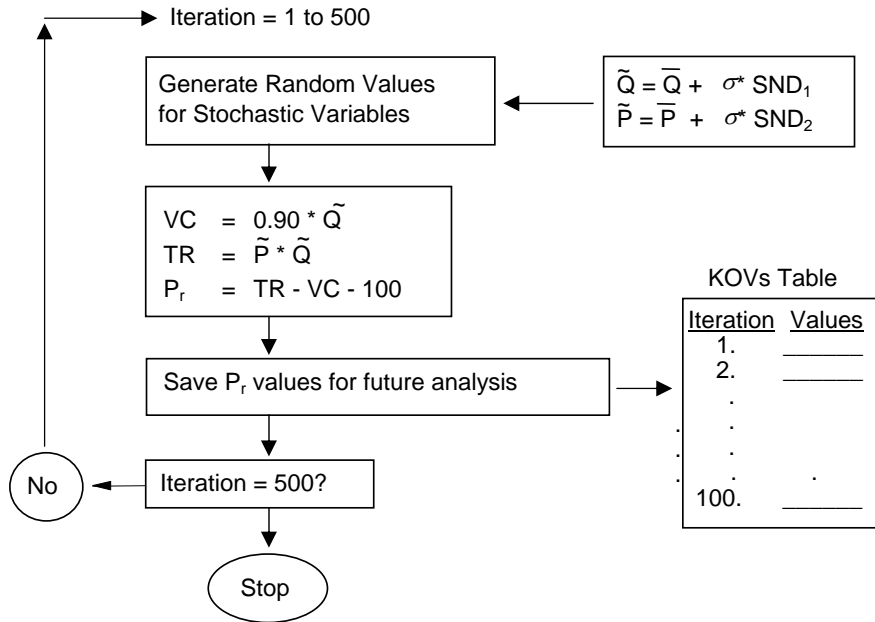


Figure 4.2. Flowchart for Simulating a Model for 500 Iterations.

Simple Statistics

- $\bar{X} = \bar{P}_r$
- $\sigma = \hat{\sigma}_{P_r}$
- min = MnP_r
- max = MxP_r

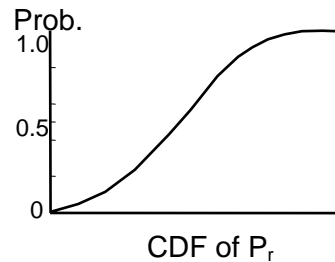
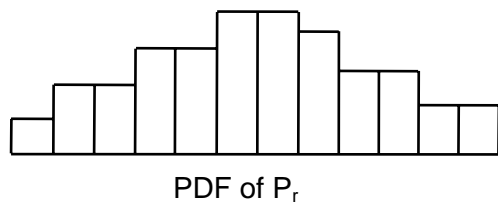


Figure 4.3. Different Ways to Summarize Simulation Results for a 500 Iteration Run of a Stochastic Model.

The results of simulating a model with Simetar are stored in a worksheet named SimData. The results consist of the summary statistics and the actual simulated values for each iteration. If the Simetar dialog box for Simulation was programmed to collect the output for 10 output variables over 500 iterations, the SimData worksheet would have 10 columns with the results for 10 KOVs and 500 rows of numbers for each variable. Each row of values in SimData constitute one iteration of simulated values for the output variables. The model depicted in Figure 4.2 was simulated for 500 iterations in the Excel program, Business Model With Risk Demo.XLS, to

demonstrate the concept of iterations. The output variables specified for Simetar were: Q, P, VC, TR, and PR. If one examines a particular iteration in SimData, you can use a calculator to verify how profit is calculated as a function of the random Q and P values for that particular iteration. (This process of checking an iteration with a calculator is part of model verification and is the responsibility of the modeler.)

Number of Iterations

The number of iterations to simulate a stochastic model is a researchable problem. The number of iterations is different for each model depending on the number of random variables, the degree of correlation among the random variables, the number and type of equations in the model, and sensitivity of the endogenous variables to the random variables. Generally, 100 to 250 iterations is sufficient to accurately estimate an empirical distribution for the key output variables in the model.

Once a model has been validated, the analyst needs to determine the number of iterations to use for production runs with the model. Simulate the model for a range of iteration numbers (say, 25, 50, 75, 100, 200, 500, 1,000, 5,000) and compare the summary statistics for the stochastic and key output variables. Compare the standard deviation for the key output variables across the alternative iterations. As the number of iterations increases, the standard deviation for the output variables changes until it reaches an equilibrium, as in Table 4.1. The iteration number where the standard deviation stabilizes should be the minimum number of iterations used for the model. The number of iterations for Business Model With Risk Demo.XLS was determined based on the results in the “IterNoTest” worksheet, using the steps described above.

Table 4.1. Comparison of the Mean and Standard Deviation for Profit as Iteration Number Increases for the Model Depicted by the Flowchart in Figure 4.2.

<u>Number of Iterations</u>	<u>Mean</u>	<u>Std. Deviation</u>
25	211.88	74.90
50	212.14	78.20
75	217.73	92.77
100	214.29	86.98
200	215.65	88.01
500	214.45	86.27
1,000	214.41	83.77
5,000	215.21	86.88

The final check for selecting the number of iterations is to examine the CDF’s for the output variables. The more iterations for a model the smoother the CDF graphs. Examine the CDFs for alternative numbers of iterations in Business Model With Risk Demo.XLS worksheets 25IterCDF to 1000IterCDF to see how iteration number affects the smoothness of the estimated CDFs.

Monte Carlo vs. Latin Hypercube Sampling

The number of iterations a model is simulated can be greatly reduced if the simulation package uses the Latin Hypercube procedure rather than the Monte Carlo procedure to sample the probability distributions. The Monte Carlo procedure randomly selects values from the probability distributions. As a result the procedure samples a greater percent of the random values from the area about the mean and under samples the tails. When Monte Carlo sampling is used it is recommended that a large number of iterations be used to minimize the effects of under sampling the tails of the probability distribution.

An alternative technique for sampling probability distributions is the Latin Hypercube sampling procedure (Inman, Davenport and Zeigler). This technique segments the distribution into N intervals and makes sure that at least one value is randomly selected from each interval. The number of intervals, N , is the number of iterations. By sampling from N intervals, the Latin Hypercube insures that all areas of the probability distribution are considered in the simulation. Therefore, when the Latin Hypercube sampling procedure is used, the number of iterations necessary to reproduce the parent distributions is less than when sampling with the Monte Carlo procedure.

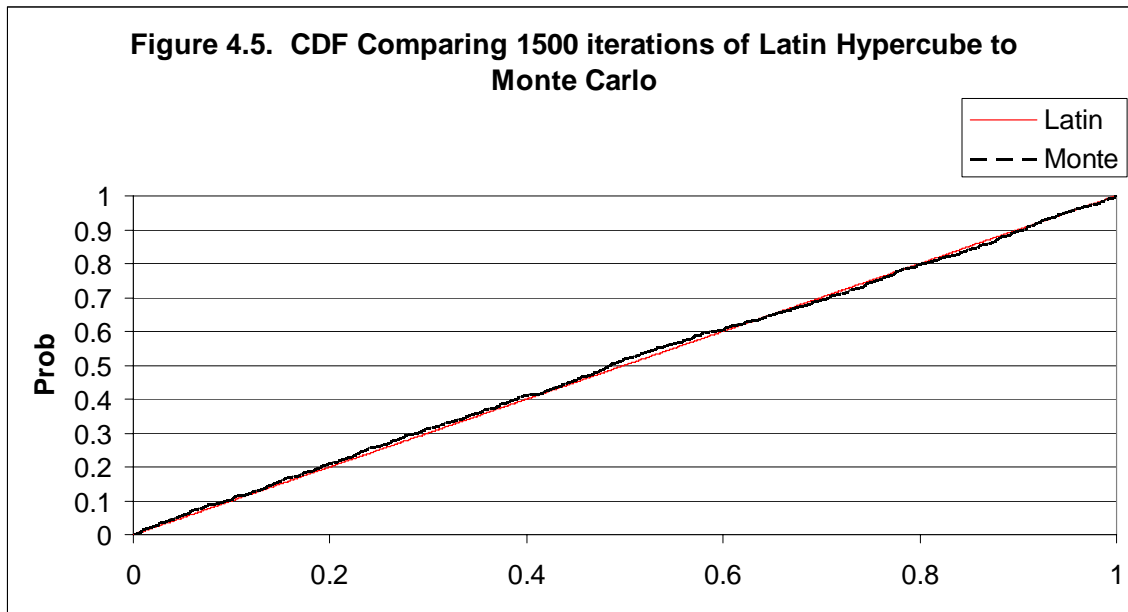
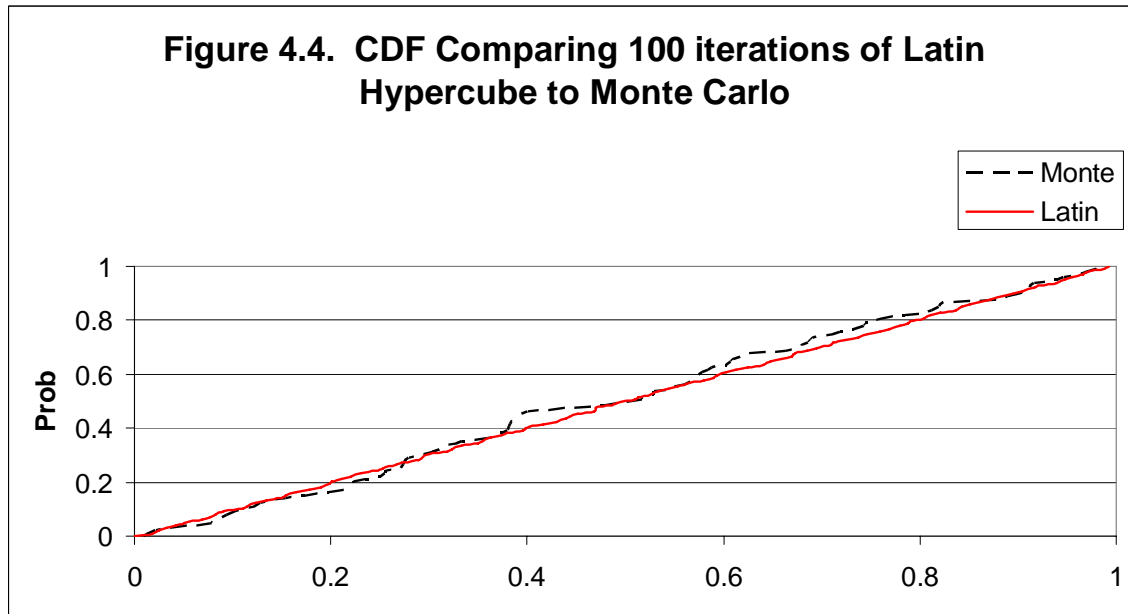
A simple example of the two sampling procedures is provided in Figures 4.4 and 4.5. These figures show the CDFs from simulating a uniform (0,1) distribution with the Monte Carlo and Latin Hypercube sampling procedures. A perfect sampling procedure will result in a straight line CDF between zero and one. With a sample size of 100 iterations the CDF for the Latin Hypercube is very close to a straight line (Figure 4.4). However, for the Monte Carlo procedure the CDF for 100 iterations is fairly uneven showing considerable bias in several segments of the distribution. At a sample size of 1,500 the Latin Hypercube CDF is a perfectly straight line while the Monte Carlo procedure still has some bias (Figure 4.5).

The Simetar simulation engine is programmed to only use the Latin Hypercube sampling procedure. The @Risk© package requires the analyst to specify which of the two sampling procedures to use. When using @Risk© be sure to select the Latin Hypercube sampling procedure.

Pseudo-Random Number Generator

To build stochastic simulation models one needs to know only a few of the details about how random numbers are actually generated. In simulation we use pseudo-random numbers. Pseudo-random numbers are generated by a process which guarantees that each "seed" will generate the "same sequence" of random numbers. This may not sound random to you, but it is.

A random number generator which generates a different sequence of random numbers each time it runs would be useless for conducting business analyses. Such a random number generator would develop a different set of results every time you run the model, even if the input values did not change. An even greater problem occurs when simulating two different management actions, because the differences in the results would be due to the assumptions and the particular sequence of random numbers. To avoid these problems we use pseudo random number generators so the only difference between runs is due to our assumptions about management's actions.



A pseudo random number generator uses a seed value to initiate its random number sequence. Each seed is associated with a unique sequence of random numbers, or realizations. The seed is a five digit, odd number, as 31517. I use this same seed for all simulation analyses and it works, meaning that the random numbers do not appear to be biased, do not degenerate to zero, and do not repeat themselves after a large number of iterations. (A note is that this is the 3,392 prime number.)

With Simetar you can generate pseudo-random numbers by specifying and using the same seed via the Simulation Settings dialog box. Any positive integer can be used as the random number seed. Changing the seed results in a different sequence of random numbers, as shown in Pseudo Random Number Generators Demo.XLS. In the Summary worksheet there are three columns of random numbers; each column is a sample of 500 pseudo random numbers distributed Uniform (0,1) using a different seed. The summary statistics indicate that all three samples are statistically equivalent, however, visual inspection reveals that the random numbers are different as the seed changes. It is recommended that you pick a seed and stay with it for all simulations.

Feldman and Valdez-Flores describe how the congruential method is used to generate pseudo random numbers. This method starts with a seed and then uses a fixed algorithm to calculate successive random numbers and seeds. Because the algorithm is a mathematical process the sequence of random numbers is always the same unless the seed is changed. The algorithm described by Feldman and Valdez-Flores is quoted directly (page 80) as follows:

“... let a and b be two fixed integers, and let L denote the largest possible (signed) integer that the computer can store. Let S be a random seed and let S_{next} be the next seed to be determined. The random number associated with the seed is:

$$R = S/L,$$

and the next seed is:

$$S_{\text{next}} = (aS + b) \bmod L.$$

For example, for a 16-bit computer, $L = 32767 = (2^{15}-1)$ and we might set $a = 1217$, set $b = 0$, and let the initial seed be $S_0 = 23$. For this situation, the random number sequence is generated by the following calculations:

$$S_1 = (1217 * 23) \bmod 32767 = 27991$$

$$R_1 = 27991/32767 = 0.85424 \text{ (first random number),}$$

$$S_2 = (1217 * 27991) \bmod 32767 \\ = 34065047 \bmod 32767 = 20134$$

$$R_2 = 20134/32767 = 0.61446 \text{ (second random number), ...}”$$

The random numbers generated by the congruential method are distributed uniform (are USDs) that can be converted to standard normal deviates (SNDs) using the inverse transform method. Because Simetar does all of this work for you during the simulation process all you have to know is that the sequence of random numbers will remain constant until the random number seed is changed.

Simulating Uncertainty

Risk and uncertainty are out of the decision makers control and most often they are the forces that affect the success of a business decision. Risk is the portion that can easily be modeled using stochastic simulation. Risky variables have probability distributions that define the nature of their risk. For example, sales per period may be normally distributed with a mean of 10,000 and a standard deviation of 300.

Uncertainty is risk which can not be defined by a probability distribution. An uncertain variable is one which does not have a known distribution. One way to think about uncertainty is that the mean and standard deviation for the variable's probability distribution are risky. Using this construct, the risky probability distribution for an uncertain variable X could be stated in terms of its uncertain mean ($\bar{X} \sim N(\bar{Y}, \sigma_y)$) and uncertain standard deviation ($\hat{\sigma}_x \sim N(\bar{\sigma}_z, \sigma_z)$) or $\tilde{X} \sim N(N(\bar{Y}, \sigma_y), N(\bar{\sigma}_z, \sigma_z))$. The resulting probability distribution for the example variable X is risky in terms of its mean and variance, but what if one did not even know the type of distribution X followed, then X would be truly uncertain.

Because uncertain variables do not have known distributions and parameters they can not be simulated directly. Uncertainty can be incorporated into simulation models two ways. The first is to use an example from catastrophe theory and assume the worst outcome happens at random with a probability of P . For example, a plant manager may know that the main power supply can fail but has no idea when it will occur. This type of uncertain event can be simulated using a Bernoulli (P) distribution. Each period the power failure variable is simulated, if it equals 1 then the uncertain event occurred.

Another way to try to stress test a business for uncertainty is to test the probability distributions for risky variables by making their means stochastic using a sensitivity analysis. A range of means for each distribution can be simulated to determine which is most critical to the business decision. This approach can be extended to the variability parameter for each risky variable.

References

@Risk©: Advanced Risk Analysis for Spreadsheets. Newfield, NY: Palisade Corporation, 2000.

Feldman, R. and C. Valdez-Flores. Applied Probability and Stochastic Process. Boston: PWS Publishing Company, 1996.

Inman, R.L., J.M. Davenport, and D.K. Zeigler. “Latin Hypercube Sampling (a Program User’s Guide).” Technical Report SAND 79-1473, Sandia Laboratories, Albuquerque, 1980.

Jones, G.T. Simulation and Business Decisions. Middlesex, England: Penguin Books Ltd., 1972.

Law, A.M. and W.D. Kelton. Simulation Modeling and Analysis. Third Edition. New York: McGraw-Hill, Inc., 2000.

Reutlinger, S. Techniques for Project Appraisal Under Uncertainty. Baltimore: The Johns Hopkins Press, 1970.

Vose, D. Risk Analysis: A Quantitative Guide. New York: John Wiley & Sons, Ltd. Second Edition, 2000.

Winston, W.L. Simulating Modeling Using @Risk©. New York: Duxbury Press 1996.

Chapter 5 Distributions Frequently Used for Simulation

Probability distributions are classified as being continuous or discrete and closed or open form. Continuous distributions are smooth functions (lines) that do not have breaks or jumps from their minimums to their maximums. These functions behave nicely, can generally be integrated and conform to mathematical functions that define their properties. Discrete probability distributions are discontinuous in at least one point and these distributions have to be integrated with summation notation. Discrete distributions are often referred to as being non-parametric.

Open and closed form probability distributions indicate whether a probability distribution has a finite minimum and maximum. Open form distributions have no finite end points. A normal distribution has no end points other than plus and minus infinity. A closed form distribution has specified end points, such as an empirical distribution.

Probability distributions are also classified as univariate or multivariate. Univariate distributions refer to one variable while multivariate distributions have more than one random variable. Several random variables that are independent (uncorrelated) would be included in a simulation model as univariate distributions. Other random variables that are correlated to each other would be simulated as a multivariate distribution. The job of estimating parameters for univariate distributions is the subject of Chapter 6. Parameter estimation and simulation of multivariate distributions is described in Chapter 7.

The mathematical and statistical properties for probability distributions frequently used for simulation are reviewed in this chapter. Each distribution is presented in terms of its parameters, density function (pdf), cumulative distribution (cdf), key properties, and Simetar command for simulation. (Recall that the cdf of a random variable is the integral of the pdf for that variable.) The Probability Distribution Demo.XLS and Simulate Alternative Distributions Demo.XLS workbooks demonstrate how to simulate most of the distributions. Distributions not demonstrated in these worksheets can be found in Simulate All Probability Distributions Demo.XLS in Chapter 16. Parameters, in the context of simulation, are the values that define the distribution.

Continuous Distributions

– Uniform Distribution $U(\text{min}, \text{max})$

Each equal length interval of X over the minimum to maximum range has an equal probability of being observed. The parameters are the minimum and maximum for variable X , so it is a closed form distribution. For example, a uniform random variable X distributed over the range of 10 to 20 is denoted as $X \sim U(10,20)$. Simulate a uniform distribution in Simetar using the command =UNIFORM (min, max).

A special case of the uniform distribution is $X \sim U(0,1)$ which produces a uniform standard deviate (USD). A USD on the 0-1 scale is used to simulate random numbers for all probability distributions, via the Inverse Transform method of generating random variables and is covered at the end of this Chapter.

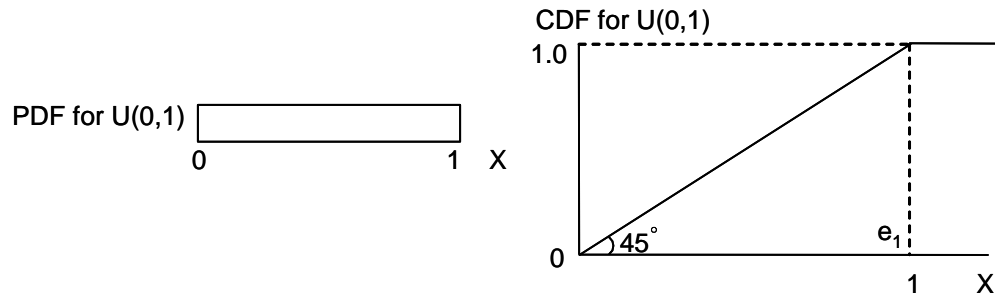


Figure 5.1. PDF and CDF for a Uniform Distribution.

– Normal Distribution $N(\text{mean}, \text{std dev})$

The normal distribution produces a bell shaped pdf with set probabilities. The normal function reaches to plus and minus infinity so it is an open distribution. This function is widely used and the integral of the pdf is found in the Z table in most statistics textbooks. Most analysts have memorized several probabilities about the normal distribution, such as:

- 66% of observations within \pm one standard deviation of \bar{X}
- 95% of observations within \pm two standard deviations of \bar{X}
- 50% of observations $> \bar{X}$ or $< \bar{X}$

Parameters for the normal distribution are the mean and standard deviation (\bar{X} and σ). Simulate a normal distribution in Simetar with =NORM (mean, standard deviation).

A special case of the normal distribution is $X \sim N(0,1)$ which results in a standard normal deviate (SND). Variations on a normal distribution is a truncated normal which cuts off one or both tails. For example, a normal distribution with a finite minimum is $X \sim \text{TNORM}(\text{mean}, \text{std dev}, \text{min})$ where 8 is a minimum. If X is a normally distributed random variable with a finite maximum it could be described as $X \sim \text{TNORM}(\text{mean}, \text{std dev}, \text{max})$.

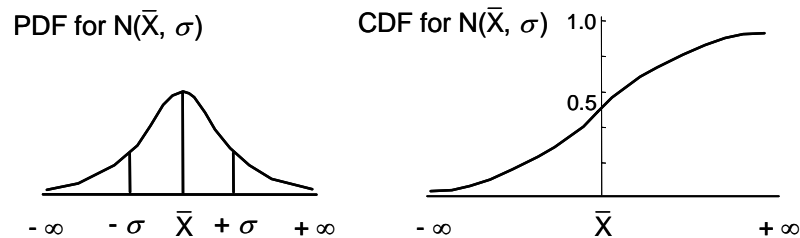


Figure 5.2. PDF and CDF for a Normal Distribution.

– **Empirical Distribution $E(\text{sorted } X_i \text{ values; } P(X_i\text{'s}))$**

Non-parametric empirical distributions are generally used when a random variable has too few observations to estimate the parameters for a parametric distribution. The distribution has a finite minimum and maximum based on observed values so it is closed form. The shape of the distribution is defined by the data. The function's input data are discrete, however, interpolation between segments is done during simulation to make the cdf continuous.

Parameters for the empirical distribution are the sorted values of X (or S_i) and the cumulative probabilities for the sorted values ($P(S_i)$). Simulate the empirical distribution in Simetar with `=EMPIRICAL (Si, P(Si))`.

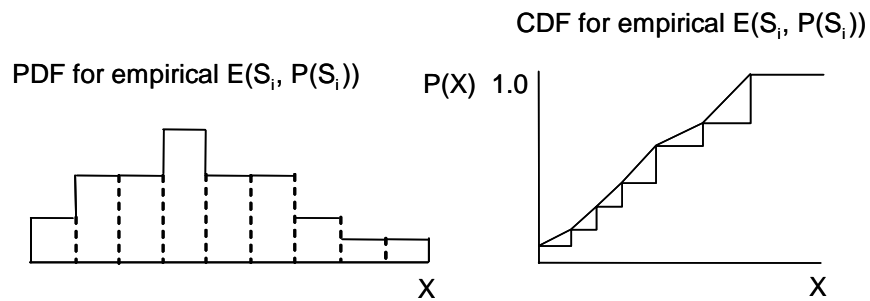


Figure 5.3. PDF and CDF for an Empirical Distribution.

– **GRKS Distribution $GRKS(\text{min, mid point, max})$**

GRKS distribution was developed by Gray, Richardson, Klose, and Schumann to simulate subjective probability distributions based on minimal input data. Business managers can provide estimates of three points on a distribution of possible outcomes (min, mid point, max), but they often admit things could be worse or better than they expect. The distribution is a closed form distribution.

Parameters for the GRKS are the minimum, mid point, and maximum. Simulate the GRKS distribution in Simetar with `=GRKS(min, mid point, max)`.

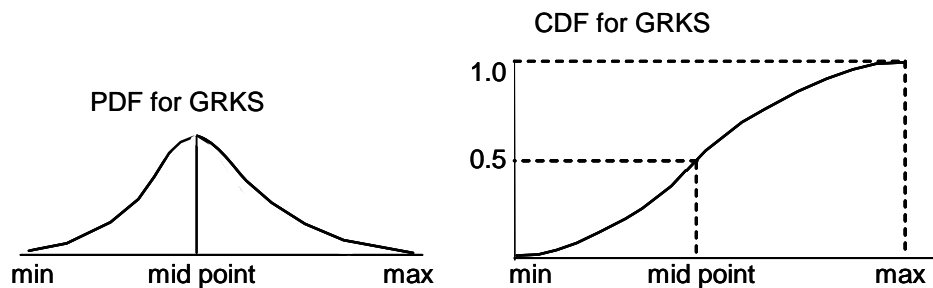


Figure 5.4. PDF and CDF for a GRKS Distribution.

Three parameters are used to estimate the “rest” of the parameters for the GRKS distribution, based on the assumed properties or shape for the GRKS. The properties of the GRKS are:

- 50 percent of the simulated observations are less than the mid point.
- About 95 percent of the simulated observations are between the minimum and the maximum.
- 2.2 percent of the simulated observations are less than the minimum and 2.2 percent are greater than the maximum.
- There are four equal distance intervals between the mid point and the minimums.
- There are four equal distance intervals between the mid point and the maximum.
- There are two intervals below the minimum and two above the maximum and they are the same distance as the other intervals on their respective side of the mean.

Given a GRKS(20, 50, 60), the full parameters are demonstrated in Table 5.1. The $P(X_i)$ values are based on probabilities in a Z table for a standard normal distribution with the minimum and maximum being -2 and $+2$, respectively, standard deviations from the mid point. The X_i values for the distribution are defined by two formulas:

- Width of intervals in the lower half of the distribution is:
(mid point – minimum) / 4 or $7.5 = (50 - 20) / 4$
- Width of intervals for the upper half of the distribution is:
(maximum – mid point) / 4 or $2.5 = (60 - 50) / 4$

Table 5.1. Summary of Parameters for a GRKS (20, 50, 60) Distribution.

<u>Interval</u>	<u>X_i</u>	<u>$P(X_i)$</u>
1 infinity	5.0	0.0000
2	12.5	.0062
3 minimum	20.0	.0227
4	27.5	.0668
5	35.0	.1587
6	42.5	.3085
7 mid point	50	.5000
8	52.5	.6915
9	55.0	.8413
10	57.5	.9332
11 maximum	60.0	.9773
12	62.5	.9938
13 + infinity	65.0	1.0000

– **Exponential Exp(β)**

The exponential distribution is used to simulate random values for growth or decay functions. The function is simulated using the functional form $1 - e^{-\frac{x}{\beta}}$. The distribution generates a random value with mean β and variance β^2 .

Parameters for the exponential distribution are the mean (β). Simulate the exponential distribution in Simetar with =EXPONINV(beta).

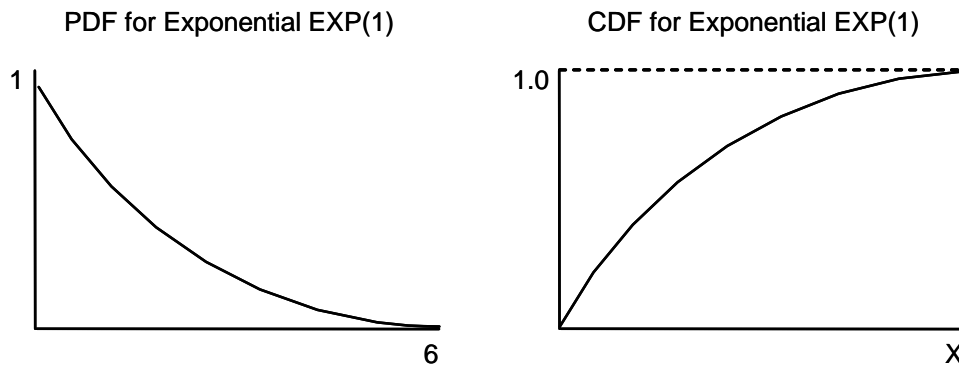


Figure 5.5. PDF and CDF for an Exponential Distribution.

– **Other Continuous Distributions**

There are many other continuous probability distributions that can be used for simulating random variables, such as: beta, gamma, poisson, log-log, log normal, and weibull. An Excel demonstration program for viewing and experimentation on these and other key probability distributions is provided in View Distributions Demo.XLS.

Figure 5.6 presents an example of how the View Distributions Demo.XLS program works. The user enters 10 observations for data series (probability distribution) to be analyzed in B6:B15. Next the user selects the distributions to be tested/viewed in cells D7:D9 via drop down menus in each of these cells. The user can choose among 12 distributions in each cell. Simetar functions calculate the parameters for the selected distributions, simulates the distribution, and shows the distribution as a CDF in the chart. The lines in the chart are color coded to match the colors for the three distributions the user selected. The =CDFDEV() value is a loss function to indicate how closely the assumed distribution fits the user's data; a zero is a perfect fit. The 12 distributions that can be viewed in View Distributions Demo.XLS are: beta, double exponential, exponential, gamma, logistic, log-log, log-logistic, lognormal, normal, pareto, uniform, and weibull.

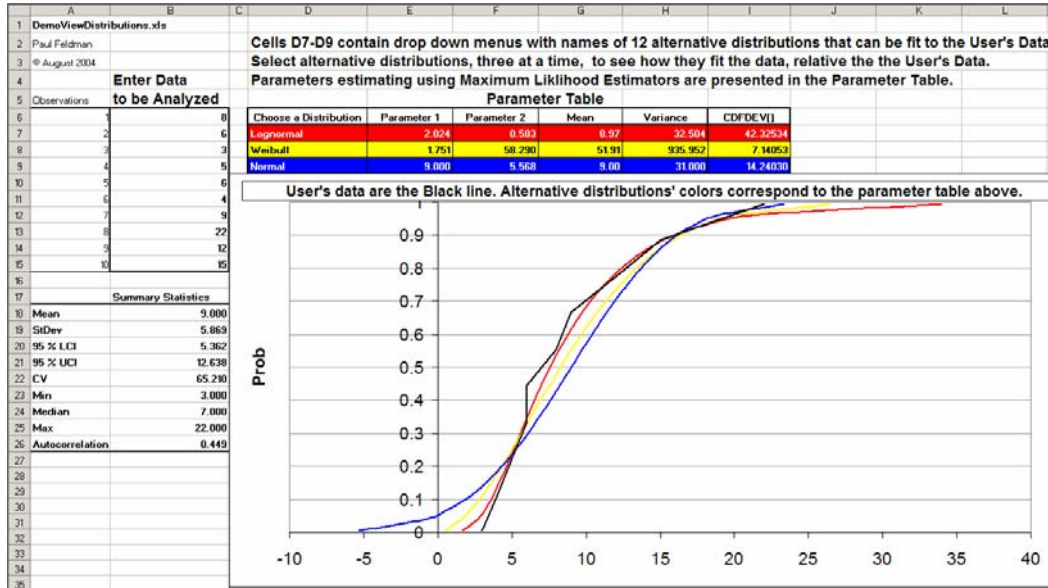


Figure 5.6. Sample Output for View Distributions Demo.XLS.

The Simetar functions used to simulate more than 40 probability distributions are summarized in Section 2.0 of Chapter 16.

Discrete Distributions

– Bernoulli B(p)

A Bernoulli distribution is used to simulate variables with two values, either $X = 0$ or $X = 1$. The probability of X equaling 1 (or true) is the probability P and the probability of X equaling zero (or false) is $1-P$. This probability distribution makes a good “on” or “off” switch for a conditional random variable. For example: the variable could be rain or dry, dead cow or a live one, prices increase or decrease, machine fails or works, etc., each with the probability of P for a particular outcome.

Parameter is simply the probability of $X=1$ or P . Simulate a Bernoulli distribution in Simetar with =BERNOULLI (P).

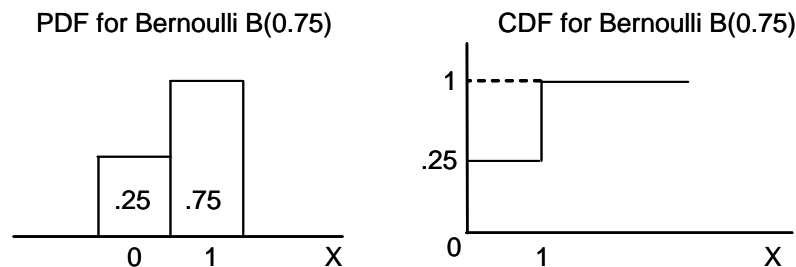


Figure 5.7. PDF and CDF for a Bernoulli Distribution.

– **Discrete Empirical DE (sorted X_i values)**

A discrete empirical random variable X can take on fixed values with an equal probability, or $X \sim DE(3.1, 4.4, 6.6, 7.8)$ for a random variable X which has an equal probability of being 3.1, 4.4, 6.6, or 7.8. The X value is not interpolated so only these X_i values can be observed.

Parameters are the sorted X values or S_i from minimum to maximum. Simulate a discrete empirical distribution in Simetar using =DEMPIRICAL(S_i).

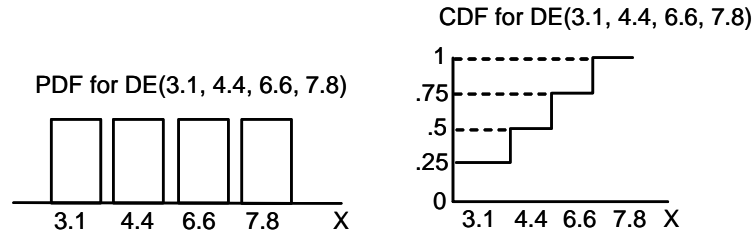


Figure 5.8. PDF and CDF for a Discrete Empirical Distribution.

Conditional Probability Distributions

The distributions described and demonstrated thus far in Chapter 5 are unconditional distributions. In other words, the distributions are not directly dependent upon another distribution. In simulation modeling we frequently encounter cases that must be modeled using conditional distributions.

Conditional distributions occur where a random variable can take on different distributions rather than just different values in one distribution. In theoretical terms, consider the case of a random variable X which has two possible distributions depending upon another random variable Y or:

$$X \sim \begin{cases} E(1, 2, 3, 4, 5) & \text{if } Y = 1 \\ \text{or} \\ E(10, 12, 13, 14, 15) & \text{if } Y = 0 \end{cases}$$

where $Y = 1$ with probability of P and $Y = 0$ with probability $1-P$. Both X and Y are random variables but X is conditional on Y .

Conditional distributions can occur when simulating rainfall, or mechanical failures, or even market demand for new products. In the case of rainfall, think of the Y (or switch) distribution being a true or false indicator of whether it rained or not over a particular time period. In this case, X is the random quantity of rainfall, if it rained, or:

$$\text{Rainfall} = \begin{cases} \text{TNORM}(3.0, 1.0, 0, 4.5) & \text{if } Y = 1 \\ \text{or} \\ 0.0 & \text{if } Y = 0 \end{cases}$$

where $Y = 1$ with probability of P and $Y = 0$ with probability $1-P$.

In the case of simulating mechanical failures the Y switch distribution determines whether there was a failure or not, over a given period. The X distribution could be either the time loss or repair cost caused by the failure. For example, the repair cost to a generator could be \$100, \$3,000, or \$5,000, with equal probability, if a mechanical failure occurs. The probability of a mechanical failure is 10 percent. This problem can be restated using the discrete uniform (DU) distribution as:

$$\text{Cost} = \begin{cases} \text{DU}(100, 3000, 5000) & \text{if } Y = 1 \\ 0 & \text{if } Y = 0 \end{cases}$$

where $Y = 1$ with probability of 10 percent and $Y = 0$ with a 90 percent probability.

To simulate the probability of a mechanical failure and its related cost in Simetar, first use the Bernoulli distribution to determine if there was a breakdown and then simulate the cost of the repair in a second distribution. Let cell A1 have the following formula:

=BERNOULLI(0.10)

and then use the resulting value from A1 in the A2 formula:

=RANDSORT(100, 3000, 5000) * A1

The result in A2 is a random cost of repair that equals zero 90 percent of the time and either \$100, \$3,000, or \$5,000 about 33.33 percent of the time.

The Conditional Probability Distribution Demo.XLS includes two examples of conditional probability distributions. The first example is the mechanical failure for a generator described above. The second example is a case where a salesman is paid a base monthly salary plus a bonus. The bonus is equal to 8.5 percent of sales receipts after the first 1,000 units sold each month. Each unit sells for \$10. Historical data for sales in the region are provided and the task is to estimate the probability distribution for the salesman's salary. The problem can be stated as follows:

$$\text{Salary} = \begin{cases} \$5,000 + (\text{Sales}-1,000) * \$10 * 0.085 & \text{if Sales} > 1,000 \\ \text{or} \\ \$5,000 & \text{if Sales} \leq 1,000 \end{cases}$$

The first step is to estimate the probability that monthly sales exceed 1,000 units. This is calculated by using the =EDF() function. This yields a probability of 62.84 percent. The second step is to estimate the parameters for a distribution of sales. In this case the distribution is made up of sales data for only months when sales exceeded 1,000 units (see columns E and H in the workbook). Because the test for normality failed to reject normality, the distribution was simulated as a truncated normal. (The distribution was truncated at 1,000 to be consistent with the conditional nature of the distribution.)

The salesman's salary problem is simulated in two steps. First the switch variable is simulated to see if sales exceed the threshold value (E65). Second the bonus for sales over the threshold is calculated. The result is that the expected salary is \$5,085 per month. See SimData1 for the simulated results.

When are these Distributions Used?

- **Uniform distribution** is used if each observation of the random variable between the minimum and maximum has an equal chance of occurrence or you have no idea what type of distribution to use.
- **Normal distribution** is used if the random variable is the error term for a regression equation or the data have been tested statistically and you cannot reject the null hypothesis of a normal distribution. Use the normal distribution if you have lots of observations and have tested the data to insure the variable is normally distributed. (See Chapter 16 for Simetar's tests for determining if a data series is distributed normal.) After simulating the random variable test that the simulated minimums and maximums do not produce irrational results, such as negative prices, yields or interest rates or unrealistically large values.
- **Empirical distribution** is used if the random variable can take on any value within a finite range and there are too few observations to estimate the parameters for the true distribution. Usually 20 or more observations are required to prove conclusively that a distribution is normally distributed or to estimate the parameters of a distribution with a high degree of certainty. This is not usually the case in business as it is hard to get 10 observations under the same economic policy, management regime, farm program, or trade policy. Simulating crop yields as an empirical distribution when you have only 10 historical values, is a good example. We know yield can be any positive value. We do not have enough observations to test for normality. We know the 10 yields were observed with a probability of 1/10, or one each year, therefore the distribution looks like Figure 5.9. The shape of the distribution is specified by the data which leads to the name – non-parametric empirical distribution.

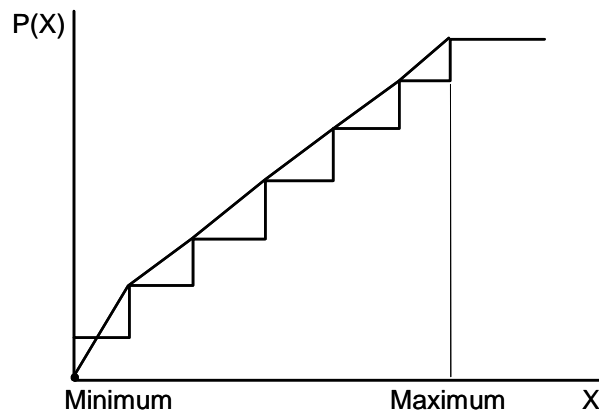


Figure 5.9. CDF of an Empirical Distribution.

- **Exponential distribution** is useful when needing to simulate a variable which is subject to decay or growth. Decaying values such as the benefits of advertising or the population of grasses on a range site are examples of such variables. Growth functions such as populations of insects or bacteria growth also can be simulated as exponential distributions.

$$\tilde{X}_{\text{decay}} \sim E_{\text{XP}} (\text{Mean})$$

or

$$\tilde{X}_{\text{growth}} \sim \frac{1}{E_{\text{XP}} (\text{Mean})}$$

- **Bernoulli distribution** is a perfect on/off switch which activates another random variable or a decision. The Bernoulli distribution returns a 1 P percent of the time and 0 (1-P) percent of the time. It's therefore often used to simulate conditional distributions.
- **Discrete Empirical distribution** is useful if the stochastic variable has a finite number of observations, the variable is able to take on only certain values, and there is no indication of alternative values the variable can take on, e.g., a factory can hire labor in fixed quantities of 2, 4, 6 and 8 hours for each employee, or the capacity for the next cattle truck in line at a feedlot could be 50, 75, 100, or 125 head of steers.

A variation on this problem is the one of randomly sorting a finite number of values, such as drawing the order of contestants. Each contestant represents a discrete random variable that can be picked in any order. This is the distribution used to simulate playing dice in a simulation model. See Games Of Chance Demo.XLS and Simulate All Probability Distributions Demo.XLS for an example of how this distribution is used to flip a coin, roll dice, play poker, play Bingo, play the lottery, and spin the slot machine.

- **GRKS distribution** is used when dealing with very limited information about the random variable. The decision maker may only be able to provide values for the mid point, minimum, and maximum. These three values define a subjective distribution that can be used until something better is developed. It is recommended that if the GRKS distribution is used, the decision maker/expert be consulted regarding its simulated values during the validation process.
- **Triangular distribution** is often used in simulation when modelers do not know better. It is easy to use because it is fully defined by the minimum, mode and maximum. The problem with the triangle distribution comes from the probability in the tails. If farmers are asked "what was the minimum or maximum yield they experienced over 10 years," they give values that were actually observed 1 in 10 years. That means the simulation model needs to simulate the minimum 10% of the time and the same for the maximum. The triangle pdf will never simulate the minimum or the maximum with more than a 1 percent probability. An alternative to the triangle distribution is the GRKS distribution when dealing with subjective data about the random variable.

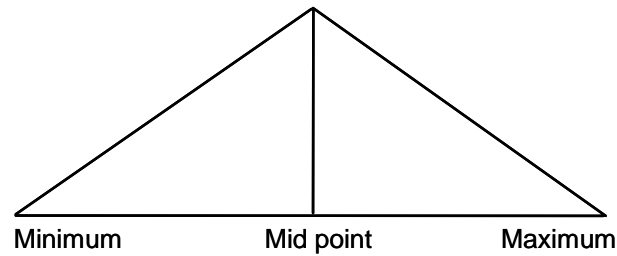


Figure 5.10. PDF for a Triangle Distribution.

- **Conditional distribution** is used when the stochastic variable is a function of another variable, such as weather, market conditions, or outcomes for certain variables in the system. This distribution allows the model to join together different distributions to simulate a single variable, thus more accurately representing the real situation. A Bernoulli distribution is generally used to activate a conditional distribution. For example, a conditional pdf for rainfall in a given region could be simulated as: if it rains ($0.25 = P(\text{rain})$) then the amount of rain is distributed normal with a mean of one inch and a standard deviation of 0.20 inches. This type of conditioning is referred to as a “Conditional Probability Distribution” and is dependent upon a Bernoulli distribution to determine if the event occurs.

Inverse Transform Method for Simulating Random Variables

The Inverse Transform for simulating random variables is the procedure for linearly transforming a random uniform standard deviate (USD) into a random value for another distribution. Most all random number generators use a USD and transform it to the desired distribution using this procedure. The purpose of this section is to describe how the Inverse Transform works for uniform, normal, empirical, Bernoulli, discrete empirical, and triangle distributions.

– Inverse Transform for the Uniform Distribution

Inverse Transform Demo.XLS contains an example of how the Inverse Transform procedure is applied to simulate a random value distributed $U(10,15)$. The steps to using the Inverse Transform procedure for the uniform distribution are:

- In cell A1 generate a USD using
= UNIFORM ()
- In cell A2 enter the Inverse Transform formula for a uniform distribution
= Min + (Max – Min) * A1
For a random variable $X \sim U(10,15)$ the formula in A2 is
= 10 + (15 – 10) * A1
or
= 10 + (15 – 10) * UNIFORM ()

Note: The former formula is used for advanced simulation techniques when correlating random variables.

The Inverse Transform procedure in the formulas above is demonstrated in Figure 5.11.

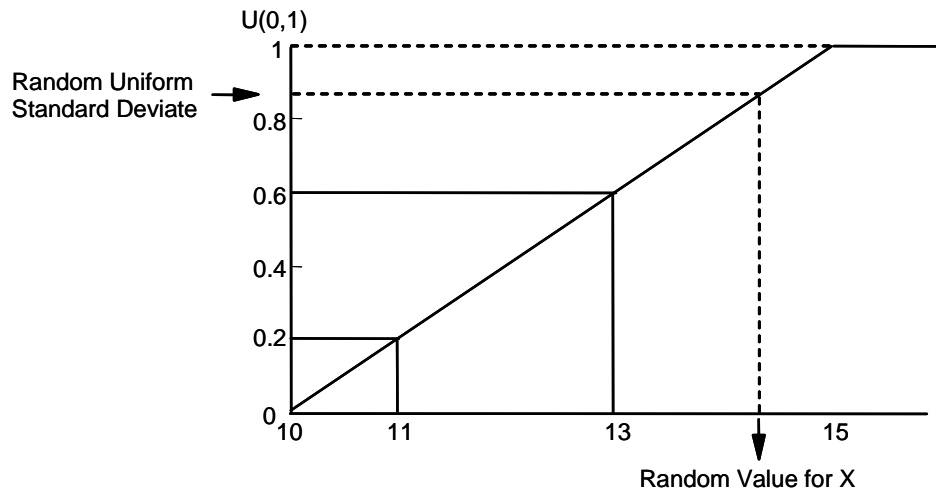


Figure 5.11. Inverse Transform for a Uniform Distribution.

Use a calculator and check the Inverse Transform formula as follows:

if the USD = 0.6
then $X = 13$, given that $X = 10 + (5 * 0.6)$

if the USD = 0.2
then $X = 11$, given that $X = 10 + (5 * 0.2)$

– Inverse Transform for the Normal Distribution

A uniform standard deviate (USD) can be used to simulate (or generate) a standard normal deviate (SND) and a normal distribution using the Inverse Transform. The SND transformation from a USD is demonstrated graphically in Figure 5.12. A USD is generated at random between 0 and 1 with =UNIFORM() and applied to the inverse normal distribution to solve for the number of standard deviations from the zero mean. For example, if a USD of 0.8 is drawn the corresponding SND is 0.84 and if the random USD is 0.25 the random SND equals -0.67 .

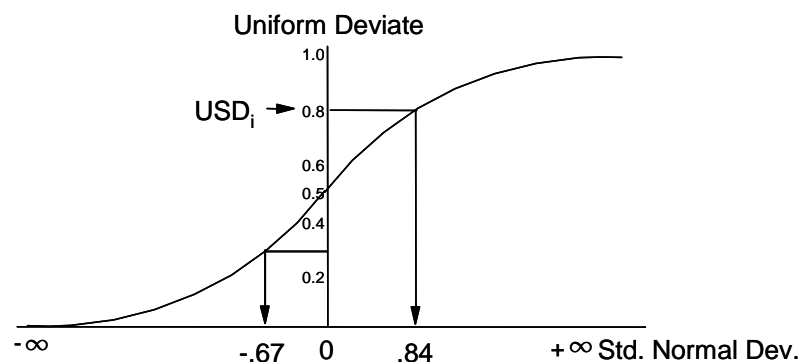


Figure 5.12. Inverse Transform for Generating a SND from a USD.

In the Inverse Transform Demo.XLS spreadsheet a table is provided so the user can type in USD's between 0 and 1 in B63-B73 to see their associated unique SNDs. The sample Z values here were generated using the example table:

USD of 0.00001	yields SND of	-4.26
USD of 0.25	yields SND of	-0.67
USD of 0.50	yields SND of	0.00
USD of 0.80	yields SND of	0.84
USD of 0.96	yields SND of	1.75
USD of 0.9999	yields SND of	3.719

The Inverse Transformation formula for a SND is very complex so Excel has provided a function to perform this calculation. The Excel equation to transform a USD to a SND is:

= NORMSINV(USD)

The Inverse Transform procedure is used to simulate the normal distribution in Simetar. The NORM command has an optional parameter which allows the analyst to supply the USD used to simulate the normal distribution. This can be demonstrated three ways for simulating a normal distribution for $X \sim N(10, 3)$:

- Generate a USD in a cell and then use that value in the =NORM() function
 - In cell A4 enter
=UNIFORM()
 - In cell A5 enter
=NORM(10,3,A4)

- Generate a USD in one cell, generate its associated SND in another cell, and calculate the random X using the Inverse Transform formula for a normal distribution
 - In cell A7 enter
=UNIFORM()
 - In cell A8 enter
=NORMSDIST(A7)
 - In cell A9 enter
=10+(3 * A8)
 - In this case the NORMSDIST function converts the USD to a SND
 - As a check use the equation below in a cell. The answer will be the same as the result in A9.
=NORM(10, 3, A7)

- Generate the random value in one step
 - In a cell enter =NORM(10, 3, UNIFORM())
or
 - In a cell enter =NORM(10, 3)

The first two procedures are used for more advanced simulation techniques which involve correlation of random variables.

– Inverse Transformation for the Empirical Distribution

The Inverse Transform procedure is used to simulate the empirical distribution. An empirical distribution is defined as $X \sim E(S_i, F(S_i))$ which is the definition a CDF for a random variable, i.e., S_i represents the horizontal axis and $F(S_i)$ represents the vertical axis. With this in mind and the fact that $F(S_i)$ is a series of probabilities from zero to one, it is easy to see how a USD can be used to simulate an S_i with the Inverse Transform procedure. There are two methods for simulating empirical distributions with Simetar.

- Direct simulation procedure calls for a sorted array of random values and their associated probabilities, referred to, respectively, as the S_i and $F(S_i)$ where S_i refers to the sorted stochastic values and $F(S_i)$ is the cumulative probability function for the S_i values. The Simetar function is

=EMPIRICAL (range of S_i , range of $F(S)$)

- The more detailed formula which is used for advanced simulation applications allows the user to specify the USD to be used for the simulation. This form of the function is used when correlating random variables.

=EMPIRICAL (range of S_i , range of $F(S)$, USD)

The steps that are programmed into the EMPIRICAL function are listed below to demonstrate how the Inverse Transform procedure works for the empirical distribution.

- Generate a USD using =UNIFORM().
- Match the USD into its interval on the probability $F(X_i)$ scale between F_U and F_L in Figure 5.13 and the table below.
- Match up the corresponding X interval, between X_L and X_U (where X_L is the X value for the lower part of the interval and X_U is the X value for the upper part of the interval)
- Interpolate to calculate \tilde{X} given the stochastic USD using the F_U , F_L , X_L and X_U
- Repeat the process for additional iterations or USD values.

The formula for interpolating a USD to \tilde{X} is best demonstrated using a distribution, such as:

	Obs	X	F(X)
Pmin	0	9.999	0.0
	1	10	.10
	2	13	.30
	3	18	.50
	4	22	.70
	5	30	.90
Pmax	6	30.001	1.00

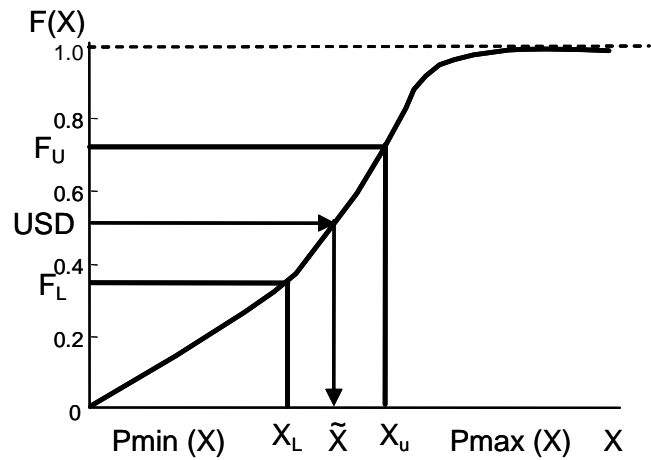


Figure 5.13. CDF of an Empirical Distribution.

To simulate five random numbers from the example distribution, draw a sample of five USD's (values in bold below) and interpolate as follows:

Iter	USD	\tilde{X}_i	The interpolation formula for the empirical cdf is: $\tilde{X}_i = X_L + (X_U - X_L) * ((\mathbf{USD} - F_L) / (F_U - F_L))$
1	0.10	10.0	$10 = 10 + (13 - 10) * [(\mathbf{.10} - .10) / (0.3 - 0.1)]$
2	0.80	26.0	$26 = 22 + (30 - 22) * [(\mathbf{.8} - .7) / (.9 - .7)]$
3	0.55	19.0	$19 = 18 + (22 - 18) * [(\mathbf{.55} - .5) / (.7 - .5)]$
4	0.11	10.15	$10.15 = 10 + (13 - 10) * [(\mathbf{.11} - .10) / (.3 - .1)]$
5	0.25	12.25	$12.25 = 10 + (13 - 10) * [(\mathbf{.25} - .10) / (.3 - .1)]$

The formula for simulating an empirical distribution via the Inverse Transform method is programmed in Empirical Distribution Demo.XLS. In the example, a simple 5 interval pdf is simulated for selected USDs to demonstrate the formula. Change the sample USD's in rows 24-28 to see how the random X_i values change. A second part of Empirical Distribution Demo.XLS shows how to use the inverse transform formula with a table lookup function to simulate an empirical distribution. Step 5 in the demo program demonstrates how to simulate an empirical distribution using Simetar. The actual Simetar command in B84 to simulate the distribution in Step 5 is displayed in cell E81.

– **Inverse Transform for the Bernoulli Distribution**

The Bernoulli distribution can be simulated as a special case of the uniform distribution. A USD is generated in one cell and an =IF statement uses the USD in another cell to determine if the simulated outcome is true (1) or false (0). For example, if X is distributed Bernoulli with probability 0.75 it is simulated as

- In cell A1 enter
= UNIFORM()
- In cell A2 enter
= IF (A1 <= 0.25, 0, 1)

The value in A2 will be one, 75 percent of the time and zero, 25 percent of the time. The distribution can be simulated more directly using the function =BERNOULLI(0.75).

A graphical depiction of how the Bernoulli (0.75) is simulated using the Inverse Transform is shown in Figure 5.14. Any USD between 0.25 and 1.0 results in a value of 1 while USD values less than or equal to 0.25 are assigned the value of zero. This of course results in a value of one, 75 percent of the time.

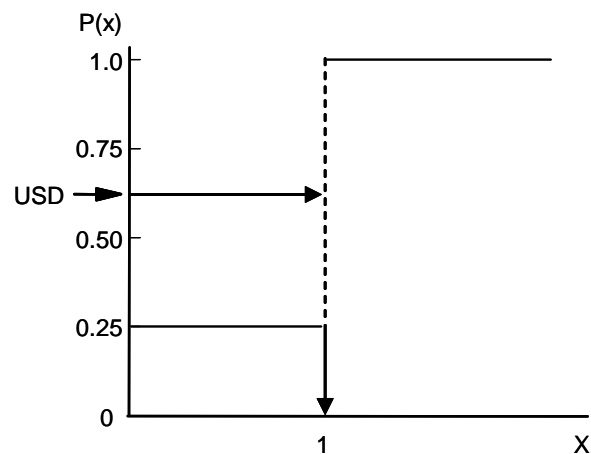


Figure 5.14 Inverse Transform for a Bernoulli Distribution B(0.75).

– Inverse Transform for the Discrete Empirical Distribution

The discrete empirical distribution assumes the random values for X can take on specific values x_1, x_2, x_3, x_4 , etc. As a result, interpolation as indicated in Figure 5.13 for the empirical distribution is not applicable for this distribution. Random values for a distribution $X \sim DE(3, 4, 8, 10)$ can be simulated using the Inverse Transform as follows:

- In cell A1 enter
= UNIFORM()
- In cell A2 enter
= IF (A1 <= 0.25, 3, 0)
- In cell A3 enter
= IF (and (A1 > 0.25, A1 <= 0.50), 4, 0)
- In cell A4 enter
= IF (and (A1 > 0.5, A1 <= 0.75), 8, 0)
- In cell A5 enter
= IF (A1 > 0.75, 10, 0)
- In cell A6 enter
= SUM (A2:A5)

The value in cell A6 is distributed DE(3, 4, 8, 10) with each value having a 25 percent chance of being observed. The DE function simulated in the example via the Inverse Transform as depicted in Figure 5.15. A USD is compared to intervals on the probability axis and assigned values for X. When the USD is between 0.25 and 0.50, as indicated in the figure the random value for X equals 4. If the USD is greater than 0.75 X is assigned the value of 10. If the X variable has more possible values, say 10, the probability (0 to 1) axis is divided into 10 equal intervals and random X's are assigned accordingly.

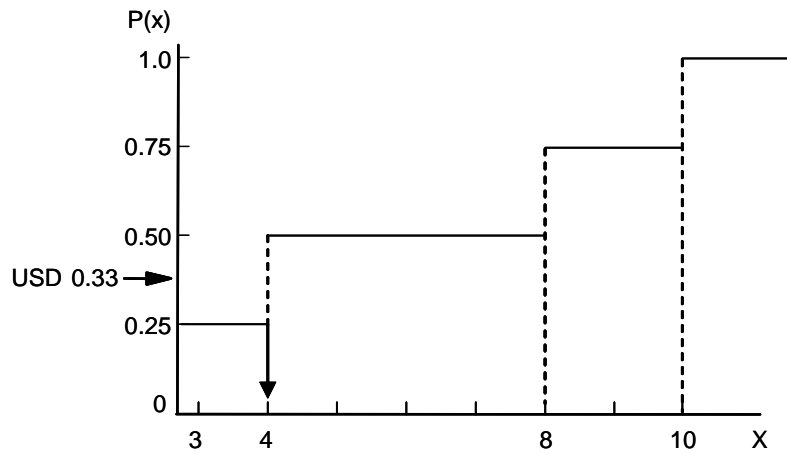


Figure 5.15. Inverse Transform for a Discrete Empirical Distribution DE(3, 4, 8, 10).

A short cut to simulating the discrete empirical distribution is to use the Simetar function rather than program the =IF() functions in Excel.

=DEMPIRICAL(3, 4, 8, 10)

– **Inverse Transform for the Triangle Distribution**

The triangle distribution is defined by the minimum, mean and maximum or a, b, c parameters, respectively. The probability of being less than the b parameter equals $P(x < b) = (b - a) / (c - a)$. To insure the random value of X is drawn from this lower quadrant a USD is used and then the USD is used in a split Inverse Transform formula. An example of a triangle would be $T \sim (2, 5, 15)$.

- In cell A1 enter
= (5-2)/(15-2)
- In cell A2 enter
= UNIFORM()
- In cell A3 enter
= IF (A2 <= A1, 2 + SQRT(USD * (15-2) * (5-2)), 0)
- In cell A4 enter
= IF (A2 > A1, 15 - SQRT((1-USD) * (15-2) * (15-5)), 0)
- In cell A5 enter
= SUM (A3:A4)

The CDF for the triangle distribution is not a pair of straight lines with a kink at the mean. Rather the square root in the function adds a curve to each segment. Figure 5.16 depicts the Inverse Transform for simulating a triangle distribution, $T(2, 5, 15)$. For USD values greater than 0.23 the random value is calculated using the upper segment AB (see Figure 5.16), for USDs less than 0.23 the random value is calculated using segment OA. (The fraction 0.23 is arrived at by the formula $(\text{mid point} - \text{min}) / (\text{max} - \text{min})$ for the $T(2, 5, 15)$ distribution.) This forces the distribution to be skewed according to the min, mean, max parameters for X.

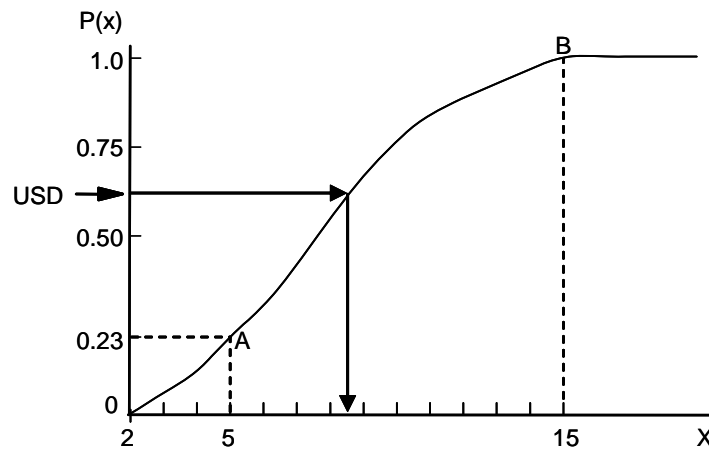


Figure 5.16. Inverse Transform for a Triangle Distribution $T(2, 5, 15)$.

– Summary

Based on the range of examples presented here, it should be evident that all random variables are generated using a USD and the Inverse Transform method. For most distributions the procedure is done automatically so modelers often forget that the stochastic source of all random variables is a USD. For advanced applications it is essential to explicitly deal with the USD so it is included as an optional parameter in the Simetar functions to simulate each random variable. For example, the fully defined Simetar function to simulate the most popular distributions described in this Chapter are:

- Uniform = UNIFORM (min, max, USD)
- Normal = NORM (mean, std. dev., USD)
- Empirical = EMPIRICAL (S_i , $F(S_i)$, USD)
- Discrete Empirical = DEMPIRICAL (S_i , USD)
- Triangle = Triangle (min, mean, max, USD)

See Chapter 16 for other probability distributions simulated by Simetar and note the optional parameter (USD) that can be specified for each distribution.

References

- Clemen, R.T. and T. Reilly. Making Hard Decisions: With Decision Tools®. Pacific Grove, CA: Duxbury, 2001.
- Law, A.M. and W.D. Kelton. Simulation Modeling and Analysis. New York: McGraw-Hill, Third edition, 2000.
- Mjelde, J.W., D.P. Anderson, K. Coble, B. Mauflik, J.L. Outlaw, J.W. Richardson, J.R. Stokes, and V. Sundarapotes. "Tutorial on Density Function Estimation and Use." Department of Agricultural Economics, Texas A&M University, FP-94-2, 1993.
- @Risk Advanced Risk Analysis for Spreadsheets. Newfield, NY: Palisade Corporation, 2000.
- Users Guide to Best Fit. Palisade Corporation, New Field: Palisade Corporation, 1995.
- Vose, D. Risk Analysis: A Quantitative Guide. New York: John Wiley & Sons, Ltd., Second Edition, 2000.
- Winston, W.L. Simulating Modeling Using @Risk©. New York: Duxbury Press 1996.

Chapter 6

Parameter Estimation for Univariate Probability Distributions

Prior to simulating a random variable you must estimate the parameters to define the probability distribution (pdf) for the variable. The parameters estimated differ from one distribution to another, however, the basic methodology is the same.

You might ask why estimate parameters for a random variable? Or ask isn't the whole variable random by definition, so why estimate the parameters? A random variable has two parts: a deterministic component (systematic variability) and a stochastic component (random variability). The parameters basically separate and quantify these components. In simulation we exogenously control or endogenously calculate the deterministic component and simulate the stochastic component. Three examples of these two components of a random variable X are:

- Assume X is distributed normal without a trend
 - The deterministic component is the mean or \bar{X} .
 - The stochastic component is the unexplained variability (risk) about the deterministic component ($\hat{\epsilon}$) and is quantified by the standard deviation parameter or $\hat{\sigma}$.

- Assume X is distributed normal about a trend
 - The deterministic component is the trend or $\hat{X} = \hat{a} + \hat{b} T$
 - The stochastic component is the unexplained variability about the trend line ($\hat{\epsilon}$) and is measured by the standard deviation of the residuals about the trend line.

- Assume X is distributed normal about a structural equation that relates X to exogenous variables Y and Z .
 - The deterministic component is the regression equation or $\hat{X} = \hat{a} + \hat{b} Y + \hat{c} Z$
 - The stochastic component is the residual about the regression equation ($\hat{\epsilon}$) and is estimated as the standard deviation of the residuals about the econometric equation.

The deterministic component of a random variable is fixed once it is estimated, as indicated in the three examples. During simulation the deterministic component of the random variable can be changed by the endogenous variables in the model, such as T , Y , and Z for the three examples. The deterministic component of a random variable can also be changed by the user to analyze alternative management strategies or policies, and to conduct sensitivity analyses.

The stochastic component of X is a measure of the dispersion along the number scale and about the positional or deterministic parameter. A key to stochastic simulation is correctly isolating the stochastic component ($\hat{\epsilon}$) from the deterministic component, in other words, separating the systematic variability from the random variability. This process is called whitening the data. We often "whiten" economic data using ordinary least squares (OLS). In econometrics the objective is to estimate the best unbiased parameters for the deterministic component of variable X , this generally amounts to minimizing the sum of the squared residuals. Unless the equations are to be simulated in a stochastic model, the stochastic component (residual) from the econometric model is discarded after estimating the parameters for the deterministic component.

The first step in parameter estimation for a probability distribution is to whiten the data so we can sort out the deterministic and the stochastic components. There are numerous procedures for "whitening" the data for a random variable and a few are outlined below for variables that have no trend, have a trend, are functions of other variables, or are a function of cycles and seasonal patterns. Alternative methods that can be used to remove systematic variability are:

- Deviations from the mean serve to whiten data series that are too short (fewer than 10 observations) to use an ordinary least squares (OLS) regression or data series that are stationary. In this case, the parameter for the deterministic component is \bar{X} and the stochastic component is what is left or $\hat{e}_i = X_i - \bar{X}$
- An OLS regression can be used to remove a systematic trend from a data series: $\hat{X} = \hat{a} + \hat{b}T$ which results in the stochastic component of $\hat{e}_i = X_i - \hat{X}_i$
- An OLS regression can remove systematic variability caused by association with other variables: $\hat{X} = \hat{a} + \hat{b}Y + \hat{c}Z$ resulting in the stochastic component of $\hat{e}_i = X_i - \hat{X}_i$
- Time series models can explain longer-term trends and seasonal variations, $X = \hat{a} + \hat{b}X_t + \hat{c}X_{t-1} + \hat{d}X_{t-2} + \dots$ resulting in the stochastic component of $\hat{e}_i = X_i - \hat{X}_i$
- First differencing the data and moving averages can also remove systematic variability, resulting in the stochastic component of $\hat{e} = X_t - X_{t-1}$

If the random variable has fewer than 10 values use the residuals from the mean to whiten the data. When there are 10 or more values for the random variable, check the trend OLS regression for a significant trend. Use the \hat{e} residuals from the OLS trend equation only if the slope coefficient is statistically different from zero, otherwise use \hat{e} calculated as the residuals from the mean. If the random variable is a function of other variables, use the residuals from the OLS equation as the stochastic component for the random variable.

This chapter provides detailed examples of how to estimate parameters for different types of univariate random variables. All of these examples are presented in the Trend Regression To Reduce Risk Demo.XLS workbook.

Random Variables Without Trend

Random variables with fewer than 10 observations and random variables with no statistically significant trend should be whitened using deviations or residuals from the mean. The mean is the deterministic portion of the random variable and the residuals from the mean are the stochastic component:

- Deterministic component
 $\bar{X} = (\sum X_i) / N$
- Stochastic component
 $\hat{e}_i = X_i - \bar{X}$

Variability about the mean, or $\hat{\epsilon}$, is the unexplained portion of the random variable. A series without trend and its residuals about the mean is depicted in Figure 6.1. The residuals about the mean constitute the stochastic portion of the random variable. Note that the residuals about zero in the bottom panel exactly match the positive and negative variations about the mean in the top panel.

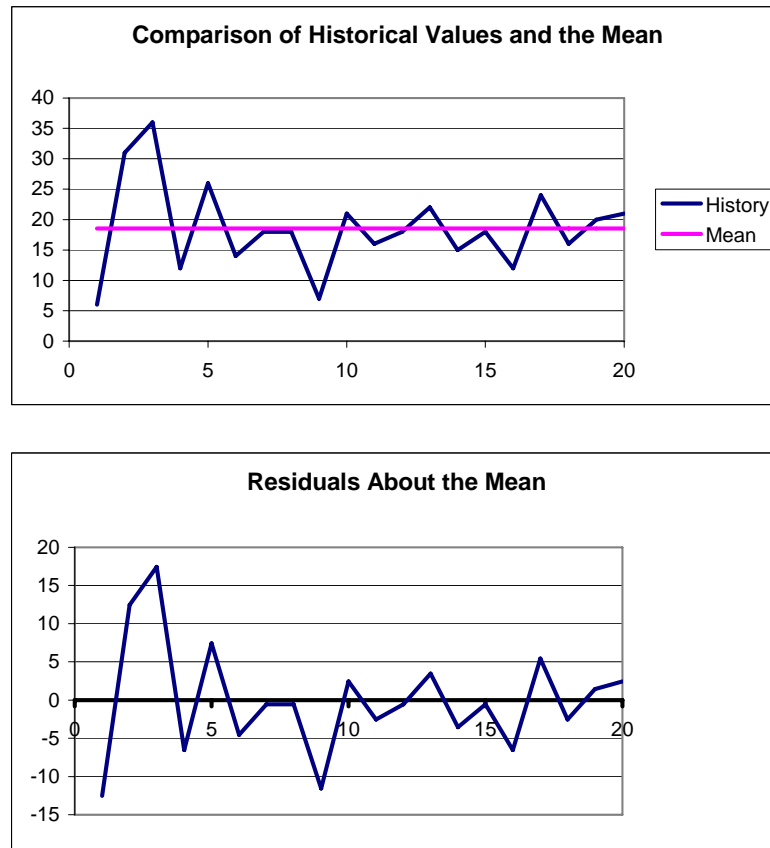


Figure 6.1. Residuals About the Mean for a Series Without Trend.

Parameter estimation for simulating a random variable with few observations or has no trend is described in this section. Parameters for simulating random variables that are distributed normal, empirical, uniform or Bernoulli are described in detail. For each distribution the following steps are used:

- Define the parameters for the distribution,
 - Describe the process for estimating the parameters,
 - Outline the steps for estimating the parameters in Simetar, and
 - Describe how the parameters are used in simulation.
- **Normal Distribution**

The parameters for the normal distribution are:

- Deterministic component is the mean \bar{X} .
- Stochastic component is the measure of dispersion about the mean or $\sigma_{\hat{\epsilon}}$.

Parameter estimation for the normal distribution is quite simple. Calculate the simple mean for the random variable without regard for whitening. Calculate the standard deviation for the stochastic component, which amounts to calculating the standard deviation of the residuals about the mean. Be sure to use the population standard deviation (=STDEVP()) rather than the sample statistic.

$$\hat{\epsilon}_i = X_i - \bar{X} \text{ for } i = 1, 2, 3, \dots, n$$

$$\hat{\sigma}_\epsilon = \text{standard deviation of the } \hat{\epsilon}_i.$$

For parameter estimation of a normal distribution, in Simetar, select the Simple Statistics icon or menu item. To be sure you calculate the correct standard deviation, select the option for the population standard deviation.

	A	B	C	D	E	F	G	H	I
25	Table 2. Estimate the Parameters and Simulate a Normal Distribution.								
26				Residuals	Formulas				
27				Mean	From Mean	in D	Parameters for Normal Distribution		
28	Mean	2.7350	2.74	-0.34	=B13-C28	Mean	2.74		
29	StDev	1.7461	2.74	-1.71	=B14-C29	Std Dev	1.7461		
30	95 % LCI	1.2807	2.74	2.25	=B15-C30				
31	95 % UCI	4.1893	2.74	-2.74	=B16-C31	Simulate Normal Distribution			
32	Min	0.0000	2.74	2.42	=B17-C32	Normal	0.675	=NORM(G28,G29)	
33	Median	2.6000	2.74	1.67	=B18-C33				
34	Max	5.1500	2.74	1.05	=B19-C34				
35	Skewness	-0.1298	2.74	-2.26	=B20-C35				
36	Kurtosis	-1.3232	2.74	-0.41	=B21-C36				
37			2.74	0.06	=B22-C37				

Figure 6.2. Parameter Estimation for a Normal Distribution.

There are three ways to simulate the normal distribution in Simetar. The formula for simulating a normal distribution is:

$$\tilde{X}_i = \bar{X} + \hat{\sigma}_\epsilon * \text{SND}$$

where SND is a standard normal deviate distributed N(0,1).

In a Simetar model the formula for a normal distribution can be simulated as:

$$= \bar{X} + \hat{\sigma}_\epsilon * \text{NORM}()$$

or

$$= \text{NORM}(\bar{X}, \hat{\sigma}_\epsilon)$$

The steps for estimating the parameters for a normal distribution are presented in Figure 6.2 and the Trend Regression To Reduce Risk Demo.XLS spreadsheet. The random variable X has 10 observations with a mean equal to 2.735. The residuals from \bar{X} are in column D of Table 1 of the Demo. The standard deviation for the residuals about the mean, 1.746, is the second parameter for the normal distribution.

– **Empirical Distribution Without a Trend**

The parameters for the empirical distribution without a trend are:

- The deterministic component is the mean \bar{X} .
- The stochastic component is the dispersion about the mean, which is expressed as the sorted fractional residuals or S_i .
- Pseudo minimum and maximum provide end points for the distribution or $Pmin_{\hat{\epsilon}}$ and $Pmax_{\hat{\epsilon}}$.
- Assign probabilities for each of the sorted residuals $P(S_i)$ with end point probabilities for the pseudo minimum and maximum equal to 0 and 1, respectively.

To estimate these parameters for the empirical distribution follow the steps outlined below and in Figure 6.3.

1. Calculate the mean for the random variable.
2. Calculate the residuals about the deterministic component ($\hat{\epsilon}_i$) and then divide the residuals by the mean to calculate the fractional deviates ($F\hat{\epsilon}_i$).
3. Sort the fractional deviates to get the sorted values S_i .
4. Calculate the pseudo minimum and maximums by multiplying the minimum and maximum S_i residuals by 1.0001 or:

$$Pmin_{\hat{\epsilon}} = (\text{minimum } S_i) * 1.0001$$

$$Pmax_{\hat{\epsilon}} = (\text{maximum } S_i) * 1.0001$$

	A	B	C	D	E	F	G	H	I	
39	Table 3. Estimate the Parameters for an Empirical Distribution.									
40	Parameters for an Empirical Distribution Without Trend						Pmin &	Final Si	Probabilities	
41	Year	X	Mean	Residuals	Res/Mean	Sorted Fi	Pmax	Values	P(Si)	
42		Pmin	X-Bar	E-hat	Fi	Si	-1.0001	1.0001	0	
43	1987	2.40	2.74	-0.34	-0.1225	-1.0000		-1.0000	0.05	
44	1988	1.03	2.74	-1.71	-0.6234	-0.8245		-0.8245	0.15	
45	1989	4.98	2.74	2.25	0.8208	-0.6234		-0.6234	0.25	
46	1990	0.00	2.74	-2.74	-1.0000	-0.1481		-0.1481	0.35	
47	1991	5.15	2.74	2.42	0.8830	-0.1225		-0.1225	0.45	
48	1992	4.40	2.74	1.67	0.6088	0.0238		0.0238	0.55	
49	1993	3.78	2.74	1.05	0.3821	0.3821		0.3821	0.65	
50	1994	0.48	2.74	-2.26	-0.8245	0.6088		0.6088	0.75	
51	1995	2.33	2.74	-0.41	-0.1481	0.8208		0.8208	0.85	
52	1996	2.80	2.74	0.06	0.0238	0.8830		0.8830	0.95	
53		Pmax					1.0001	0.8831	1	
54	Simulate the Empirical Distribution									
55	EMP	0.8005	=C52*(1+EMP(H42:H53,I42:I53))							

Figure 6.3. Parameter Estimation for an Empirical Distribution.

The parameter estimation for an empirical distribution can be done by hand using Excel commands as demonstrated in Figure 6.3 or by selecting the EMP icon on the Simetar toolbar. Next select the option to calculate the dispersion parameters as “Percentage Deviations from the Mean.” This Simetar function is particularly useful when there are several random variables in the model that will be simulated as empirical. The EMP icon function can calculate the parameters for 250 random variables in one step. The EMP icon can also estimate the parameters assuming the data are different from the mean and as actual data.

The formulas for simulating a univariate empirical distribution with Simetar that does not have a trend differs based on the form (assumption) for the S_i values.

- The S_i values are actual data:

$$=EMP(\text{Historical Data Series})$$

- The S_i values are differences from the mean:

$$=\bar{X} + EMP(S_i, P(S_i))$$

- The S_i values are fractional deviations from the mean:

$$=\bar{X} * (1 + EMP(S_i, P(S_i)))$$

The steps for estimating these parameters are summarized in Figure 6.3 and in the Trend Regression To Reduce Risk Demo.XLS workbook. The residuals, $\hat{\epsilon}_i$, values in column D of the Demo are divided by the mean to get the $F_i = \hat{\epsilon}_i/\bar{X}$ values in column E and then column E is sorted in column F to get the S_i values (Figure 6.3). The parameters for the empirical distribution are summarized on the right side of Figure 6.2 for an empirical distribution. The mean and the sorted residuals as a fraction of the mean constitute all of the parameters for the distribution. In addition to this long-hand procedure for estimating the parameters, the Simetar EMP procedure is demonstrated in the workbook.

– **Uniform**

The parameters for the uniform distribution are:

- Minimum
- Maximum

Parameter estimation for the uniform distribution is quite simple. Review the data for the random variable and identify the minimum and the maximum. Parameter estimation in Simetar is done by using the Simple Statistics icon.

The uniform distribution is simulated using the formula:

$$\tilde{X}_i = X_{\min} + (X_{\max} - X_{\min}) * USD$$

where USD is a uniform standard deviate distributed $U(0,1)$.

With Simetar the uniform distribution can be simulated two ways:

- In an Excel model the formula can be

$$= X_{\min} + (X_{\max} - X_{\min}) * \text{UNIFORM}()$$

or

$$= \text{UNIFORM}(X_{\min}, X_{\max})$$

See Table 4 in Trend Regression To Reduce Risk Demo.XLS for an example of how the parameters for a uniform distribution are calculated and simulated.

– Bernoulli

The only parameter for a Bernoulli distribution is the probability that an event occurs or p. Parameter estimation for the Bernoulli distribution amounts to counting the number of times an event occurs and dividing by the number of possible outcomes in the period observed. This could involve counting the number of days it rains in June over a 10 year period, or it could be the number of times a price is less than 2.00. These two examples would be for simulating the chance of rain in June or of an unacceptably low price.

Simetar has a function that can be used to easily calculate the p parameter for a Bernoulli distribution. The =EDF() function calculates the probability that X is less than some critical value or $p = P(X < \text{Critical Value})$. The Simetar function is programmed as:

$$=\text{EDF}(\text{Historical Data}, \text{Critical Value})$$

where Critical Value is the cut off, such as 2.00 price in the above example.

Simulate the Bernoulli function in Simetar using the function

$$=\text{BERNOULLI}(P)$$

Pressing the F9 key will produce a zero or 1. The cell with the =BERNOULLI function has a 1 p percent of the time and has a 0 (1-p) percent of the time.

For an example of how the Bernoulli parameter was calculated using Simetar's EDF function and a critical value of 2.0, see Table 5 in Trend Regression To Reduce Risk Demo.XLS.

Random Variables With a Trend

Random variables that have a significant trend must be de-trended prior to estimating parameters for the distributions. Usually the best method for de-trending a random variable is to use an ordinary least squares (OLS) regression with time as the dependent variable. First differencing the data is an alternative method for removing a trend from the data. See Table 6 in the Trend Regression To Reduce Risk Demo.XLS workbook for an example of using an OLS regression to de-trend a random variable. The regression equation to de-trend a random variable X with a linear trend is:

$$X_i = \hat{a} + \hat{b}T_i$$

where T is a series of values 1, 2, 3, or
T can be values for the years 1990, 1991, 1992,

Prior to using the trend adjusted series be sure that the slope (\hat{b}) is statistically different from zero, at least, at the 10 percent level. If the random variable (X) has a statistically significant trend, then the deterministic and stochastic components are:

- The deterministic component is the trend predicted values for X , or

$$\hat{X}_i = \hat{a} + \hat{b}T_i$$
 for any T_i
- The stochastic component is the residual about the deterministic component, or the residuals from trend

$$\hat{e}_i = X_i - \hat{X}_i$$

which is the same as

$$\hat{e}_i = X_i - (\hat{a} + \hat{b}T_i)$$

A graph depicting the variability about a trend regression is provided in Figure 6.4 to show how residuals about a trend are the stochastic portion of the random variable. Notice that the original data ranged from about 90 to more than 130 while the variability about the trend regression is on the range of -5 to +4. Thus the OLS trend regression explained a large portion of the variability (90 to 130) for the random variable. Parameter estimation for simulating a random variable that has a significant trend is described in this section. Parameters for simulating random variables which are distributed normal and empirical are described in detail.

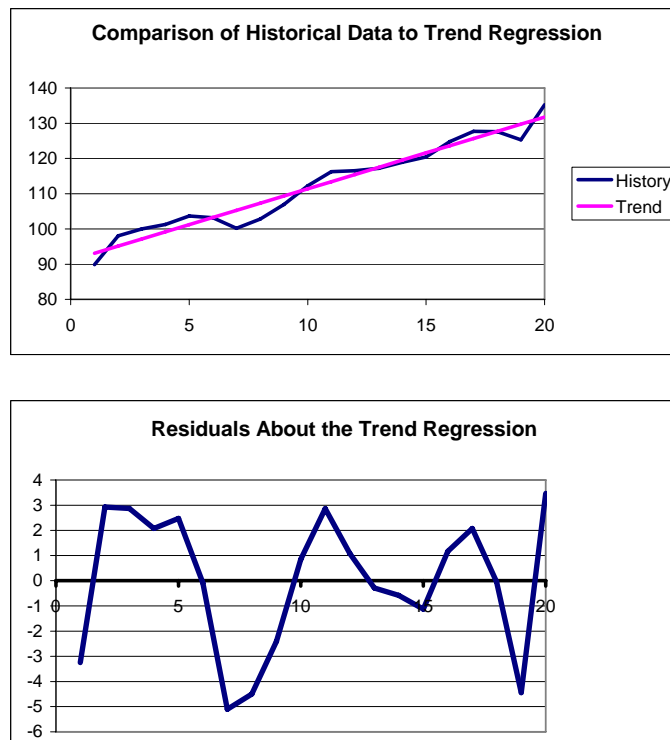


Figure 6.4. Residuals About Trend for a Series With a Trend.

– **Normal Distribution With a Trend**

Random variables that have a statistically significant trend can be simulated as a normally distributed variable. The variable itself may not be distributed normal, but the residuals about the trend regression will most likely be distributed normal. The parameters for the random variable with a trend are calculated for the deterministic and stochastic components as follows:

- Deterministic component is the trend regression or:

$$\hat{X}_i = \hat{a} + \hat{b}T_i$$

- Stochastic component is the dispersion about the trend or the $\sigma_{\hat{\epsilon}}$

Parameter estimation for the normal distribution starts by calculating the parameters for the OLS trend regression to remove the trend effect:

$$\hat{X} = \hat{a} + \hat{b} T_i$$

Next, calculate the standard deviation for the residuals from the trend line to estimate the parameter for the stochastic component:

$$\hat{\epsilon}_i = X_i - \hat{X}_i \text{ for all } i = 1, 2, 3, \dots n$$

and

$$\sigma_{\hat{\epsilon}} = \text{StdDev Population } (\hat{\epsilon}_i)$$

Parameter estimation with Simetar for a normal distribution with trend is done using the Multiple Regression icon. To start the process create a column of values to represent the trend as 1, 2, 3, ..., m. The trend values are used as the T variable in the regression $\hat{X}_i = \hat{a} + \hat{b} T_i$. Use the standard deviation for the residuals ($\hat{\sigma}_{\epsilon}$) or the standard error of the prediction (SEP_i).

To simulate the random variable with Simetar there are two options:

1. Simulate the variable using the historical mean \bar{X} :

$$=NORM(\bar{X}, \hat{\sigma}_{\epsilon})$$

2. Simulate the variable for forecasted values of X beyond the range of the historical data, such as:

$$\hat{X}_{T+i} = \hat{a} + \hat{b} T_{T+i}$$

Simulate the random variable \tilde{X}_{T+i} as:

$$=NORM(\hat{X}_{T+i}, SEP_{T+i})$$

where SEP_{T+i} is the standard error of prediction for time period T+i that is calculated by Simetar as part of the Multiple Regression calculation.

An example of simulating a random variable for the two different options is presented in Figure 6.5. The first example is the case where the historical mean and the standard deviation for the residuals is used. In the second example, an out of sample simulation is demonstrated for five years using the trend forecasted values as the mean and the standard error of the prediction for each year.

	A	B	C	D	E
134	Simulate X about the Historical Mean				
135	Average	34.13			
136	SD Residuals	2.40160618			
137	Stoch Val	37.8867485			
138					
139	Simulate X for Mean Values Outside the Range of the Data				
140	Year	Stoch Xs	Formula	X-Hat	SEPt
141	24	39.830	=NORM(D141,E141)	38.149	2.737
142	25	42.792	=NORM(D142,E142)	38.484	2.765
143	26	36.291	=NORM(D143,E143)	38.819	2.796
144	27	37.264	=NORM(D144,E144)	39.154	2.828
145	28	37.111	=NORM(D145,E145)	39.489	2.862

Figure 6.5. Example of Simulating a Random Variable With Trend as a Normally Distributed Variable.

– Empirical Distribution With a Trend

A random variable with a trend can be simulated using an empirical probability distribution. The parameters for this type of random variable are:

- The deterministic component estimated from the trend regression or:

$$\hat{X}_i = \hat{a} + \hat{b} T_i$$

- The stochastic component is the unexplained variability about the deterministic component or:

$$\hat{e}_i = X_i - \hat{X}_i$$

The \hat{e}_i series is transformed to be the sorted deviations from the deterministic component or:

$$S_i = \text{Sorted}(\hat{e}_i / \hat{X}_i)$$

- The empirical distribution also requires calculation of pseudo minimum and maximum values for the empirical distribution or:

$$P_{\min} = S_{\min} * 1.0001$$

$$P_{\max} = S_{\max} * 1.0001$$

- The final parameter for the empirical distribution are the probabilities, $P(S_i)$, for the S_i values. The $P(S_i)$ range from 0 at the P_{\min} to 1 for the P_{\max} . The $P(S_i) = 0$ indicates there is a zero probability that a random value of S_i will be less than S_{\min} . By the same logic there is a zero probability that a random S_i will exceed S_{\max} .

Simetar calculates the parameters for a random variable with trend that is to be simulated as an empirical random variable. The EMP icon can calculate the parameters for one or more random variables with a trend. If the analyst specifies that the random variable(s) has a trend, Simetar calculates the regression equation for a linear trend, reports the slope and intercept, the residuals ($\hat{\epsilon}_i$), calculates the fractional deviates from trend (F_i), the sorted fractional deviates (S_i), the P_{\min} and P_{\max} , and finally the probabilities ($P(S_i)$) to define the empirical distribution.

Figure 6.6 summarizes the parameter estimation process for the empirical probability distribution of a random variable with trend. In the first panel the linear trend line through the actual data suggests the presence of a statistically significant trend. The residuals from the trend regression line, as a fraction of the trend forecasted values, in the second panel show how removing the trend reduces the variability of the variable to plus and minus 15 percent. The fractional residuals are graphed as a PDF in the third panel to show the range of variability about the deterministic component in simulation as a probability distribution. The fourth panel is developed using the S_i and $P(S_i)$ values to show the CDF of fractional deviates for the stochastic component.

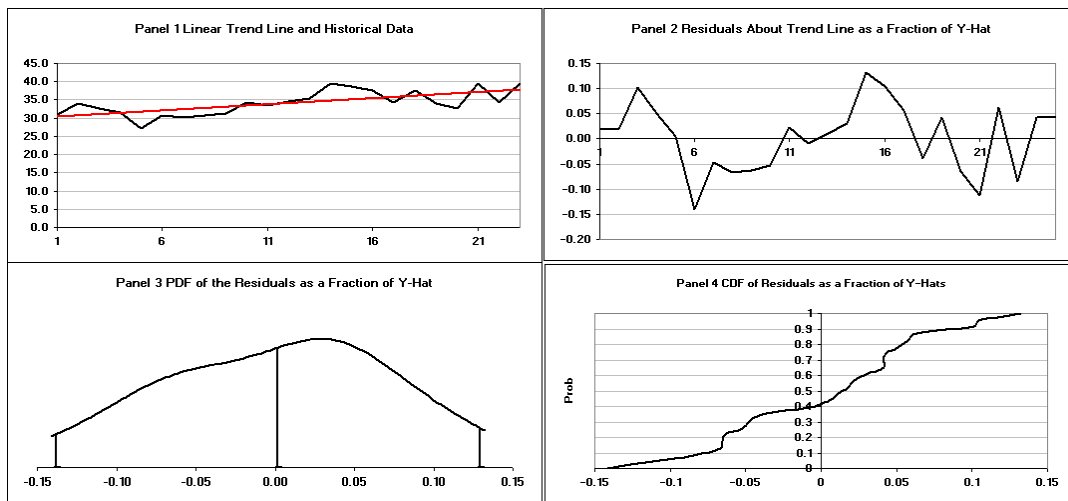


Figure 6.6. Parameter Estimation for a Trend Dependent Variable Assuming an Empirical Distribution.

The random variable can be simulated for an “in sample” or “out of sample” forecasted \hat{X} value or:

$$\tilde{X} = \bar{X} * (1 + EMP(S_i, P(S_i)))$$

or

$$\tilde{X}_{T+i} = \hat{X}_{T+i} * (1 + EMP(S_i, P(S_i)))$$

During simulation the =EMP() function generates a random uniform standard deviate on the 0-1 scale and interpolates the CDF defined by S_i and $P(S_i)$ in panel 4 of Figure 6.6 to simulate a stochastic \tilde{S}_i . Because the \tilde{S}_i is a fraction of \hat{X} by adding \tilde{S}_i to 1.0 and multiplying it by the \hat{X}_{T+i} the random value for the $T+i$ out of sample period can be simulated.

The result of detrending X is to reduce its variability by removing the systematic component. The effect of expressing the \hat{e}_i 's as a fraction of the \hat{X} s and simulating them as empirical causes the simulated \tilde{X} s for any period $T+i$ to have the same coefficient of variation (CV) as the historical period. The benefit of a constant CV is that the simulated variables do not increase or decrease the relative risk in the model. See Chapter 9 for a further discussion of this topic.

Random Variables as a Function of Other Variables

Random variables in economic models are usually a function of other variables. For example, yield and exports are considered endogenous variables in agricultural sector models. In stochastic simulation these endogenous variables can be treated as stochastic variables. We know there is a deterministic and a stochastic component to these variables because they are not completely random. The deterministic component is removed through econometrics, leaving a stochastic (residual) component that can be simulated as the "random" portion of the variable.

The example in Table 8 of the Trend Regression To Reduce Risk Demo.XLS workbook is for US wheat acreage. Wheat acreage is a function of lagged wheat prices, lagged wheat acres, and wheat acres set aside or idled by the CRP. The OLS equation for the variable is:

$$\hat{X}_i = 30.28 + 5.88 P_{i-1} + 0.29 A_{i-1} - 0.285I_i$$

The multiple regression equation defines the deterministic component of US wheat acreage for a random variable. The stochastic component of US wheat acreage is the residual about the predicted \hat{X}_i values. The estimated standard deviation of the residuals for US wheat acreage is 2.697, less than half the standard deviation for wheat acreage (7.609).

A chart depicting the residuals about an OLS regression is provided in Figure 6.7 to show how OLS reduced the variability for the stochastic portion of the variable. Note that the historical data ranged from about 45 to 80 with apparently no trend and that the residuals about the OLS model range from -4 to +6. Also, note that the residuals in the bottom panel match the misses by the OLS regression in the top panel.

An econometric model can be simulated stochastically by simulating random error terms for the equations. The analyst is cautioned to estimate the best econometric model possible for the deterministic component of each random variable. It is recommended that the error term for an econometric model be simulated assuming it is distributed normal. However, it is possible to simulate the error terms as empirical so both procedures are described in this section.

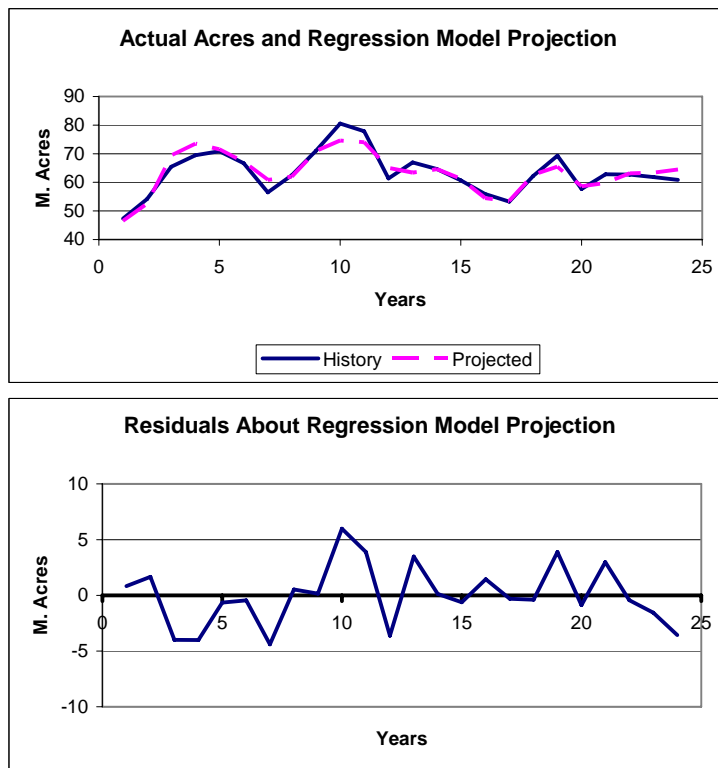


Figure 6.7. Residuals from a Multiple Regression Model.

When a multiple regression is used to forecast/explain a random variable, the residuals from the regression can be used to simulate a probabilistic forecast. The variable has two components: a deterministic and a stochastic component.

- The deterministic component is the forecasted \hat{X} value for the assumed values of the exogenous variables or:

$$\hat{X}_i = \hat{a} + \hat{b}Y_i + \hat{c}Z_i + \hat{d}T_i$$

Alternative values for the endogenous variables Y, Z, and T result in alternative forecasts of \hat{X}_i , which may or may not be considered out of sample forecasts.

- The stochastic component is the residuals or σ_ϵ from the regression or:

$$\sigma_\epsilon = X_i - \hat{X}_i$$

The assumption regarding the residuals determines how the variable will be simulated. If the residuals are assumed to be normally distributed then the regression model will be simulated as a normal distribution. If, however, the residuals do not follow a normal distribution the model can be simulated as a non-parametric (empirical) distribution. In both cases the \hat{X}_{T+i} value is the same. Prior to choosing a distribution for the residuals it is useful to examine the consequences of the two distributions.

- Assuming the residuals are distributed normal is consistent with the t test and F test to determine if the model is acceptable. However, due to the unbounded nature of the normal distribution, it can result in simulated values that are outside the historical range for the variable. Also, for out of sample forecasts the coefficient of variation (CV) will be biased relative to the historical CV.
- Assuming the residuals are distributed empirical prevents the minimum and maximum simulated values from exceeding their relative historical counterparts. Additionally, the CV for out of sample forecasts will be approximately equal to the historical CV.
- **Simulating a Multiple Regression Model as Normal**

The parameters for simulating the stochastic component, if normality is assumed, are the forecasted \hat{X}_{T+1} deterministic component and either the standard deviation of the residuals ($\sigma_{\hat{\epsilon}}$) or the standard error of the prediction (SEP). The $\sigma_{\hat{\epsilon}}$ is often the appropriate parameter if the forecast value is in the historical sample range. The SEP_{T+1} is the appropriate parameter for out of sample forecasts of \hat{X}_{T+1} . In the Multiple Regression option of Simetar both parameters are calculated.

To simulate a multiple regression model for an in sample analysis, assuming the stochastic component is distributed normal, use the following equation and function:

$$\begin{aligned}\hat{X}_t &= \hat{a} + \hat{b}Y_t + \hat{c}Z_t + \hat{d}T_t \\ &= \text{NORM}(\hat{X}_t, \sigma_{\hat{\epsilon}})\end{aligned}$$

To simulate an out of sample forecast from a multiple regression use the following equation and function:

$$\begin{aligned}\hat{X}_{T+1} &= \hat{a} + \hat{b}Y_t + \hat{c}Z_t + \hat{d}T_t \\ &= \text{NORM}(\hat{X}_{T+1}, SEP_{T+1})\end{aligned}$$

When the SEP_{T+1} is used to simulate the out of sample values the resulting CV will approximately equal the historical CV.

- **Simulating of a Multiple Regression Model as Empirical**

The parameters for simulating the stochastic component if it is assumed to be distributed as an empirical distribution are the forecasted \hat{X}_{T+1} deterministic component and the S_i values.

The S_i values are the sorted residuals divided by the \hat{X} 's or:

$$S_i = \text{Sorted}(\hat{\epsilon}_i / \hat{X}_i)$$

The S_i and $P(S_i)$ values define the empirical probability distribution for the stochastic component of the random variable (Figure 6.8). When the residuals are expressed as a fraction of the predicted \hat{X} s the stochastic, \tilde{X} , will have approximately the same CV as the historical data, even for out of sample \hat{X} values.

To simulate a multiple regression model for an in or out of sample case, assuming the stochastic component is distributed empirical, use the following formula and function:

$$\begin{aligned} \hat{X}_{T+i} &= \hat{a} + \hat{b}Y_{T+i} + \hat{C}Z_{T+i} + \hat{d}T_{T+i} \\ &= \hat{X}_{T+i} * (1 + \text{EMP}(S_i, P(S_i))) \end{aligned}$$

This formula can be used if $\hat{X}_{T+i} = \bar{X}$ or it is the forecasted value for any period $T+i$.

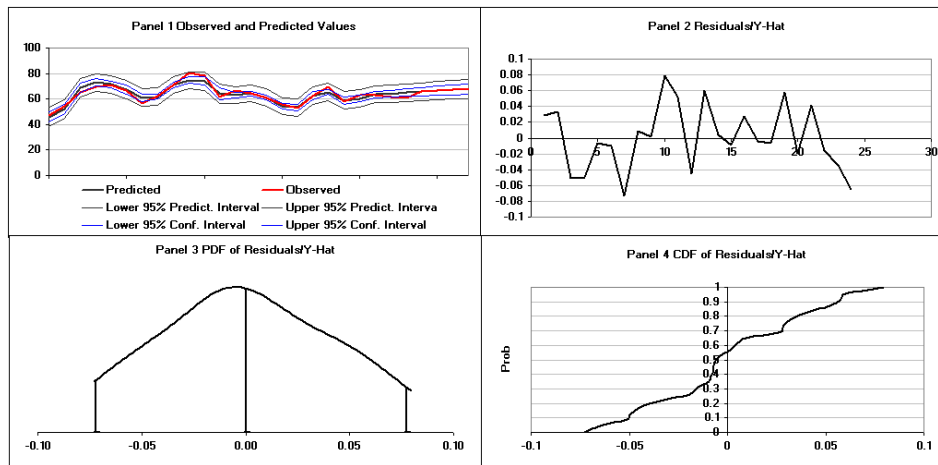


Figure 6.8. Parameter Estimation for a Random Variable Dependent on Other Variables Assuming an Empirical Distribution.

– Additional Options for Simulation of Multiple Regression Models

The two previous sections covered two methods for simulating the stochastic component of a multiple regression model. Two more situations can be simulated using a multiple regression model: stochastic betas and stochastic exogenous variables.

If the betas for the multiple regression model are stochastic, the model can be simulated as:

$$\hat{X}_{T+i} = \tilde{a} + \tilde{b}Y_{T+i} + \tilde{C}Z_{T+i} + \tilde{d}T_{T+i} + \tilde{e}$$

where \tilde{a} , \tilde{b} , \tilde{c} , and \tilde{d} must be simulated as a multivariate normal distribution. See Chapter 7 for a description of simulating multivariate probability distributions.

The \tilde{e} stochastic component in the equation can be simulated normal or empirical as described in the previous sections.

Alternatively the multiple regression model can be simulated using stochastic exogenous variables, such as:

$$\hat{X}_{T+i} = \hat{a} + \hat{b}\tilde{Y}_{T+i} + \hat{c}\tilde{Z}_{T+i} + \hat{d}\tilde{T}_{T+i} + \tilde{e}$$

where the \tilde{Y} , \tilde{Z} , and \tilde{T} exogenous variables are stochastic and simulated from their own distributions.

The exogenous variables can be distributed independent univariate or multivariate. An example of this type of model is simulating a local price based on the stochastic national price or:

$$L_{T+i} = \hat{a} + \hat{b}\tilde{N}_{T+i} + \tilde{e}$$

where L is the local price and N is the exogenous stochastic national price.

Random Variables a Function of an Autoregressive Structure

Time series models provide a convenient method for forecasting univariate random variables. Such models estimate the parameters for the deterministic component of a random variable. Residuals from a time series model represent the stochastic component for the random variable. Therefore time series models can be simulated stochastically to provide probabilistic forecasts for future periods.

The procedure for simulating a time series model follows the steps outlined in the previous section for a random variable that is a function of other variables. It is recommended that the error term for a time series model be simulated, assuming the residuals are distributed normal.

Estimating Parameters for Other Distributions

Parameters for distributions other than normal and empirical are generally more difficult to estimate. The parameters are often estimated using a maximum likelihood estimator or a methods of moments estimator owing to the fact that the functions for the parameters are generally not solved directly for a single value. Simetar provides a Univariate Parameter Estimator (UPE) that estimates the parameters for 16 parametric probability distributions. The UPE assumes the data for a random variable has no trend (i.e., that the data are stationary). Parameters are calculated for the following distributions, if they are applicable:

- Beta
- Double Exponential
- Exponential
- Gamma
- Logistic
- Log-Log
- Log-Logistic
- Lognormal
- Normal
- Pareto
- Uniform

- Weibull
- Binomial
- Geometric
- Poisson
- Negative Binomial

Simetar calculates the parameters for each distribution using a maximum likelihood estimator (MLE) and a method of moments estimator (MOM). Also, Simetar prepares the functions for simulating the random variables using the calculated parameters. If the random variable does not conform to a particular distribution type the parameters are not reported. For example, the random variable evaluated in Figure 6.9 does not conform to a Binomial so the parameters are not reported. The parameters reported in Figure 6.9 are calculated both as MLE and as MOM. The functions for simulating the random variable, assuming MLE or MOM parameters, are provided in the two right hand columns (Figure 6.9).

The results of a simulation test on the ability of the different distributions and their associated parameters to simulate the random variable are summarized in Figure 6.10. The CDFDEV values are scalars from a loss function and indicate how closely the simulated distribution fits the random variable. If the simulated distribution is a perfect fit, the CDFDEV value is zero. The results in Figure 6.10 suggest that if MLE parameters are used, the uniform and beta distributions would be best, while if MOM parameters are used the beta is best.

An interactive Simetar assisted program is available in Test Parameters Demo.XLS. An example of using the program to visualize a normal distribution with the same mean but three standard deviations is presented in Figure 6.11. To use Test Parameters Demo.XLS the user selects the distributions in cells A7-A9 from drop down menus and then types in parameters in cells B7-C9. Simetar approximates the specified distributions and displays them as CDFs and PDFs (Figure 6.11).

The interactive Simetar program can be used to view a particular distribution with three alternative parameters assumptions, as in Figure 6.11 or to compare distributions. For example, three alternative distributions (e.g., beta, gamma, and normal) can be compared and parameters specified so each has the same mean.

	A	B	C	D	E	F	G	H	I
351	Random Var.	Univariate Parameter Estimation for Random Var. at 4/6/2005 4:40:41 PM							
352	2.40	Univariate Parameter Estimation for Random Var. at 11/17/2005 9:55:49 PM							
353	2.60			MLEs		MOM Estimates		Random Variables	
354	4.98	Distribution	Parameters	Parm. 1	Parm. 2	Parm. 1	Parm. 2	MLE	MDM
355	0.50	Beta	$\alpha, \beta; A \leq x \leq B, \alpha, \beta > 0$	0.441	0.500	0.539	0.493	0.480	0.481
356	5.15	Double Expc	$\alpha, \beta; \alpha \leq x < \infty, -\infty < \mu < \infty, \beta$	2.700	1.280	2.942	1.171	-3.282	-2.531
357	4.40	Exponential	$\mu, \sigma; -\infty < x < \infty, -\infty < \mu < \infty$	0.480	2.462	1.286	1.656	0.492	1.294
358	3.78	Gamma	$\alpha, \beta; 0 \leq x < \infty, \alpha, \beta > 0$	2.188	1.345	3.155	0.932	0.180	0.350
359	0.48	Logistic	$\mu, \sigma; 0 \leq x < \infty, -\infty < \mu < \infty, c$	2.972	0.936	2.942	0.913	-2.047	-1.954
360	2.33	Log-Log	$\mu, \sigma; -\infty < x < \infty, -\infty < \mu < \infty$	2.146	1.496	2.197	1.291	-0.369	0.027
361	2.80	Log-Logistic	$\mu, \sigma; 0 \leq x < \infty, -\infty < \mu < \infty, c$	2.237	2.623	3.787	2.616	0.239	0.635
362		Lognormal	$\mu, \sigma; 0 \leq x < \infty, -\infty < \mu < \infty, c$	0.834	0.822	0.941	0.525	0.272	0.655
363		Normal	$\mu, \sigma; -\infty < x < \infty, -\infty < \mu < \infty$	2.942	1.571	2.942	1.656	-1.142	-1.363
364		Pareto	$\alpha, \beta; \alpha \leq x < \infty, \alpha, \beta > 0$	0.480	0.638	1.974	3.038	0.484	1.977
365		Uniform	$a, b; a \leq x \leq b$	0.480	5.150	0.073	5.811	0.502	0.100
366		Weibull	$\alpha, \beta; 0 \leq x < \infty, \alpha, \beta > 0$	1.844	8.960	1.842	9.079	0.179	0.180
367		Binomial	$n, p; x = 0, 1, 2, \dots, n; 0 \leq p \leq 1$						
368		Geometric	$p; x = 1, 2, \dots; 0 \leq p \leq 1$						
369		Poisson	$\lambda; x = 0, 1, \dots; 0 \leq \lambda < \infty$	2.400		2.400		0	0
370		Negative Bir	$r, p; x = 1, 2, \dots; 0 \leq p \leq 1$						

Figure 6.9. Sample Output for the Univariate Parameter Estimator.

	A	B	C	D	E	F
389	CDFDEV Values comparing the different distributions for simulating the Random Variable					
390	and the two methods for estimating parameters for the distributions.					
391	Distribution	EV if MLE	CDFDEV if MOM			
392	Beta	4.80693	4.218			
393	Double Exponential	4.55308	4.268			
394	Exponential	5.54597	5.002			
395	Gamma	4.86276	4.645			
396	Logistic	4.24102	4.268			
397	Log-Log	4.56101	4.597			
398	Log-Logistic	4.72936	4.679			
399	Lognormal	5.27803	4.765			
400	Normal	4.26800	4.268			
401	Pareto	7.99689	4.914			

Figure 6.10. Comparison of Simulating a Random Variable with Alternative Distributions Using MLE and MOM Estimates of the Parameters.

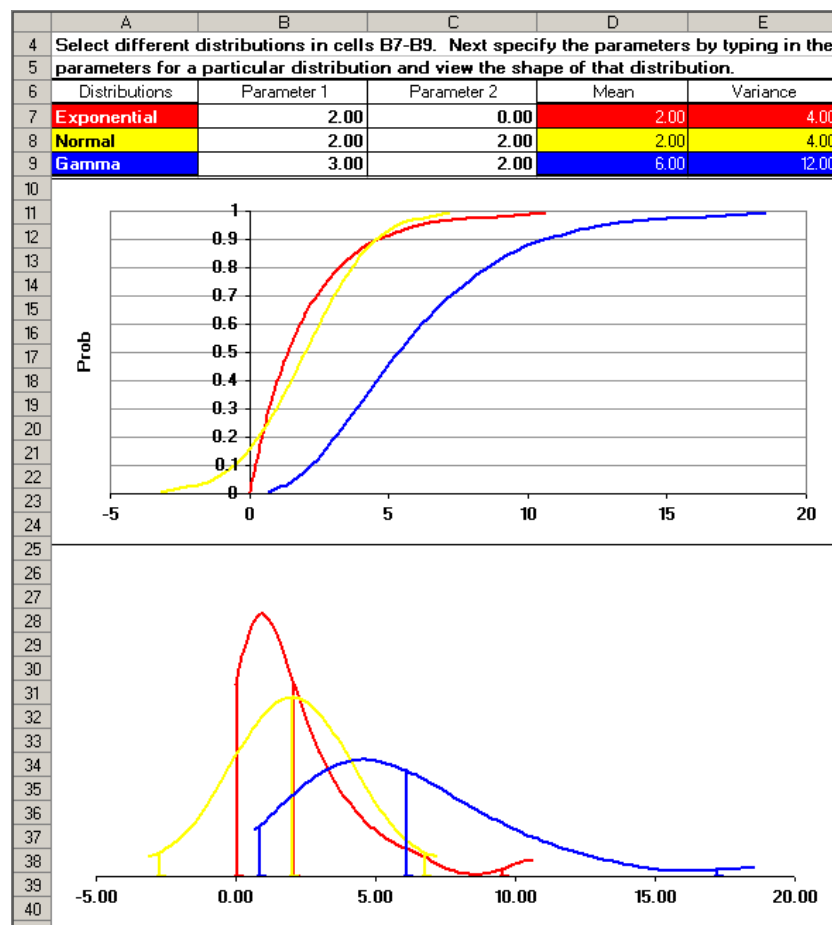


Figure 6.11. Sample Output for Test Parameters Demo.XLS.

Summary

The procedure for estimating parameters for a univariate probability distribution can be summarized as follows:

- Estimate the best model possible to quantify the deterministic component of the random variable. This model may range from the mean (\bar{X}) to a multiple regression ($\hat{a} + \hat{b}Y + \hat{c}Z$).
- Calculate the stochastic component for the random variable by subtracting the deterministic forecasted values from their respective observed values to get \hat{e} .
- Calculate the appropriate parameter(s) to quantify \hat{e} . The parameter for a normal distribution is the standard deviation ($\hat{\sigma}$) while the parameters for the empirical distribution are sorted fractional deviates (S_i) and their associated probabilities ($P(S_i)$).

These steps are logical if you think about separating a random variable into its two components: the deterministic and stochastic parts. The deterministic component is the systematic variability and the stochastic component is the uncontrollable or random variability.

References

- Hardaker, J.B., Huirne, R.B.M., and J.R. Anderson. Coping with Risk in Agriculture. New York: CAB International, 1997.
- McCarl, B. "Forming Probability Distributions." Department of Agricultural Economics, Texas A&M University, 1988.
- Mjelde, J.W., D.P. Anderson, K. Coble, B. Maufluk, J.L. Outlaw, J.W. Richardson, J.R. Stokes, and V. Sundarapothes. "Tutorial on Density Function Estimation and Use." Department of Agricultural Economics, Texas A&M University, FP-94-2, 1993.
- O'Brien, D., M. Hayenga, and B. Babcock. "Deriving Forecast Probability Distributions of Harvest-Time Corn Futures Prices." Review of Agricultural Economics, 18(1996): 167-180.

Chapter 7

Parameter Estimation for Multivariate Probability Distributions

Univariate probability distributions are for individual random variables so multivariate (MV) probability distributions are for two or more random variables that are dependent on one another. Multivariate distributions are the rule in economic analysis models because most variables are correlated to each other. The purpose of this chapter is to describe and demonstrate how to estimate and apply parameters for multivariate distributions. This chapter builds on Chapter 6, which describes how to simulate univariate probability distributions.

The chapter is separated into three parts: multivariate normal (MVN) distributions, multivariate empirical (MVE) distributions, and simulating very large MVE distributions. The MVN and MVE sections deal with correlating random variables within years or intra-temporal correlation. For the problem of simulating inter-temporally correlated random variables see Chapter 8, after working through this chapter. Chapter 8 provides a comprehensive treatment of intra- and inter- temporal correlation and is recommended for advanced work in simulation.

Ignoring Correlation

If two random variables are correlated and their correlation is ignored in simulation the model will either over or under state the variance and mean for the system's KOVs. The direction of the bias introduced on the variance is inversely related to the correlation. Ignoring a positive correlation between \tilde{X} and \tilde{Y} will understate the variance for Z if $Z = \tilde{X} + \tilde{Y}$. Ignoring a negative correlation between \tilde{X} and \tilde{Y} will overstate the variance for Z in the same case.

The reason why the variance of Z is inversely biased relative to the correlation between X and Y is due to the variance formula for variable Z :

Let $Z = \tilde{X} + \tilde{Y}$ where \tilde{X} and \tilde{Y} are random variables,

the expected value of Z is

$$E(Z) = E(X) + E(Y)$$

the variance for Z is

$$V(Z) = \sigma_X^2 + \sigma_Y^2 + 2 \text{Cov}(X, Y) = \sigma_X^2 + \sigma_Y^2 + 2 \rho_{xy} * \sigma_x * \sigma_y$$

where ρ_{xy} is the correlation between X and Y .

When X and Y are negatively correlated the $\text{Cov}(X, Y)$, or ρ_{xy} is negative and reduces $V(Z)$ so ignoring the correlation overstates the true variance of Z . The opposite is true when X and Y are positively correlated because the $\text{Cov}(X, Y)$ or ρ_{xy} is positive. In both cases the mean, $E(Z)$, is unbiased by ignoring the correlation between X and Y .

On the other hand if the KOV is a function of products of random variables, as $Z = \tilde{X} * \tilde{Y}$, the mean will be biased if correlation between \tilde{X} and \tilde{Y} is ignored. In this case the expected value and variance of Z is:

$$E(Z) = E(X) + E(Y) + \text{Cov}(X, Y) = E(X) + E(Y) + 2 \rho_{xy} * \sigma_x * \sigma_y$$

$$V(Z) = \sigma_x^2 + \sigma_y^2 + 2 \text{Cov}(X, Y) = \sigma_x^2 + \sigma_y^2 + 2 \rho_{xy} * \sigma_x * \sigma_y$$

The mean and variance for Z are over or under estimated, inversely with respect to the sign on the correlation coefficient, if correlation is ignored in simulation.

As demonstrated in this chapter simulating a multivariate probability distribution is very easy and automatically corrects for the potential of biasing the mean and variance. The procedure described for simulating multivariate distributions insures that the random variables are “appropriately” correlated, meaning that the historical correlation is maintained in the simulation process.

Multivariate Normal (MVN) Distribution

Two or more normally distributed random variables that are correlated must be simulated as a MVN distribution to prevent biasing the model results. Check for correlation of the random variables by calculating the simple correlation coefficients among the variables. If the correlation coefficients are significantly different from zero the variables must be simulated MVN. A Student-t test is used to test each correlation coefficient in the correlation matrix to determine if it is statistically different from zero at, say, the 95 percent level. Simetar provides a Student-t test of correlation coefficients when calculating the correlation matrix (Figure 7.1). (See Chapter 16 and Correlation Demo.XLS for an example of this test.) The example in Figure 7.1 uses a critical t value of 2.20 for a 95% confidence test. The calculated t-statistics (in the lower matrix) which are larger than the critical value indicate their corresponding correlation coefficient is statistically different from zero. For ease of interpretation these calculated t values are bold.

	A	B	C	D	E	F	G
20	Correlation Matrix						
21		Corn P	Wheat P	Sorg P	Corn Y	Wheat Y	Sorg Y
22	Corn P	1	0.87	0.98	-0.30	-0.18	-0.53
23	Wheat P		1	0.82	-0.16	-0.37	-0.34
24	Sorg P			1	-0.29	-0.17	-0.56
25	Corn Y				1	0.53	0.59
26	Wheat Y					1	0.47
27	Sorg Y						1
28							
29	Correlation Coefficient t-values.						
30	Significanc	95%		t-critical	2.20		
31		Corn P	Wheat P	Sorg P	Corn Y	Wheat Y	Sorg Y
32	Corn P		5.83	16.79	1.04	0.60	2.06
33	Wheat P			4.68	0.54	1.31	1.21
34	Sorg P				1.00	0.58	2.27
35	Corn Y					2.08	2.40
36	Wheat Y						1.78
37	Sorg Y						
38	Bold values indicate statistical significance at the specified level.						

Figure 7.1. Statistical Test of Correlation Coefficients.

A MVN distribution has three parameters (components) to be quantified and is described here for a model with four random variables. The three components for a four variable MVN distribution are:

- Deterministic component for each of the four variables is the mean, or forecast, or \hat{X}_j for $j = 1, 2, 3, 4$.
- Stochastic component for each of the four variables is the standard deviation about the mean or forecast, or $\hat{\sigma}_{\epsilon_j}$ for $j = 1, 2, 3, 4$
- Multivariate component for the four variables is represented by a 4x4 correlation matrix, or ρ (or the covariance matrix or Σ).

- **Parameters for a MVN Distribution**

The deterministic component of a MVN can be the mean or the predicted value from a trend regression, multiple regression, or time series model for each of the random variables, such as:

$$\hat{X}_{ij} = \hat{a} + \hat{b}_1 T_i + \hat{b}_2 X_{i-1} + \hat{b}_3 Z_i$$

or simply the mean

$$\hat{X}_{ij} = \bar{X}_{ij}$$

where \hat{X}_{ij} are the predicted values for all random variables $X_j, j = 1, 2, 3, \dots m$, and i denotes the periods (years, months, etc.) over which the variable is to be simulated.

The stochastic component for the MVN distribution is the measure of the dispersion about the deterministic component. The dispersion measure for a normal distribution is the standard deviation (σ_{ϵ}). The standard deviation is calculated using the residuals about the mean or forecast and is defined for each j variable in the distribution as:

$$\hat{\epsilon}_{ij} = X_{ij} - \hat{X}_{ij}$$

$$\hat{\sigma}_{\epsilon_j} = \text{standard deviation for the } \hat{\epsilon}_{ij} \text{'s.}$$

where $\hat{\sigma}_{\epsilon_j}$ is the standard deviation of the residuals for each of the random variables $X_j, j = 1, 2, 3, \dots m$.

The $\hat{\sigma}_{\epsilon_j}$ is calculated over the T historical periods used to calculate the deterministic component, the \hat{X}_j .

The multivariate component for the MVN distribution is generally the correlation matrix of rank m for the m random variables. The correlation matrix must be calculated using the residuals ($\hat{\epsilon}_{ij}$), i.e., the stochastic component. For a 4 variable model the ρ matrix is:

$$\rho = \begin{bmatrix} 1.0 & \rho_{12} & \rho_{13} & \rho_{14} \\ & 1.0 & \rho_{23} & \rho_{24} \\ & & 1.0 & \rho_{34} \\ & & & 1.0 \end{bmatrix}$$

An alternative method for simulating a MVN distribution uses the covariance matrix for the multivariate component. For a 4 variable MVN model the covariance matrix, Σ , is:

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ & \sigma_{22}^2 & \sigma_{23} & \sigma_{24} \\ & & \sigma_{33}^2 & \sigma_{34} \\ & & & \sigma_{44}^2 \end{bmatrix}$$

– Parameter Estimation for the MVN Distribution

The steps for estimating the parameters for a MVN distribution are:

1. Calculate the best model possible to predict each of the random variables, whether this is simply the mean, a trend regression, a multiple regression, or a time series model.

$$\hat{X}_{ij} = \text{econometric model}$$

2. Calculate the residuals, $\hat{\epsilon}_{ij}$, from the econometric forecast for each random variable.
3. Calculate the standard deviations, $\hat{\sigma}_j$, for each random variable using their residuals.
4. Calculate the correlation matrix (ρ) and calculate the covariance matrix (Σ) for the random variables using the residuals. (Note: Use the residuals to calculate the matrices because the residuals are the stochastic component of the variables to be correlated. Calculating the correlation and covariance matrices from the actual data is equivalent to calculating the multivariate measures about the mean which is not the same as the correlation for the residuals.)

– Parameter Estimation Using Simetar

The Simple Statistics and Multiple Regression options in Simetar will most often be used to estimate the deterministic components for MVN distributions. When the Multiple Regression function is used, Simetar forecasts the random variable, \hat{X}_{ij} , and estimates the standard deviation for the residuals, $\hat{\sigma}_j$. Additionally, Simetar calculates the standard error of prediction, $\hat{\sigma}_{jp}$, or SEP_j which should be used in place of a standard deviation of the residuals for simulating a variable distributed normal.

Use the residuals provided in the Multiple Regression function’s output to calculate the correlation matrix and the covariance matrix. Simetar provides a Correlation function for calculating the ρ matrix and the Σ matrix and testing the correlation coefficients for significance.

An example of estimating the parameters for a 3 variable MVN distribution is included in Multivariate Normal Distribution Demo.XLS. Each of the four steps for MVN parameter estimation are identified and the distribution is simulated different ways using Simetar. The example begins with the data for the three random variables in rows 8-34. In Step 1, OLS regression results show significant trends for all three random variables. The residuals from trend for each random variable are calculated using the Simple Regression option in Simetar in Step 1. Standard deviations for the residuals are calculated using an Excel function in line 108. The unsorted residuals are used in Step 2 to calculate the correlation matrix.

– **Simulating a MVN Distribution**

Three methods for simulating a MVN probability distribution are presented here. The technical description of what is involved in simulating the MVN distribution is provided in an Appendix at the end of this Chapter.

The first method for simulating a MVN distribution uses the correlation matrix to simulate CUSD’s. An example of this method for a three variable MVN distribution is presented in Figure 7.2 (see Multivariate Normal Distribution Demo.XLS). For this method a vector of CUSD’s is simulated using =CUSD (Correlation Matrix). The CUSD’s are used individually to simulate MVN random variables using the Simetar function:

$$=NORM(\hat{X}_j, StdDev_j, CUSD_j)$$

	A	B	C	D	E	F	G	H	I	J	K	L	M
118	Use Correlation Matrix to Generate CUSDs and =NORM() Function to Simulate MVN Distribution												
119	Random				Correlation Matrix								
120	Yields	Means	Std Dev	CUSDs	Corn	Wheat	Sorghum	Formulas in Column B				Formula for Column E	
121	Corn	110.0931	107.559	11.44636	0.5876	1	0.23	0.69	=NORM(C121,D121,E121)				=CUSD(F121:H123)
122	Wheat	34.0981	35.312	2.3554	0.303102		1	0.24	=NORM(C122,D122,E122)				
123	Sorghum	65.2169	60.006	6.072636	0.804565			1	=NORM(C123,D123,E123)				

Figure 7.2. Simulating a MVN Distribution Using the Correlation Matrix.

The second method for simulating a MVN distribution uses the covariance matrix to simulate stochastic correlated deviations (or CDEVs). A CDEV is the number of deviations from the mean that the random value lies. For this method a vector of CDEV’s is simulated using =CSND (Covariance Matrix). The CDEV’s are used individually in the formula:

$$\tilde{X}_i = \hat{X}_i + CDEV_i$$

This method is demonstrated for simulating a three variable MVN distribution in Figure 7.3. (See the Multivariate Normal Distribution Demo.XLS for this example.)

	A	B	C	D	E	F	G	H	I	J	K	L	
125	Use the Covariance Matrix to Generate Correlated Deviations (CDEVs)												
126		Random	Histroical										
127	Year 1	Yields	Means	CDEVs	Covariance Matrix			Formulas in Column B		Formula in Column D			
128	Corn	124.5414	107.559	16.98221	302.6983	42.96493	114.6301	=C128+D128			=CSND(E128:G130)		
129	Wheat	38.1642	35.312	2.851893	0	13.38022	17.57007	=C129+D129					
130	Sorghum	65.31816	60.006	5.311842	0	0	62.536	=C130+D130					

Figure 7.3. Example of Using Covariance Matrix to Simulate MVN Distribution in Two Steps.

The third method for simulating a MVN distribution does it all in one step using Simetar's MVNORM function. The one step formula is the easiest method for simulating a MVN distribution over a multiple year planning horizon. The MVNORM function simulates the MVN values for the i^{th} forecast period as:

$$\hat{X}_{ij} = \text{MVNORM}(\text{Vector of } \hat{X}_{ij}, \text{Covariance Matrix})$$

The one step method for simulating a MVN distribution is demonstrated in Figure 7.4.

	A	B	C	D	E	F	G	H	I	
110	A One Step Approach to Simulating a MVN Distribution.									
111	Assemble the parameters and use the =MVNORM() function									
112		Random		Covariance Matrix						
113		Yields	Means	Corn	Wheat	Sorghum	Formula in Column B			
114	Corn	84.24	107.559	302.6983	42.96493	114.6301	=MVNORM(C114:C116,D114:F116)			
115	Wheat	31.52	35.312		13.38022	17.57007				
116	Sorghum	51.46	60.006			62.536				

Figure 7.4. Example of the One Step Procedure for Simulating a MVN Distribution for One Year.

The one step method for simulating a MVN distribution is demonstrated in Figure 7.5 for a three variable MVN distribution that is simulated for four years.

	A	B	C	D	E	F	G	H
141	Simulate 5 Years for a MVN distribution in One Step							
142	Forecast Means for Each Year to Simulate							
143	Years	Corn	Wheat	Sorghum				
144	27	131.1439	40.34983	69.12418				
145	28	132.8909	40.72298	69.79958				
146	29	134.6379	41.09613	70.47498				
147	30	136.3849	41.46928	71.15038				
148	Covariance Matrix							
149		Corn	Wheat	Sorghum				
150	Corn	302.6983	42.96493	114.6301				
151	Wheat		13.38022	17.57007				
152	Sorghum			62.536				
153	Simulate Stochastic Forecast of the MVN for 4 Years							
154	Years	Corn	Wheat	Sorghum				
155	Y27	106.5542	39.55058	66.31111	=MVNORM(B144:D144,\$B\$150:\$D\$152)			
156	Y28	141.2978	45.432	76.61209	=MVNORM(B145:D145,\$B\$150:\$D\$152)			
157	Y29	103.6693	35.8383	48.31042	=MVNORM(B146:D146,\$B\$150:\$D\$152)			
158	Y30	139.0412	42.41304	65.17593	=MVNORM(B147:D147,\$B\$150:\$D\$152)			

Figure 7.5. Example of the One Step Procedure for Simulating a MVN Distribution for Four Years.

– **Results of Simulating MVN Distribution**

After simulating a multivariate probability distribution one must first test the simulated values to be sure they are appropriately correlated. This involves calculating the correlation matrix for the simulated random variables and comparing this correlation matrix to the original. Simetar includes a statistical test to assist in validating the correlation of multivariate distributions. The validation procedure for a MVN distribution is demonstrated using the example in Multivariate Normal Distribution Demo.XLS.

The correlation test used in validation compares the correlation matrix used for the simulation to the implicit correlation matrix in the simulated variables. For a model with three random variables the Check Correlation in Simetar’s Hypothesis Testing for Data menu, the user provides the location for the three simulated random variables, say B:8; B108 of the SimData worksheet and the location for the correlation matrix to simulate the three random variables. The null hypothesis for the test is that each correlation coefficient for the simulated variables equals the original or assumed correlation coefficient or:

$$H_0: \hat{\rho}_{ij} = \rho_{ij}$$

The test uses a Student –t test and the critical value at the alpha equal 5% or less based on sample size. To reject the null hypothesis the calculated t-test statistic must exceed the critical value of, say, 2.43 at the 98.3% confidence level. The calculated t-test statistics in Figure 7.6 are all less than 2.43 so we fail to reject the null hypothesis and say the simulated correlation coefficients are statistically equal to the assumed (historical) correlation coefficient at the 98% level. This test must be done for each year simulated by the MVN distribution.

	O	P	Q	R	S	T
2	Test Correlation Coefficients for Simulated Values in Year 27					
3	Confidence Level		98.3048%			
4	Critical Value		2.43			
5						
6		Wheat	Sorghum			
7	Corn	0.84	0.25			
8	Wheat		0.47			

Figure 7.6. Example of a Student –t Test for a Correlation Matrix.

Simetar provides three non-parametric tests for validating MVN distributions. The tests are:

- Two Sample Hotelling T² Test – tests the historical mean vector vs. the mean vector for the simulated variables.
- Box’s M Test – tests the historical covariance matrix vs. the covariance matrix for the simulated variables.
- Complete Homogeneity Test – simultaneously tests the historical means vector and covariance matrix vs. the means vector and covariance matrix for the simulated variables.

The tests are described in Chapter 3 and 16 in more detail. If the tests “Fail to Reject the Null Hypothesis” that the simulated parameters equal the historical parameters, then the simulation process simulated the MVN distribution appropriately. An example of using the non-parametric tests to validate a MVN distribution is provided in Figure 7.7. When the Test Values are less than the Critical Values, the tests will Fail to Reject the null hypothesis and this is what we want.

	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	A
28	Distribution Comparison of Corn & Corn													
29	Confidence Level		95.0000%											
30		Test Value	Critical Val	P-Value										
31	2 Sample Hotelling T ² Test	0.00	8.17	1.000	<i>Fail to Reject the Ho that the Mean Vectors are Equal</i>									
32	Box's M Test	0.61	12.59	0.996	<i>Fail to Reject the Ho that the Covariance Matrices are Equivalent*</i>									
33	Complete Homogeneity Test	0.58	16.92	1.000	<i>Fail to Reject the Ho that the Mean Vectors and Covariance Matrices are Equivalent, Respectively*</i>									
34	<i>*Based on asymptotic distribution</i>													

Figure 7.7. Non-Parametric Tests for Testing Simulated Values from MVN Distributions.

Multivariate Empirical (MVE) Distribution

Two or more correlated random variables can be simulated as a multivariate distribution, even if the variables are not normally distributed. For example, non-normal MV distributions can be simulated as a MVE distribution. The generalized MV procedure presented here for the MVE distribution allows one to correlate non-normal distributions in a simulation model. Richardson and Condra first introduced the procedure in 1978 and an extension of their original procedure is used here and expanded upon in Chapter 8. King later reported the procedure to simulate a multivariate beta distribution. The workbook Multivariate Empirical Demo.XLS demonstrates the steps for MVE parameter estimation and simulation described in this section.

A MVE distribution has three parameters or components to be estimated. The MVE is described here for a model with four random variables, even though the procedure can easily be expanded for an m variable MVE. The three components/parameters are:

- Deterministic component for each of the four variables, \hat{X}_j for $j = 1, 2, 3, 4$.
- Stochastic component for each of the four variables, $S_{\hat{e}_j}$ for $j = 1, 2, 3, 4$
- Multivariate component for the four variables $\rho_{4 \times 4}$.

– Parameters for a MVE Distribution

The deterministic component is the projected value based on the mean, trend regression, multiple regression, or time series model for each of the random variables, such as:

$$\hat{X}_{ij} = \hat{a} + \hat{b}_1 T_i + \hat{b}_2 X_{i-1} + \hat{b}_3 Z_i$$

or simply the mean:

$$\hat{X}_{ij} = \bar{X}_{ij}$$

where \hat{X}_{ij} are the predicted values for all random variables X_j , $j = 1, 2, 3, \dots, m$, and i denotes the periods (years, months, etc.) over which the variable is to be simulated.

The stochastic component for the MVE distribution is the measure of the dispersion about the deterministic component or $S_{\hat{\epsilon}}$. The dispersion measure for an empirical distribution is the vector of sorted deviations from the deterministic component, expressed as a fraction of the forecasted values at each historical period i . The $S_{\hat{\epsilon}}$ values are calculated for each random variable as:

$$\hat{\epsilon}_{ij} = X_{ij} - \hat{X}_{ij}$$

$$F_{\hat{\epsilon}_{ij}} = \hat{\epsilon}_{ij} / \hat{X}_{ij}$$

$$S_{\hat{\epsilon}_{ij}} = \text{Sorted}(F_{\hat{\epsilon}_{ij}})$$

where $S_{\hat{\epsilon}_{ij}}$ are the sorted fractional residuals for each of the random variables X_j , $j = 1, 2, 3, \dots, m$ over the historical period $i = 1, 2, 3, \dots, T$

Multivariate component for the MVE distribution is the correlation matrix of rank m for the m random variables. The correlation matrix must be calculated using the unsorted residuals ($\hat{\epsilon}_{ij}$), i.e., the stochastic component. For a 4 variable model the ρ matrix is:

$$\rho = \begin{bmatrix} 1.0 & \rho_{12} & \rho_{13} & \rho_{14} \\ & 1.0 & \rho_{23} & \rho_{24} \\ & & 1.0 & \rho_{34} \\ & & & 1.0 \end{bmatrix}$$

– **Parameter Estimation for the MVE Distribution**

The steps for estimating the parameters for a MVE distribution are:

1. Calculate the best model possible to predict each of the random variables, whether this is simply the mean or a complex econometric model, such as, a trend regression, a multiple regression, or a time series model.
 \hat{X}_{ij} = econometric model
2. Calculate the residuals, $\hat{\epsilon}_{ij}$, from the econometric forecast for each random variable.
3. Calculate the $m \times m$ correlation matrix for all of the random variables using the unsorted residuals. (Note: Use the residuals to calculate the correlation matrix because the residuals are the stochastic component of the variables to be correlated. Calculating the correlation matrix from the actual data is equivalent to calculating the multivariate measures about the mean which is not the same as the correlation for the residuals.

4. Calculate the fractional residuals for each variable and then sort these values for each of the random variables. Calculate the pseudo minimums and maximums for each variable using the sorted fractional residuals.
5. Assign probabilities to each of the sorted fractional residuals including a zero to the pseudo minimum and a one to the pseudo maximum.

– Parameter Estimation Using Simetar

The EMP icon on the Simetar toolbar can be used to calculate all of the parameters for an MVE distribution. The EMP icon calculates the S_{eij} values for multiple variables assuming the MVE uses either actual data, deviations from mean or deviations from trend. If more complex forecasting models are needed to project \hat{X}_{ij} , use the residuals from the econometric model and select the Actual Data option in the EMP icon to estimate the parameters.

The steps for estimating the parameters for a MVE distribution are presented in Multivariate Empirical Demo.XLS. The worksheet starts with six random variables and goes through the steps described above for estimating the parameters. The data do not have statistically significant trends so fractional deviations from the mean are used to estimate the S_i parameters for the empirical distributions. The resulting MVE distribution is simulated three different ways in Steps 7-9.

– Simulating a MVE Distribution

A MVE distribution can be simulated several ways in Excel using Simetar functions. Three methods are presented here starting with the easiest and proceeding to the most complex. But first the general procedure is described in matrix notation for completeness.

The first step is to simulate an $M \times 1$ vector of correlated uniform standard deviates or CUSD's. The Simetar array function =CUSD() performs the necessary calculations and simulation. In its most simple form the CUSD function is =CUSD (Correlation Matrix) so the result is m cells with CUSDs as demonstrated in Figure 7.8.

	E	F	G	H	I	J	K	L	M
160	CUSDs			Correlation Matrix					Formula in Column E
161	0.3354	1.000	0.869	0.981	-0.299	-0.175	-0.529	=CUSD(F161:K166)	
162	0.5353	0.000	1.000	0.816	-0.159	-0.367	-0.342		
163	0.2788	0.000	0.000	1.000	-0.288	-0.169	-0.565		
164	0.8504	0.000	0.000	0.000	1.000	0.531	0.587		
165	0.3809	0.000	0.000	0.000	0.000	1.000	0.473		
166	0.8180	0.000	0.000	0.000	0.000	0.000	1.000		

Figure 7.8. Example of Simulating a Vector of CUSD's.

The final step in simulating a MVE distribution is to use the CUSD's in an empirical distribution. The formula varies depending upon the format of the S_j values:

- S_j are actual data

$$\tilde{X}_j = \text{EMP}(S_j, P(S_j), \text{CUSD}_j)$$

- S_j are absolute deviations from means

$$\tilde{X}_j = \bar{X}_j + \text{EMP}(S_j, P(S_j), \text{CUSD}_j)$$

- S_j are fractional deviations from means

$$\tilde{X}_j = \bar{X}_j + \bar{X}_j * \text{EMP}(S_j, P(S_j), \text{CUSD}_j)$$

- S_j are fractional deviations from a forecast

$$\tilde{X}_j = \hat{X}_j + \hat{X}_j * \text{EMP}(S_j, P(S_j), \text{CUSD}_j)$$

The EMP() Simetar function simulates an empirical distribution defined by S_j and $P(S_j)$ using the uniform standard deviate indicated by the CUSD_j .

There are two ways to simulate the MVE distribution with Simetar. The first method uses the =MVEMP() function and it takes only one step. The one step method calculates all of the parameters in the background and also generates the CUSD's. Program the one step MVE distribution function as:

$$=\text{MVEMP}(\text{Actual Data},,,, \text{Forecasted Xs}, \text{Code})$$

where: Actual Data is the location for the original historical data,
 The ,,,, must be provided and indicate optional parameters that are not provided,
 Forecasted Xs represent the \hat{X}_{ij} for the i^{th} period to simulate,
 Code is a switch to specify the format for the Sjs, as 0 for actual data,
 1 for fractional deviates from mean,
 2 for fractional deviates from trend, and
 3 for differences from mean.

The =MVEMP() function is an array function so highlight m cells representing the m random variables in the MVE distribution. Additionally, the function is completed by pressing Control Shift Enter. An example of simulating a six variable MVE distribution using the one step method is presented in Figure 7.9.

	A	B	C	D	E	F	G
148		Random					
149	Names	Values	Formulas Used to Simulate the MVE in One Step				
150	Corn Pr	2.98	=MVEMP(B11:G23,,,B144:G144,1)				
151	Wheat Pr	3.63					
152	Sorghum Pr	2.64	Note B11:G23 is the actual data and B144:G144 is the				
153	Corn Yld	127.04	forecasted deterministic component of the MVE distribution				
154	Wheat Yld	43.05					
155	Sorghum Yld	67.53					

Figure 7.9. One Step Method to Simulate an MVE Distribution.

The second method for simulating an MVE distribution with Simetar is a two step process. First generate an $m \times 1$ vector of CUSD's. Next use the CUSD's in the EMP function applying the appropriate formula based on the form of the S_j to each of the m random variables:

$$=\hat{X}_j + \hat{X}_j * \text{EMP}(S_j, P(S_j), \text{CUSD}_j)$$

Repeat this formula for each of the m variables in the distribution.

This method is demonstrated in Figure 7.10 for a six variable MVE distribution where the S_j 's are specified as fractional deviations from the mean.

	A	B	C	D	E	F	G	H	I	J	K
159		Random		Correlated							
160	Names	Values	Means	Deviates as %	CUSDs			Correlation Matrix			
161	Corn Pr	2.23	2.325	-0.0395	0.3354	1.000	0.869	0.981	-0.299	-0.175	-0.529
162	Wheat Pr	3.32	3.294	0.0063	0.5353	0.000	1.000	0.816	-0.159	-0.367	-0.342
163	Sorghum Pr	1.90	2.128	-0.1066	0.2788	0.000	0.000	1.000	-0.288	-0.169	-0.565
164	Corn Yld	132.49	118.363	0.1194	0.8504	0.000	0.000	0.000	1.000	0.531	0.587
165	Wheat Yld	36.04	37.152	-0.0300	0.3809	0.000	0.000	0.000	0.000	1.000	0.473
166	Sorghum Yld	69.96	64.855	0.0787	0.8180	0.000	0.000	0.000	0.000	0.000	1.000
167		=C166+C166*D166			=CUSD(F161:K166)						
168				=EMP(G\$122:G\$136,\$A\$122:\$A\$136,E166)							

Figure 7.10. Example of Using a Two Step Method to Simulate a MVE Distribution.

If there is a one step method, what is the need for the two step method? The two step method is used when not all of the random variables have the same form for the S_j 's. For example, if the first three X s have no trend so they are simulated as actual data and the next three X 's have a trend and must be simulated as fractional deviations from trend.

– Results of Simulating MVE Distribution

A validation test to insure that the random variables are appropriately correlated should be done before using the random values in a decision model. Simulate the MVE distribution collecting values for the stochastic variables and then test the correlation implicit in the simulated values against the correlation matrix for the historical data. Validation of the simulated random variables is particularly important for non-normal distributions because the procedure is not widely used and understood. Simetar provides a correlation test and three non-parametric tests for validating MVE distributions. All four tests should be used to validate all multivariate distributions.

The six variable MVE distribution simulated in the Multivariate Empirical Demo.XLS is used to demonstrate the four validation tests in the SimData worksheet. The results of the four tests are summarized in Figure 7.11. The results of the correlation test reveal that all of the correlation coefficients implicit in the simulated variables are statistically equal to their counterparts in the actual data's correlation matrix, at the 99% level. Additionally the mean vectors and covariance matrices for the simulated data and the actual data are statistically equal at the 95 percent level.

	I	J	K	L	M	N	O	P	Q	R	S	T	U	
7	Test Correlation Coefficients													
8	Confidence Level	99.6586%												
9	Critical Value	3.00												
10														
11		Wheat Pr	Sorghum F	Corn Yld	Wheat Yld	Sorghum Yld								
12	Corn Pr	0.38	1.19	0.58	0.83	1.27								
13	Wheat Pr		1.13	0.68	0.68	0.55								
14	Sorghum Pr			0.72	0.70	0.93								
15	Corn Yld				1.55	0.37								
16	Wheat Yld					1.29								
17														
18	Distribution Comparison of Corn Pr & Corn Pr													
19	Confidence Level	95.0000%												
20	Test Value	Critical Val	P-Value											
21	2 Sample t	1.77	13.73	0.945	<i>Fail to Reject the Ho that the Mean Vectors are Equal</i>									
22	Box's M Test	12.83	32.67	0.914	<i>Fail to Reject the Ho that the Covariance Matrices are Equivalent*</i>									
23	Complete L	14.69	40.11	0.974	<i>Fail to Reject the Ho that the Mean Vectors and Covariance Matrices are Equivalent, Respectively*</i>									
24					<i>*Based on asymptotic distribution</i>									

Figure 7.11. Validation Tests for MVE Distributions.

Mixed Multivariate Probability Distributions

When parameter estimation (or the problem being analyzed) requires simulation of a distribution where the random variables are not all normal or all empirical, what do you do? Ignoring the correlation of these variables would bias the key output variables in the model by over-or under-stating their means and risk. The procedure presented in this section allows for the appropriate correlation of random variables with different distributions. For example, one variable can be normal, another can be empirical, and another can be uniform or beta and yet the probability distribution can be appropriately correlated in simulation, so historical variability and correlation will be observed in the simulation. Richardson and Condra reported this procedure in 1987.

– Parameter Estimation

The deterministic component for each random variable must be quantified using the best model possible, such as, mean, trend regression, multiple regression, or time series model. The stochastic component for each variable is the residual from the deterministic component and must be calculated for each variable in the MV distribution, i.e., estimate the $\hat{\epsilon}_{ij}$'s. The multivariate component of the MV distribution (ρ matrix) must be estimated using the unsorted residuals or $\hat{\epsilon}_{ij}$'s. The parameters to quantify the stochastic component for each of the random variables are estimated based on the appropriate parameters for each variable's assumed distribution, i.e., $\hat{\sigma}$ for the normal, S_i and $P(S_i)$ for the empirical, minimum and maximum for the uniform, and so on.

– Simulation Steps

1. Generate m correlated uniform standard deviates using the correlation matrix and the Simetar command =CUSD for an $m \times 1$ array of $CUSD_i$, where $i = 1, 2, 3, \dots, m$.
2. Use the inverse transform formulas to simulate random values for each variable by applying the appropriate parameters for the variable's deterministic and stochastic components and the variable's respective $CUSD_i$.

– Simulating a Mixed Multivariate Distribution

Assume the mixed MV distribution to simulate has 4 random variables, defined and distributed as follows:

$X \sim \text{Normal}(\bar{X}, \hat{\sigma})$

$Y \sim \text{Empirical}(S_i, P(S_i))$

$Z \sim \text{Empirical}(S_i, P(S_i))$

$W \sim \text{Uniform}(\text{min}, \text{max})$

The multivariate component for the mixed MV distribution is the correlation matrix for the four random variables. The correlation coefficients for the ρ matrix must be estimated using the residuals from the deterministic component for the X , Y , and Z variables and the historical data for variable W . The actual historical data has to be used for W because of the nature of the uniform distribution. The resulting correlation matrix is:

$$\rho = \begin{bmatrix} \rho_{\hat{e}x_t, \hat{e}x_t} & \rho_{\hat{e}x_t, \hat{e}y_t} & \rho_{\hat{e}x_t, \hat{e}z_t} & \rho_{\hat{e}x_t, w_t} \\ & \rho_{\hat{e}y_t, \hat{e}y_t} & \rho_{\hat{e}y_t, \hat{e}z_t} & \rho_{\hat{e}y_t, w_t} \\ & & \rho_{\hat{e}z_t, \hat{e}z_t} & \rho_{\hat{e}z_t, w_t} \\ & & & \rho_{w_t, w_t} \end{bmatrix}$$

Generate a vector of correlated uniform standard deviates (CUSDs) of size $m \times 1$ using the array function =CUSD as follows:

Block a 4×1 array and type the command =CUSD (Correlation Matrix Range), then press the Control Shift Enter keys.

The resulting 4×1 array has four values that change as the F9 key is pressed and during simulation, Simetar will generate new values for each iteration. The four values in the array are correlated based on the correlation matrix.

Apply the appropriate Simetar function to simulate the random numbers for each random variable being sure to use each variables $CUSD_i$,

$$\tilde{X} = \text{NORM}(\hat{X} + \hat{\sigma}, \text{CUSD}_1)$$

$$\tilde{Y} = \hat{Y} + \hat{Y} * \text{EMP} (S_{y_i}, F(S_i), \text{CUSD}_2)$$

$$\tilde{Z} = \hat{Z} + \hat{Z} * \text{EMP} (S_{z_i}, F(S_i), \text{CUSD}_3)$$

$$\tilde{W} = \text{UNIFORM}(\text{Min}, \text{Max}, \text{CUSD}_4)$$

The Simetar functions for generating random variables all allow for specification of a random uniform standard deviate (see Chapter 16). In the case of a MV distribution you must specify a CUSD rather than an independent USD.

The Multivariate Mixed Probability Distribution Demo.XLS spreadsheet provides an example of how to simulate random variables that have different probability distributions, as a multivariate distribution. The first variable is assumed to be normally distributed, the second and third are distributed empirical and the last is distributed uniform. The MV distributions simulated as indicated in (cells B79-B82) use their respective CUSD's in the inverse transform formulas Figure 7.12.

	A	B	C	D	E	F	G	H	I	J
77		Random								
78		Values				Mean	Std Dev	Min	Max	CUSDs
79	Var X Normal	2.1823438	=NORM(F79,G79,J79)			2.127	0.369			0.5595
80	Var Y EMP	2.5497697	=F80*(1+EMP(F32:F43,E32:E43,J80))			2.361				0.7477
81	Var Z EMP	172.11025	=F81*(1+EMP(G32:G43,E32:E43,J81))			160.7				0.6124
82	Var W Uniform	89.959372	=UNIFORM(H82,I82,J82)					61.310	92.740	0.9115

Figure 7.12. Example of Simulating a MV Mixed Distribution Using CUSD's and the Inverse Transform Functions.

Four validation tests were used to statistically determine if the MV mixed distribution reproduced the historical correlation matrix, means vector, and covariance matrix. The null hypothesis that the simulated test statistic (ρ , \bar{X} and Σ) equal their historical counterparts were not rejected, indicating that the procedure worked. The test statistics are reported in Figure 7.13.

	H	I	J	K	L	M	N	O	P	Q	R	S	T
8	Test Correlation Coefficients												
9	Confidence Level	99.1488%											
10	Critical Value	2.64											
11													
12		Var Y EM	Var Z EMP	Var W Uniform									
13	Var X Nor	1.58	0.15	0.34									
14	Var Y EMP		0.49	1.06									
15	Var Z EMP			0.71									
16													
17	Distribution Comparison of Var X Normal & Var X												
18	Confidence Level	95.0000%											
19		Test Value	Critical Value	P-Value									
20	2 Sample	4.52	9.62	0.344	Fail to Reject the Ho that the Mean Vectors are Equal								
21	Box's M T	1.42	18.31	0.999	Fail to Reject the Ho that the Covariance Matrices are Equivalent*								
22	Complete	2.00	23.68	1.000	Fail to Reject the Ho that the Mean Vectors and Covariance Matrices are Equivalent, Respectively								
23					*Based on asymptotic distribution								

Figure 7.13. Validation Tests for a MV Mixed Distribution.

Simulating Large Multivariate Distributions

For multivariate (MV) distributions containing a large number of variables, it is often impossible to use the procedures described for the MVN and MVE distributions. The symptom that the MVE and MVN methods will not work, is that Excel returns “# VALUE” in the CUSD or CSND arrays. The function fails because the correlation matrix is not positive definite, and therefore the correlation matrix (ρ) can not be factored by the square root method (Choleski decomposition). A factored matrix (ρ) is required to simulate the MVE and MVN distributions. The equations for simulating the MVE and MVN distributions show the dependence of the method on the R matrix:

$$\begin{aligned} \text{Let } \rho &= N \times N \text{ correlation matrix} \\ \text{and } R_{(N \times N)} &= \sqrt{\rho}_{(N \times N)} \\ \text{CSND}_{(N \times 1)} &= R_{(N \times N)} * \text{ISND}_{(N \times 1)} \\ \text{also } \text{CUSD}_{(N \times 1)} &= \text{ERF}(R_{(N \times N)} * \text{ISND}_{(N \times 1)}) \end{aligned}$$

where: ERF is the error function for integrating the area under a standard normal distribution from $-\infty$ to z and is calculated using Excel's function =NORMSDIST.

The Choleski decomposition of the correlation matrix is calculated in Simetar using the =MSQRT (Correlation Matrix) function. If the correlation matrix is positive definite, the MSQRT function will return a non-zero value in every cell in the upper right triangle and the main diagonal of the result matrix. On the other hand, if MSQRT returns a matrix with zeros in the upper right triangle or main diagonal (or #VALUE) the distribution cannot be simulated using the MVE or MVN procedure. See the correlation matrix and its factored matrix in Figure 7.14 for an example of a matrix which is not positive definite. The example comes from the Bad Correlation Matrix Demo.XLS.

As a further test of a problem matrix, the determinate of the full symmetric correlation matrix or the covariance matrix should be calculated to make sure it is positive or that it is not so close to zero it causes exponent overflows in =MSQRT. The correlation icon on the Simetar toolbar can be used to calculate the full symmetric correlation or covariance matrix. The determinate of a square matrix can be calculated using the determinate function in the Simetar Matrix Operations dialog box. If the determinate for either the correlation or covariance matrix is negative or almost zero the matrix cannot be factored. This result is generally due to the fact that there are too many “large” correlation coefficients or the number of correlation coefficients outside the ± 0.50 causes the MV distribution to be over specified.

When the MVE and MVN procedures described in this Chapter cannot be used, there are two options for simulating a MV distribution: (a) use a bootstrap simulation technique or (b) re-arrange the correlation matrix.

	A	B	C	D	E	F	G	H	I	J	K
21	Correlation Matrix of 10x10 Variables										
22		Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Var 10
23	Var 1	1	0.21	-0.05	0.30	-0.12	0.35	0.09	0.19	-0.47	0.00
24	Var 2		1	0.78	0.76	0.89	0.64	0.84	0.60	0.17	0.52
25	Var 3			1	0.80	0.93	0.79	0.88	0.26	0.68	0.62
26	Var 4				1	0.79	0.80	0.92	0.33	0.26	0.70
27	Var 5					1	0.66	0.91	0.35	0.50	0.60
28	Var 6						1	0.79	0.26	0.43	0.40
29	Var 7							1	0.32	0.34	0.69
30	Var 8								1	-0.06	0.46
31	Var 9									1	0.31
32	Var 10										1
33											
34		Factor a 10x10 Square, Symmetric Matrix. 12/14/2003 3:46:01 PM									
35		#####	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
36		#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
37		#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
38		#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
39		#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
40		#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
41		#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
42		#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
43		#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
44		#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!

Figure 7.14. Example of a Correlation Matrix that Can Not be Factored by the Square Root Method.

– **Bootstrapping a Multivariate Distribution**

Bootstrap methodology can be used to simulate a multivariate empirical distribution if there are a large number of observations for the random variables. Bootstrap simulation is based on Elfon’s (1979) work on bootstrapping univariate distributions. For a complete description of univariate bootstrapping see Chapter 11. If applied properly bootstrap simulation of MV distributions maintains not only the correlation among variables, but also higher order moments and any multi-modal characteristics of the variables. To simulate a MV distribution using bootstrap simulation methodology, do the following:

1. Prepare the historical data in a table with all random variables contiguous and in the proper temporal order, $X_{1i}, X_{2i}, \dots X_{mi}$ for m variables, with years i for the rows. See Figure 7.15 for a data table in the proper format.
2. Use Simetar’s =BOOTSTRAPPER function to randomly draw rows from the data matrix. The array function is programmed as:

=BOOTSTRAPPER(Data Matrix, TRUE)

where: TRUE instructs Simetar to draw m values from the m columns of the Data Matrix and that all of M values must come from one row.

Simetar draws rows of values from the data matrix at random if the function is used properly as an array function and is concluded by pressing Control Shift Enter. See the example in Figure 7.15 taken from the Bootstrap Multivariate Distribution Demo.XLS.

	A	B	C	D	E	F	G	H	I	J
77		Random								
78		Values				Mean	Std Dev	Min	Max	CUSDs
79	Var X Normal	2.1823438	=NORM(F79,G79,J79)			2.127	0.369			0.5595
80	Var Y EMP	2.5497697	=F80*(1+EMP(F32:F43,E32:E43,J80))			2.361				0.7477
81	Var Z EMP	172.11025	=F81*(1+EMP(G32:G43,E32:E43,J81))			160.7				0.6124
82	Var W Uniform	89.959372	=UNIFORM(H82,I82,J82)					61.310	92.740	0.9115

Figure 7.15. Using the Bootstrapping Function to Simulate a MV Distribution.

The benefits to the bootstrap simulation methodology for a MV empirical distribution are: (a) it draws random \tilde{X}_j observations in historically observed paired groups so the historical intratemporal correlation is maintained, (b) the historical data are used to simulate each variable's distribution so no assumptions about distributional shapes or types are required, and (c) the method is efficient in that no parameters must be calculated. The disadvantage to the methodology is that the values simulated are the discrete historical values because interpolation between observed points is not employed. To minimize this being a significant disadvantage, the original sample in the data matrix must be sufficiently large to define the population being simulated, as no new values are generated by interpolating among observed values.

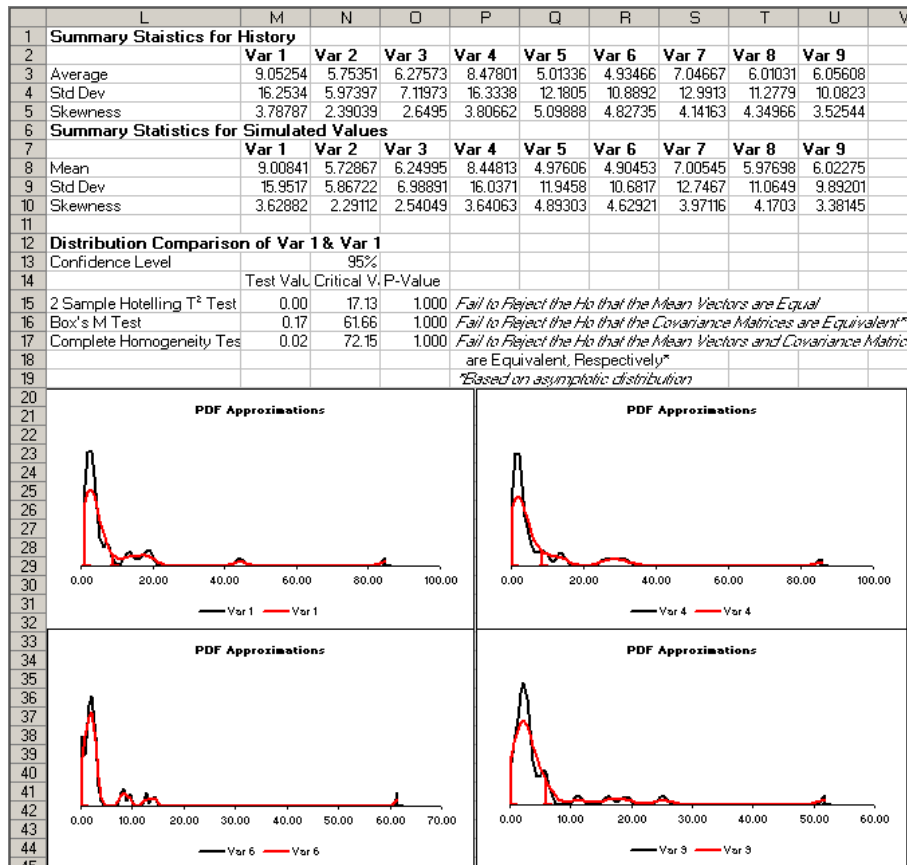


Figure 7.16. Validation Tests and PDFs for a Bootstrapped MV Distribution.

When using bootstrap simulation it is recommended that the number of iterations should be expanded to 1000 or more (Conover). The question remains, how well does this method work? For the example MVE in Figure 7.15, the results from a 1,000 iteration simulation are summarized in Figure 7.16. The means for the simulated and historical series are statistically equal as are the covariance matrices, based on three non-parametric validation tests. The skewness for the distributions are also very similar between the simulated and the historical distributions. The PDFs for four of the random variables are presented to show that the simulated distributions conform closely to the historical distributions. In particular the PDFs show that the multi-modal aspects of the historical distributions are matched by the simulated values.

– Two Stage Correlation for Multivariate Distributions

The second method for simulating a MV distribution defined by a correlation matrix that will not factor is to make two correlation matrices and factor the two matrices. The new correlation matrix is re-estimated after eliminating one or two variables and a second correlation matrix between the deleted variables and one remaining variable is estimated.

To demonstrate the procedure, the 10x10 correlation matrix in Figure 7.14 is used because it will not factor. The 3rd variable in the original data set is removed and a new correlation matrix is estimated (Figure 7.17). The 3rd variable was deleted because of its large number of correlation coefficients outside the ± 0.5 range. (The 5th variable could have been the eliminated variable for the same reason.) The resulting 9x9 correlation matrix factors as demonstrated by the third part of Figure 7.17. The second correlation matrix is a 2x2 between variables 3 and 5 (or any other variable remaining in the 9x9) (see Figure 7.18).

Once the two correlation matrices have been estimated and factored, the CSNDs for the 10 variables are generated in three steps.

1. Use the first correlation matrix and the =CSND (9x9 Correlation Matrix) function to generate an array of nine CSNDs (Step 1 in Figure 7.18).
2. Use the second correlation matrix and the =CSND (2x2 Correlation Matrix, 2x1 ISND Array) function to generate two CSNDs for variables 3 and 5 (Step 2 in Figure 7.18). The two ISNDs in the CSND function are actually one ISND for variable 3 (=NORM()) and the CSND₅ generated for variable 5 in Step 1.
3. Assemble the 10 CSNDs into a final 10x1 array using the CSNDs for variables 1-2 and 4-9 from Step 1 and variable 3's CSND from Step 2. In Figure 7.18 the final array is G96:G105.

The assembled 10x1 CSND array can be used to generate MVN values or converted to CUSDs using =NORMSDIST() and used to generate MVE values.

	A	B	C	D	E	F	G	H	I	J	
52	Drop out the Var 3 Column and re-build the Correlation Matrix										
53		Var 1	Var 2	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Var 10	
54		1	25	22	38	25	31	24	31	32	30
55		2	33	34	35	30	25	22	44	29	33
56		3	23	30	32	30	22	23	30	24	26
57		4	30	38	35	33	32	23	44	32	25
58		5	12	45	45	50	33	36	40	45	35
59		6	37	40	44	43	41	34	32	49	35
60		7	4	23	29	28	18	16	26	40	25
61		8	25	42	40	42	37	33	33	33	26
62		9	5	41	40	45	33	35	42	45	40
63		10	6	27	31	33	31	22	33	54	25
64	Correlation Matrix for 9 Variables without Var 3										
65		Var 1	Var 2	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Var 10	
66	Var 1	1	0.21	0.30	-0.12	0.35	0.09	0.19	-0.47	0.00	
67	Var 2		1	0.76	0.89	0.64	0.84	0.60	0.17	0.52	
68	Var 4			1	0.79	0.80	0.92	0.33	0.26	0.70	
69	Var 5				1	0.66	0.91	0.35	0.50	0.60	
70	Var 6					1	0.79	0.26	0.43	0.40	
71	Var 7						1	0.32	0.34	0.69	
72	Var 8							1	-0.06	0.46	
73	Var 9								1	0.31	
74	Var 10									1	
75	Factor a 9x9 Square, Symmetric Matrix. 12/14/2003 3:52:01 PM										
76		0.000345	0.3424751	0.0763134	0.3545068	0.6721691	0.2108404	0.1024603	-0.494406	0.0033277	
77		0	0.0405048	-0.063409	0.3152329	-0.137327	0.6590514	0.418288	0.0062927	0.5164832	
78		0	0	0.2937199	0.098093	0.1947553	0.6069121	0.0147362	0.0366814	0.7044656	
79		0	0	0	0.1823569	-0.259104	0.6366731	0.1673576	0.3257839	0.6000957	
80		0	0	0	0	0.5225801	0.6628211	0.1638904	0.3198373	0.398023	
81		0	0	0	0	0	0.7101914	0.0308386	0.1284647	0.6915013	
82		0	0	0	0	0	0	0.8622241	-0.210661	0.4606425	
83		0	0	0	0	0	0	0	0.9497636	0.3129682	
84		0	0	0	0	0	0	0	0	1	

Figure 7.17. Correlation Matrix and Its Factored Matrix After Deleting Variable Three.

Before using the CSNDs for generating random numbers, they need to be tested to insure that they are appropriately correlated. To test the correlation select the assembled array of CSNDs as the simulation output variables, simulate the workbook, and test the implicit correlation for the simulated values against the original 10x10 correlation matrix. The results of the correlation test for the problem in Figure 7.18 indicate that all but two of the correlation coefficients in the matrix are statistically correlated as they were over the historical period. If the correlation test indicates that many of the correlation coefficients are statistically different from the original matrix, select a different variable to eliminate and repeat the process. Experience suggests that the variable with the most large correlation coefficients is the best candidate for elimination and for some problems more than one variable must be eliminated and then added back via the second matrix. The variable selected to correlate back to (5 in this example) must be highly correlated to the eliminated variables and to the remaining variables in the correlation matrix. The Bad Correlation Matrix Demo.XLS workbook contains the example described in this section.

	A	B	C	D	E	F	G	H	I
87	Use the correlation coefficient from the original								
88	matrix between Var 3 and 5, to create a new correlation matrix.								
89		Var 3	Var 5						
90	Var 3	1	0.9311	Note that Variable 5 is the independent variable in the matrix					
91	Var 5		1	and Variable 3 depends upon Variable 5.					
92									
93	Step 1. CSNDS for 9 variable								
94	Correlation Matrix				Step 3. Assemble the CSNDS for the				
95		CSND			10 Variables and use the CSNDS.				
96	Var 1	-0.13614	=CSND(B66:J74)		Var 1	-0.13614			
97	Var 2	-0.4416			Var 2	-0.4416			
98	Var 4	-0.01293			Var 3	-0.67602			
99	Var 5	-0.20708			Var 4	-0.01293			
100	Var 6	-0.71311			Var 5	-0.20708			
101	Var 7	-0.18983			Var 6	-0.71311			
102	Var 8	-0.21889			Var 7	-0.18983			
103	Var 9	0.08733			Var 8	-0.21889			
104	Var 10	0.85187			Var 9	0.08733			
105					Var 10	0.85187			
106	Step 2. Calculate CSNDS for Variables 3 and 5								
107	For this step use the CSND generated for Variable 5 in Step 1								
108		CSND for Var 3		CSNDs					
109	Var 3	-1.324787	=NORM()	-0.676024	=CSND(B90:C91,B109:B110)				
110	Var 5	-0.207082	=B99	-0.207082					

Figure 7.18. Steps for Simulating CSNDS from a 9x9 and a 2x2 Correlation Matrix.

Appendix: Simulation of a MVN Distribution

In matrix notation a four variable MVN distribution is simulated using the R matrix, which is the factored correlation matrix. The Choleski decomposition can be used to factor the correlation matrix, i.e., to calculate the square root of the correlation matrix. Appendix Figure 7.1 shows the matrix notation for the steps to simulate a four variable MVN distribution. The ISND vector is an Mx1 vector of independent standard normal deviates which is multiplied by the R matrix to calculate a vector of correlated standard normal deviates (CSND's). Multiplying the MxM standard deviation diagonal matrix by a vector of CSND's and adding the product to the means gives a vector of correlated stochastic values that are distributed MVN. The CSND vector is not presented in Appendix Figure 7.1 as this is an intermediate step, and is shown in Step 8 of the Multivariate Normal Distribution Demo.XLS.

$$\underbrace{\begin{bmatrix} \tilde{X}_1 \\ \tilde{X}_2 \\ \tilde{X}_3 \\ \tilde{X}_4 \end{bmatrix}}_{\text{Stochastic Results}} = \underbrace{\begin{bmatrix} \hat{X}_1 \\ \hat{X}_2 \\ \hat{X}_3 \\ \hat{X}_4 \end{bmatrix}}_{\text{Deterministic Component}} + \underbrace{\begin{bmatrix} \hat{\sigma}_1 & & & \\ & \hat{\sigma}_2 & & \\ & & \hat{\sigma}_3 & \\ & & & \hat{\sigma}_4 \end{bmatrix}}_{\text{Stochastic Component}} * \underbrace{\begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} \\ & r_{22} & r_{23} & r_{24} \\ & & r_{33} & r_{34} \\ & & & r_{44} \end{bmatrix}}_{\text{Correlation Component Results in CSNDs}} * \underbrace{\begin{bmatrix} \text{ISND}_1 \\ \text{ISND}_2 \\ \text{ISND}_3 \\ \text{ISND}_4 \end{bmatrix}}_{\text{ISND Vector}}$$

Appendix Figure 7.1. Simulation of a MVN Distribution.

The MVN distribution simulation depicted in Appendix Figure 7.1 can be written as individual equations for each random variable, much like a univariate normal variable. Recall that a univariate normal variable is simulated as:

$$\tilde{X} = \hat{X} + \sigma_{\tilde{e}} * \text{SND}$$

where SND is an independent standard normal deviate distributed N(0,1).

For a MVN distribution each random variable is simulated as:

$$\tilde{X}_i = \hat{X}_i + \hat{\sigma}_{\tilde{e}_i} * \text{CSND}_i$$

where CSND_i is the ith correlated standard normal deviate in the CSND vector.

Simetar calculates the CSND vector two ways. Both methods are used in Multivariate Normal Distribution Demo.XLS and are described here:

1. The user can specify the vector of ISND's. The 3 variable MVN distribution example in Appendix Figure 7.2 shows that the vector of ISND's is provided as input in column I. The CSND function is an array function so highlight 3 cells and enter the function =CSND (correlation matrix, ISND vector) and press Control Shift Enter.

	E	F	G	H	I	J	K	L
121	CSNDs	Correlation Matrix			ISNDs	Formula in Column E		
122	1.644655	1.0000	0.2335	0.6943	1.977	=CSND(F122:H124,I122:I124)		
123	0.47652	0.0000	1.0000	0.2373	0.421			
124	0.286447	0.0000	0.0000	1.0000	0.286			

Appendix Figure 7.2. Example of CSND Calculated With the ISND Vector Provided as Input.

2. The user does not specify the vector of ISND's. The 3 variable MVN distribution example in Appendix Figure 7.3 shows that the CSND vector is calculated using only the correlation matrix as input. In this case Simetar generates the ISND's without the user explicitly including this step.

	E	F	G	H	I	J	K
113	CSNDs	Correlation Matrix				Formula in Column E	
114	-1.0016	1.0000	0.2335	0.6943		=CSND(F114:H116)	
115	-1.3480	0.0000	1.0000	0.2373			
116	-0.8221	0.0000	0.0000	1.0000			

Appendix Figure 7.3. Example of CSND Vector Calculated Without Explicitly Providing the ISND Vector.

References

- Clements, A.M., Jr., H.P. Mapp, Jr., and V.R. Eidman. "A Procedure for Correlating Events in Farm Firm Simulation Models." Technical Bulletin T-131, Oklahoma Agricultural Experiment Station, August 1971.
- Conover, W. J. Practical Nonparametric Statistics. New York: John Wiley & Sons, Inc. 1999.
- Eidman, V.R. (editor). "Agricultural Production Systems Simulation." Proceedings of a Workshop by the Southern Farm Management Research Committee. Stillwater: Oklahoma State University, May 1971.
- Elfon, B. "Bootstrap Methods: Another Look at the Jackknife." Annals of Statistics 7(1979): 1-26.
- Fackler, P.L. "Modeling Interdependence: An Approach to Simulation and Elicitation." American Journal of Agricultural Economics, 73(1991): 1091-1097.
- King, R.P. "Operational Techniques for Applied Decision Analysis Under Uncertainty." Ph.D. dissertation, Department of Agricultural Economics, Michigan State University, 1979.
- McCarl, B. "Forming Probability Distributions." Department of Agricultural Economics, Texas A&M University, 1988.
- Mjelde, J.W., D.P. Anderson, K. Coble, B. Maufluk, J.L. Outlaw, J.W. Richardson, J.R. Stokes, and V. Sundarapotes. "Tutorial on Density Function Estimation and Use." Department of Agricultural Economics, Texas A&M University, FP-94-2, 1993.
- Richardson, J.W., and G.D. Condra. "A General Procedure for Correlating Events in Simulation Models." Department of Agricultural Economics, Texas A&M University, 1978.
- Taylor, C. Robert. "Two Practical Procedures for Estimating Multivariate Nonnormal Probability Density Functions." American Journal of Agricultural Economics, 72(1990): 210-217.

Chapter 8

Simulating Inter- and Intra- Temporal Multivariate Distributions

The general procedure described in this section is a repeat of some of the material found in previous sections and is intended for more advanced students. The procedure described here is used by FLIPSIM, Farm Assistance and several other large simulation models. The procedure has been developed over the past two decades. Recent variations and improvements in the procedure benefited from my work with Steven Klose and Alan Gray. The three of us wrote an invited paper on the topic for the SAEA Annual Meeting in 2000. The description presented in this section draws heavily on that invited paper.

Some of the special problems facing firm level simulation modelers are:

- non-normally distributed random yields and prices,
- intra-temporal correlation of production across enterprises and fields,
- intra- and inter-temporal correlation of output prices,
- heteroskedasticity of random variables over time due to policy changes,
- numerous enterprises that are affected by weather and carried out over a lengthy growing season,
- government policies that affect the shape of the price distributions, and
- strategic risks associated with technology adoption, competitor responses, and contract negotiations.

The focus of this section is to describe and demonstrate an applied simulation approach for dealing with the first four problems in the list. A portion of the literature in the area of farm level simulation is reviewed prior to describing a generalized procedure for generating appropriately correlated random numbers in firm simulation models.

The relevant phrase is “appropriately correlated” and it means that what ever procedure used to simulate random variables must insure that the historical relationship between all variables is maintained in the simulated variables. This concept can be extended to include coefficient of variation stationarity which means that the relative variability for the random variables must not be changed by the simulation process.

Review of Literature

Agrawal and Heady (1972) provided a cursory treatment of simulation in their operations research book but no details were provided on how to construct a firm level simulation model. Anderson, Dillan and Hardaker (1977) suggested simulation as a tool for analyzing risky decisions but provided no detail for addressing the unique modeling problems listed above. Richardson and Nixon (1986) described the types of equations and identities used to construct the Farm Level Income and Policy Simulation Model (FLIPSIM), but provided a minimum amount of detail on how the random variables were simulated. More recently Hardaker, Huirne, and Anderson (1997) have suggested that simulation can be used as a possible tool for helping farmers cope with risk, but they did not provide details on how to build a farm level simulation model or how to simulate the random variables facing farmers.

Eidman (1971) edited a bulletin on farm level simulation that included a description of the Hutton and Hinman simulation model and various random number generation schemes. Eidman's bulletin became the basic reference material for farm level modelers during the 70s. The General Farm Simulation Model developed by Hutton and Hinman (1971) addressed many of the problems faced by farm level simulators today but did not address the problems of correlating random yields and prices and dealing with heteroskedasticity. Law and Kelton demonstrate that ignoring the correlation of random variables biases the variance for output variables as follows: a model over estimates variance if a negative correlation between enterprises is ignored, and vice versa.

Clements, Mapp, and Eidman (1971) proposed using correlated random yields and prices for firm level simulation models. However, the procedure described by Clements, Mapp, and Eidman for correlating two or more random variables only works if the variables are normally distributed; not the case for yields and prices for most agricultural firms. Richardson and Condra (1978 and 1981) reported a procedure for simulating intra-temporally correlated random prices and yields that are not normally distributed. Working independently, King (1979) reported a similar procedure for correlating multivariate non-normal distributions. King's procedure was included in an insurance evaluation program by King, Black, Benson and Pavkov (1988).¹ Taylor (1990) presented his own procedure for simulating correlated random variables that are not normally distributed.

A procedure for simulating inter-temporally correlated random variables was described by Van Tassel, Richardson, and Conner and demonstrated for simulating monthly meteorological data from non-normal distributions. Their procedure relied on mathematical manipulation of the random deviates to correlate variables from one year to the next and therefore was difficult to expand beyond two or three years for problems involving a large number of random variables.

Simulating Multivariate Non-Normally Distributed Random Variables

Assume we are faced with the analysis of a farm that has four enterprises, corn, soybean, wheat, and sorghum. This means the model will have to simulate eight variables: four yields and four prices. The farm in question only has ten years of yield history (Table 8.1). Therefore, we have an eight variable probability distribution that must be parameterized with only ten observations. To make the problem realistic, assume the model is to be simulated for three years, thus requiring the parameters for a multivariate distribution with 24 random variables.

With the limitation of only having ten observations the use of standardized probability distributions can be ruled out because there are too few observations to prove the data fit a particular distribution. The distribution we recommend in this situation is the empirical distribution defined by the ten available observations.² Assuming the data are distributed empirical avoids enforcing a specific distribution on the variables and does not limit the ability of the model to deal with correlation and heteroskedasticity.

¹ Fackler (1991) reported that the procedure described by King was similar to Li and Hammand's procedure reported in 1975.

² Law and Kelton provide an overview of the F(x) function for an empirical distribution and the inverse transform method of simulating from the F(x) for an empirical distribution.

Parameter Estimation for a MVE Probability Distribution

The first step in estimating the parameters for a multivariate empirical (MVE) distribution is to separate the random and non-random components for each of the stochastic variables. There are two ways to remove the random component of a stochastic variable: (a) use regression (or time series) analysis to identify the systematic variability, or (b) use the mean when there is no systematic variability. Yield is often a function of trend so an ordinary least squares (OLS) regression on trend may identify the deterministic component of a random yield variable. When an OLS regression fails to indicate a statistically significant non-random component, then use the simple mean (\bar{X}) of the data as defined in equations 1.1 or 1.2 and shown in column 3 of Table 8.2.³ (The steps for parameter estimation described here correspond to the steps identified in Complete Correlation Demo.XLS and the rows in the workbook are indicated for each step.)

(1) Non-Random Component of the Historical Values (rows 20-40)

$$(1.1) \hat{X}_{it} = \hat{a} + \hat{b} * \text{Trend}_t + \hat{c} * Z_t$$

or

$$(1.2) \hat{X}_{it} = \bar{X}_i$$

for each random variable X_i and each year t .

The second step for estimating parameters for a MVE distribution is to calculate the random component of each stochastic variable. The random component is simply the residual (\hat{e}) from the predicted or non-random component of the variable (column 4 of Table 8.2). It is this random component of the variable that will be simulated, not the whole variable.

(2) Random Component (rows 43-54)

$$(2.1) \hat{e}_{it} = X_{it} - \hat{X}_{it}$$

for each random variable X_i and each year t .

The third step is to convert the residuals in equation 2.1 (\hat{e}_{it}) to relative deviates about their respective deterministic components. Dividing the \hat{e}_{it} values by their corresponding predicted values in the same period results in fractions that express the relative variability of each observation as a fraction of the predicted values (column 5 in Table 8.2).

(3) Relative Variability of each Observation (Deviates) (rows 57-68)

$$(3.1) D_{it} = \hat{e}_{it} / \hat{X}_{it}$$

for each of the 10 years t and for each random variable X_i .

³ Stochastic prices in a farm level model present a unique problem. The farm receives local prices that are a function of national prices and a wedge or basis. Due to the effect of farm policy on prices, the model must simulate the national prices and then use the wedge to convert stochastic national prices to stochastic local prices. This is particularly important when simulating the affects of policy changes on farms in different regions because all of the farms must be impacted by the same prices.

The fourth step is to sort the relative deviates in equation 3.1 and to create pseudo-minimums and pseudo-maximums for each random variable. The relative deviates, D_{it} , are simply sorted from the minimum deviate to the maximum to define the points on the empirical distribution for each random variable X_i (column 6 in Table 8.2). In a standard empirical distribution the probability of simulating the minimum or maximum of the data is equal to zero (Law and Kelton). In reality these points were each observed in history with a ten percent probability, for a variable with ten years of data. The problem can be corrected by adding two pseudo observations. Pseudo-minimum and pseudo-maximum values are calculated and added to the data resulting in a 12 point empirical probability distribution. The pseudo- minimum and maximums are defined to be very close to the observed minimum and maximum and cause the simulated distribution to return the extreme values with approximately the same frequency they were observed in the past.

(4) Sorted Deviates and Pmin and Pmax (rows 71-84)

$$(4.1) \quad S_{it} = \text{Sorted } [D_{it} \text{ from min to max}] \\ \text{for all years } t \text{ and each random variable } X_i.$$

$$(4.2) \quad P_{\min_i} = \text{Minimum } S_{it} * 1.000001$$

$$(4.3) \quad P_{\max_i} = \text{Maximum } S_{it} * 1.000001$$

The fifth step is to assign probabilities to each of the sorted deviates in equations 4.1-4.3. The probabilities for the end points (Pmin and Pmax) are defined to be 0.0 and 1.0 to insure that the process conforms to the requirements for a probability distribution (column 7 in Table 8.2). Each of the ten observed deviates had an equal chance of being observed ($1/T$) in history so in the simulation process that assumption must be maintained.⁴ The intervals created by the addition of the Pmin and Pmax deviates are assigned one half of the probability assigned to the other intervals. Based on this empirical formulation, outcomes approximating the minimum are realized about 10 percent of the time, and the same for the maximum. Equation 5 illustrates the assigning of probabilities for each of the deviates.

(5) Probabilities of Occurrence for the Deviates (rows 87-100)

$$(5.1) \quad P(P_{\min_i}) = 0.0 \\ (5.2) \quad P(S_{i1}) = (1/T) * 0.5 \\ (5.3) \quad P(S_{i2}) = (1/T) + P(S_{i1}) \\ (5.4) \quad P(S_{i3}) = (1/T) + P(S_{i2}) \\ \dots \\ (5.11) \quad P(S_{i10}) = (1/T) + P(S_{i9}) \\ (5.12) \quad P(P_{\max_i}) = 1.0$$

⁴ However, the flexibility of this procedure allows for assigning any probability between 0 and 1 to the sorted deviates. Thus, elicitation processes can be incorporated to reflect management's/experts opinions about the distributions for each variable.

The sixth step for estimating the parameters for a MVE distribution is to calculate the MxM intra-temporal correlation matrix for the M random variables (Table 8.3).⁵ The intra-temporal correlation matrix is calculated using the unsorted, random components ($\hat{\epsilon}_{it}$) from equation 2.1 and is demonstrated for a 2x2.

(6) Intra-Temporal Correlation Matrix for X_i to X_j (rows 103-112)

$$\rho_{ij} = \begin{bmatrix} \rho(\hat{\epsilon}_{it}, \hat{\epsilon}_{it}) & \rho(\hat{\epsilon}_{it}, \hat{\epsilon}_{jt}) \\ 0 & \rho(\hat{\epsilon}_{jt}, \hat{\epsilon}_{jt}) \end{bmatrix}$$

The seventh step is to calculate the inter-temporal correlation coefficients for the random variables. The inter-temporal correlation coefficients are calculated using the unsorted residuals ($\hat{\epsilon}_{it}$) from equation 2.1 lagged one year, or the correlation of $\hat{\epsilon}_{it}$ to $\hat{\epsilon}_{i,t-1}$ (Table 8.3). The inter-temporal correlation coefficients are used to create a separate matrix for each random variable. The inter-temporal correlation matrices are 3x3 for a three-year simulation problem. A zero in the upper right most cell of the inter-temporal matrix assumes no second order autocorrelation of the variables, a reasonable assumption given only ten observations.

(7) Inter-Temporal Correlation Matrix for Variable X_{it} 's Correlation to $X_{i,t-1}$ (rows 115-124)

$$\rho_{i(t,t-1)} = \begin{bmatrix} 1 & \rho(\hat{\epsilon}_{it}, \hat{\epsilon}_{i,t-1}) & 0 \\ & 1 & \rho(\hat{\epsilon}_{it}, \hat{\epsilon}_{i,t-1}) \\ & & 1 \end{bmatrix}$$

The seventh step completes the parameter estimation for a MVE distribution. The parameters used for simulation are summarized in equation 8.

(8) $\left[\hat{X}_{ik} , S_{it} , Pmin_i , Pmax_i , P(S_{it}) , \rho_{ij(MxM)} \text{ and } \rho_{i(t,t-1)(KxK)} \right]$

for random variables X_i , $i = 1, 2, 3, \dots, M$,

historical years $t = 1, 2, 3, \dots, T$,

and simulated years $k = 1, 2, 3, \dots, K$.

The completed MVE probability distribution can be simulated in Excel using Simetar or in any other computer language that generates independent standard normal deviates (i.e., values drawn independently from a normal distribution with a mean of 0.0 and a standard deviation of 1.0). The steps to simulate the MVE are provided next to demonstrate how the parameters are used to simulate a MVE probability distribution.

⁵ When using the data to estimate the correlation coefficients, Fackler (1991, p. 1093) agrees that one should estimate the rank correlation coefficient directly and then calculate the appropriate random values.

Prior to simulation, the square root of the intra-temporal (ρ_{ij}) correlation matrix and each of the inter-temporal ($\rho_{i,t-1}$) correlation matrices must be calculated. The square root procedure for factoring a covariance matrix, described by Clements, Mapp and Eidman, is used to factor the correlation matrices (one intra-temporal and M inter-temporal) and is named MSQRT.⁶

(9) Factored Correlation Matrices (rows 129 - 164 and 167 - 176)

$$(9.1) \quad R_{ij(M \times M)} = \text{MSQRT}(\rho_{ij(M \times M)})$$

$$(9.2) \quad R_{i(t,t-1)(K \times K)} = \text{MSQRT}(\rho_{i(t,t-1)(K \times K)})$$

Simulation of a MVE Probability Distribution

The first step for simulating a MVE distribution is to generate a sample of independent standard normal deviates (ISND). The number of ISNDs generated must equal the number of random variables, in the case of this example 24 ISNDs are needed for eight variables and three years. The best solution to the problem of generating ISNDs is to use Simetar to generate the ISNDs. By using Simetar to generate the ISNDs, one can take advantage of Simetar's ability to manage the iterations and calculate statistics for the model's output variables, while controlling the process to simulate stochastic variables. During the simulation process Simetar fills the ISND vector each iteration with a new sample of random standard normal deviates and Excel calculates the equations for correlating the deviates.⁷

(10) Vector of ISNDs (rows 181-188)

$$\text{ISND}_{i(24 \times 1)} = \text{Norm}()$$

generate 24 ISNDs by repeating the Simetar formula in 24 cells.

The second step for simulating a MVE distribution is to correlate the ISNDs within each year of the simulation period ($k=1,2, \dots k$) by multiplying the factored correlation matrix (R_{ij}) and eight of the values in the ISND vector. The matrix multiplication is repeated once for each year (k) to be simulated, using the same R_{ij} matrix each time but a different set of eight ISNDs. The resulting eight values in each of three vectors are intra-temporally correlated standard normal deviates (CSNDs) (see Richardson and Condra (1978)). For large samples (number of iterations) the correlated standard normal deviates in equation 11 exhibit similar intra-temporal correlation to that observed in the ρ_{ij} correlation matrix in equation 8.

(11) Correlated Standard Normal Deviates for Simulated Years 1-3 (rows 191-225)

$$(11.1) \quad \text{CSND}_{i(8 \times 1)}^{k=1} = R_{ij(8 \times 8)} * \text{ISND}_{i(8 \times 1)} \text{ for the first eight ISND values,}$$

$$(11.2) \quad \text{CSND}_{i(8 \times 1)}^{k=2} = R_{ij(8 \times 8)} * \text{ISND}_{i(8 \times 1)} \text{ for the second eight ISND values, and}$$

$$(11.3) \quad \text{CSND}_{i(8 \times 1)}^{k=3} = R_{ij(8 \times 8)} * \text{ISND}_{i(8 \times 1)} \text{ for the last eight ISND values.}$$

⁶ MSQRT is a function in Simetar to factor a correlation matrix by the square root method.

⁷ While Simetar includes a correlation function, the simulation is much faster if the correlation matrices are factored ahead of time rather than having to refactor them for each iteration.

The third step in simulation is to capture the inter-temporal correlation of the random variables. The values in the three 8x1 vectors of CSNDs (equation 11) are used in a second matrix multiplication to add the inter-temporal correlation to each random variable. Equation 12 is repeated for each of the eight variables and does not significantly diminish the intra-temporal relationship established in equation 11.⁸ A single step approach to correlating random variables that combines equations 11 and 12 into one 24x24 correlation matrix would be superior. However, the problem with a single-step approach is that even for small models the (MTxMT) correlation matrix can be impossible to factor. The two-step correlation process in equations 11 and 12 overcomes that problem and allows for a large number of random variables to be appropriately correlated in a multi-year simulation model.⁹

(12) Adjusted Correlated Standard Normal Deviates for Variable X_i in Simulated Years 1-3 (rows 228-261)

$$(12.1) \begin{bmatrix} ACSND_i^{k=3} \\ ACSND_i^{k=2} \\ ACSND_i^{k=1} \end{bmatrix} = R_{it-1 (3 \times 3)} * \begin{bmatrix} CSND_i^{k=3} \\ CSND_i^{k=2} \\ CSND_i^{k=1} \end{bmatrix}$$

for each of the i random variables.

The fourth step in simulating a MVE distribution is to transform the ACSNDs from equation 12 to uniform deviates. This step is accomplished using Excel's command =NORMSDIST ($ACSND_i$) for each of the 24 values and is demonstrated in Figure 8.1. Most simulation languages contain a similar error function which can be used to integrate the standard normal distribution from minus infinity to the $ACSND_i$. Because the input for the error function ($ACSND$) is appropriately correlated, the output is a vector of correlated deviates distributed uniform zero-one.

⁸ The two step approach is an improvement over Van Tassel, Richardson, and Conner's mathematical manipulation of deviates one year at a time, because it permits a large number of variables to be correlated over 10 or more years.

⁹ The ACSNDs can be used to simulate multivariate normal (MVN) random variables by applying the adjusted correlated deviates as follows:

$$\tilde{X}_{ik} = \bar{X}_{ik} + \sigma_i * ACSND_{ik}$$

for each random variable X_i and

where $\hat{\sigma}_i$ is the standard deviation for X_i .

This procedure for simulating MVN distributions incorporates both inter- and intra- temporal correlation for large scale models with numerous variables and years in the planning horizon. If the model being simulated contains both normal and non-normal distributions, the normal distributions use the above equation and the ACSNDs while the non-normal distributions use equation 15. In this manner the procedure outlined here is capable of appropriately (intra- and inter- temporally) correlating any distribution and any combination of distributions.

(13) Correlated Uniform Deviates (rows 265-290)

$$CUD_{i(24 \times 1)} = NORMSDIST(ACSND_{i(24 \times 1)})$$

The fifth step in simulation is to use the CUD_i s to simulate random deviates for the empirical distribution of each variable X_i . Using the CUD_i along with the respective variable's S_i and $P(S_i)$ one simply interpolates among the S_i values to calculate a random deviate for variable X_i (see Figure 8.2 for a graphical interpretation). In Excel the interpolation can be accomplished using a table lookup function for each random variable X_i , thus calculating 24 fractional deviates.¹⁰ The interpolation process does not affect the correlation implicit in the CUD_i 's so the resulting random deviates are appropriately correlated fractional deviates (or CFD_i).

(14) Interpolation of an Empirical Distribution for Variable X_i Using the CUD_i (rows 294-319)

$$CFD_{ik} = \begin{bmatrix} Pmin & 0.0 \\ S_1 & P(S_1) \\ S_2 & P(S_2) \\ S_3 & P(S_3) \\ S_4 & P(S_4) \\ S_5 & P(S_5) \\ S_6 & P(S_6) \\ S_7 & P(S_7) \\ S_8 & P(S_8) \\ S_9 & P(S_9) \\ S_{10} & P(S_{10}) \\ Pmax & 1.0 \end{bmatrix} \leftarrow CUD_{ik}$$

The sixth step in simulating a MVE distribution is to apply the correlated fractional deviates to their respective projected means and make any needed adjustment for heteroskedasticity. Projected mean yields for years 1-3 can be the historical means or the projected values from the OLS regressions in equation (1). Projected mean prices for years 1-3 can be from the OLS results in equation (1) or from projections by FAPRI or any other macro model that projects national prices. The CFD_i values are fractions of the mean so as the mean changes, the MVE distribution keeps the relative variability or coefficient of variation constant.¹¹ An expansion factor (E_{ik}) is included in equation 15 to allow for managing of the coefficient of variation over time. If the variable is assumed to have the same relative variability over time the E_{ik} factors are 1.0 for all years t , however if the relative risk is assumed to increase ten percent per year the E_{ik} factors are 1.1, 1.2, and 1.3, respectively, for the first three years.

(15) Simulate Random Values in Year k for Variable X_I (rows 322-347)

$$\tilde{X}_{ik} = \hat{X}_{ik} * (1 + CFD_{ik} * E_{ik})$$

¹⁰ Addin software for Excel to simplify the interpolation step is available from the authors.

¹¹ An explanation of coefficient of variation stationarity for the empirical distribution is provided by Richardson (1999, pp. 104-111). The use of heteroscedasticity adjustments to simulate random variables is explained in the same paper (pp. 140-144).

Excel repeats the process described in simulation steps 1-5 automatically as @Risk© simulates each iteration. The resulting random values can be used in the firm level simulation model to simulate receipts and other variables of interest. The process described here to estimate the parameters and simulate a MVE probability distribution is easily expanded to accommodate models with a large number of random variables. It should be noted that as the correlation matrix gets larger it often becomes difficult to factor.

Random variables generated from the MVE distribution described here have the following properties.

- The variables are intra-temporally correlated the same as the historical period.
- The variables are inter-temporally correlated the same as the historical period.
- The variables have the same means, minimums, and maximums as their parent distributions, if the \hat{X}_{ik} values in equation (15) equal their respective historical means and the E_{ik} s equal one. If the \bar{X}_{ik} in equation 15 is not equal to the historical mean the random variable's average will equal the \bar{X}_{ik} and the minimum will be less than the mean by the same percentage as observed in the historical data.
- The random variables are coefficient of variation (CV) stationary over time if the expansion factors (E_{ik}) are equal to 1.0 for all years.
- When the expansion factors (E_{ik}) are not equal to 1.0 the coefficient of variation in any year t equals the historical coefficient of variation (CV_o) times the expansion factor, or $CV_t = CV_o * E_{ik}$.
- The standard deviations for the output variables are less likely to be overstated or understated due to ignoring the correlation among enterprises and across years.
- The distributions for the random variables are similar to their parent distributions in terms of shape.

Once the parameters for the MVE are estimated, the distribution can be used to simulate a variety of assumptions about the predicted means without changing the relative variability for the variables. This feature is particularly useful for analyzing technological changes that assume changes in the mean yields. An added feature is that the MVE procedure allows one to experiment with alternative levels of relative variability in the future, due to policy changes and or new varieties which may have more or less risk.

The steps for parameter estimation and simulation of MVE distributions are robust and perform efficiently for large scale, agricultural economics simulation models. In addition, the procedure is easily adapted to a variety of programming languages and/or software. The MVE procedure is used by FLIPSIM, Farm Assistance, POLYSYS's crops model, and FAPRI's crops model (Richardson and Nixon 1985; Klose and Outlaw; Ray, et. al.; and Adams). Gray (1998) was the first to apply the MVE procedure to a large scale agribusiness simulation model in Excel. Richardson (1999, pp. 184-245) demonstrates the use of the MVE procedure in several agricultural economics oriented simulation models that are programmed in Excel.

Numerical Application of the MVE Distribution

The Excel worksheet used to demonstrate the generalized MVE procedure presented in this section is provided in Complete Correlation Demo.XLS.

A simple farm level simulation example is presented in this section. Ten years of actual yields for a farm growing corn, soybeans, sorghum, and wheat are combined with ten years of national prices to develop an MVE yield and price distribution for a farm (Tables 8.1-8.3). The farm is simulated for three years using stochastic yields and prices to estimate the distribution of total crop receipts for the farm, assuming 100 acres planted to each crop.

The MVE distribution is simulated for three years using historical mean yields and projected national prices for 2000-2002 from the FAPRI November 1999 baseline. For the simulation, it was assumed that the relative variability of yields would be the same in the future as it has been in the past. However, the relative variability of crop prices is assumed to be 40 percent greater in the last year of the historical period. The results of the simulation are summarized in Tables 8.4 and 8.5.

A comparison of the simulated and historical distribution statistics can validate the MVE procedure. The simulated means for each crop's yield in year 1 compare very favorably to the historical means as do the other statistics. The simulated mean national prices are very close to the mean forecasts provided by FAPRI. By separating the non-random component from the random component, the MVE has the flexibility to impose the historical variability on any assumed mean. The simulated mean yields in years 2 and 3 reflect the 2 percent per year increase in the assumed mean yields.

The simulated coefficient of variation (CV) is the same as the historical CV for all yields and the first 2 years of all prices, where the expansion factors were 1.0. Using the percentage deviations as parameter estimates in the MVE forces the CV stationary process, even when the mean changes from year to year. The standard deviation for corn yield increases from 25.93 to 26.83 as the simulated mean rises from 118.61 to 121.02, in year 1 and 2, respectively, thus maintaining a 0.22 CV (Table 8.4). A process that uses a constant standard deviation would generate a declining CV. The price distributions show the CV stationary process between year 1 and 2. However, in year 3 the CV increases by 40 percent reflecting the assumed expansion factors of 1.4 (Table 8.3). Again, the flexibility of this procedure allows one to control the stochastic process in many dimensions.

The results in Table 8.4 indicate that the stochastic procedure does a good job of simulating the given means, historical relative variability, and provides flexibility in controlling the relative variability overtime. However, a significant contribution of this research centers around the multivariate process. Table 8.5 reports the simulated 24x24 correlation matrix for the random variables. The intra-temporal correlation coefficients, in the triangular areas below the outlined blocks, can be compared directly to the intra-temporal correlation matrix in Table 8.3. The bold numbers along the diagonal of each outlined box are the simulated first-order inter-temporal correlation coefficients that can be compared to the input inter-temporal correlation coefficients shown in Table 8.3.

The simulated yields and prices for 500 iterations in the SimData worksheet (Table 8.5) were tested against the original 24x24 correlation matrix using the Simetar function for comparing correlation matrices. The results of this t-test are presented in SimData and show that the inter-temporal correlation coefficients for the simulated data are not statistically different from the historical correlation coefficients, at the 95 percent level. The t-tests for other correlation coefficients in the simulated data reveal that most of the coefficients are statistically equal to their historical counter parts (see SimData).

The most encouraging result is that this procedure can incorporate a complete correlation matrix into the multivariate simulation for a non-normal distribution with limited historical data. With limited data it is often impossible to estimate a non-singular 24x24-input correlation matrix that can be factored. For this reason, among others, using the simple inter-temporal correlation procedures in Chapter 8 or using @Risk© may not work. However, the two-stage procedure described here avoids the singular matrix problem, incorporates first-order inter-temporal correlation, and produces acceptable intra- and inter-temporal correlation for all of the random variables.

To illustrate the importance of capturing the intra and inter-temporal correlation affects, a simulation of the joint distribution of revenue for the example was conducted. Assuming the farm plants 100 acres each of corn, soybeans, wheat, and sorghum, the joint distribution of price times yield was simulated 10,000 iterations.¹² This simulation was repeated for four scenarios with the assumptions of no correlation, only intra-temporal correlation, only inter-temporal correlation, and complete correlation.

The results of simulating the total receipts for a three crop farm are summarized in Table 8.6. The alternative assumptions about the correlation of the random yields and prices have very little impact on the mean of cash receipts (198.72 – 199.43). However, the inclusion/exclusion of correlation changed the minimum from a low of 126.65 with complete correlation to 135.56 with no correlation. The coefficient of variation was actually understated by ignoring the correlation, in this example, because of the positive intra- and inter-temporal correlations of the random variables. The reverse would have been the case if negative correlation among the variables had been prevalent across the correlation matrices.

SUMMARY

An application of the method was conducted using 10 years of actual farm-level historical data for corn, soybeans, wheat, and sorghum. The simulation model was run for 10,000 iterations and the simulated statistics and correlation matrix were compared to the historical input values. Analysis of the simulated statistics showed that the stochastic procedure does a good job of simulating the given means, historical relative variability, and provides flexibility for controlling the stochastic process. Further evaluation of the simulated correlation matrix indicated that the expected signs on the correlation were attained and the order of magnitude for both the intra- and inter-temporal coefficients were consistent.

Finally, an illustration of the impact of including multivariate stochastic processes was conducted using the joint distribution of revenues for an example farm. By including both intra- and inter-temporal correlation coefficients, the spread of the joint PDF increased dramatically. This result suggests that including correlation in stochastic simulation models that deal with analysis of risk management alternatives is critical. The process described in this section allows applied researchers to address risk management analysis using simulation when historical data is limited and not normally distributed.

¹² Effects of the loan deficiency payments were ignored in this analysis to illustrate the impact of multivariate simulation on the ability to more accurately characterize the joint distribution of total revenue before any risk management intervention.

Table 8.1. Historical Yields and Prices for a Representative Farm.

Years	Yields				National Prices			
	Corn	Soybean	Wheat	Sorghum	Corn	Soybean	Wheat	Sorghum
	bu.	bu.	bu.	cwt.	\$/bu.	\$/bu.	\$/bu.	\$/cwt.
1	100	29.0	48.0	45.0	2.540	7.42	3.72	2.27
2	155	38.0	46.0	61.0	2.360	5.69	3.72	2.10
3	165	40.0	48.0	55.0	2.280	5.74	2.61	2.12
4	112	33.0	54.0	75.0	2.370	5.58	3.00	2.25
5	80	28.0	65.0	5.0	2.070	5.56	3.24	1.89
6	109	40.0	52.0	37.0	2.500	6.40	3.26	2.31
7	145	45.0	50.0	25.0	2.260	5.45	3.45	2.13
8	90	26.0	48.0	12.0	3.050	6.76	4.37	2.91
9	117	47.0	72.0	60.0	2.710	7.35	4.30	2.24
10	114	46.0	50.0	59.0	2.600	6.50	3.45	2.34
Summary Statistics								
Mean	118.700	37.200	53.300	43.400	2.474	6.245	3.512	2.256
Std Dev	26.435	7.386	8.050	21.919	0.261	0.712	0.516	0.251
Coef Var	0.223	0.199	0.151	0.505	0.105	0.114	0.147	0.111
Minimum	80.000	26.000	46.000	5.000	2.070	5.450	2.610	1.890
Maximum	165.000	47.000	72.000	75.000	3.050	7.420	4.370	2.910

Table 8.2. Steps for Estimating the Parameters for an Empirical Distribution.

Observation	Random Variable X_{it}	Deterministic Component \hat{X}_t	Stochastic Component $\hat{\epsilon}_t$	Relative Variability D_{it}	Sorted Deviates S_{it}	Probability of Occurrence $P(S_{it})$
Pmin					-0.1370	0.00
1	48.0	53.3	-5.30	-0.0994	-0.1370	0.05
2	46.0	53.3	-7.30	-0.1369	-0.0994	0.15
3	48.0	53.3	-5.30	-0.0994	-0.0994	0.25
4	54.0	53.3	0.70	0.0131	-0.0994	0.35
5	65.0	53.3	11.70	0.2195	-0.0619	0.45
6	52.0	53.3	-1.30	-0.0243	-0.0619	0.55
7	50.0	53.3	-3.30	-0.0619	-0.0244	0.65
8	48.0	53.3	-5.30	-0.0994	0.0131	0.75
9	72.0	53.3	18.70	0.3508	0.2195	0.85
10	50.0	53.3	-3.30	-0.0619	0.3508	0.95
Pmax					0.3508	1.00

Table 8.3. Parameters for a Sample MVE Probability Distribution.

Sorted Fractional Deviates (Dit) and Probability of Occurrence (P(Sit))									
Observation	Yields				National Prices				Probability of Occurrence
	Corn	Soybean	Wheat	Sorghum	Corn	Soybean	Wheat	Sorghum	
Pmin	-0.3260	-0.3011	-0.1370	-0.8848	-0.1633	-0.1273	-0.2568	-0.1622	0.00
1	-0.3260	-0.3011	-0.1370	-0.8848	-0.1633	-0.1273	-0.2568	-0.1622	0.05
2	-0.2418	-0.2473	-0.0994	-0.7235	-0.0865	-0.1097	-0.1458	-0.0691	0.15
3	-0.1575	-0.2204	-0.0994	-0.4240	-0.0784	-0.1065	-0.0774	-0.0603	0.25
4	-0.0817	-0.1129	-0.0994	-0.1475	-0.0461	-0.0889	-0.0718	-0.0559	0.35
5	-0.0564	0.0215	-0.0619	0.0369	-0.0420	-0.0809	-0.0177	-0.0071	0.45
6	-0.0396	0.0753	-0.0619	0.2673	0.0105	0.0248	-0.0177	-0.0027	0.55
7	-0.0143	0.0753	-0.0244	0.3594	0.0267	0.0408	0.0592	0.0062	0.65
8	0.2216	0.2097	0.0131	0.3825	0.0509	0.0825	0.0592	0.0239	0.75
9	0.3058	0.2366	0.2195	0.4055	0.0954	0.1769	0.2244	0.0372	0.85
10	0.3901	0.2634	0.3508	0.7281	0.2328	0.1882	0.2443	0.2899	0.95
Pmax	0.3901	0.2634	0.3508	0.7281	0.2328	0.1882	0.2443	0.2899	1.00

Intra-Temporal Correlation Matrix

	Corn	Soybean	Wheat	Sorghum	Corn	Soybean	Wheat	Sorghum
Corn	1	0.5826	-0.3793	0.4829	-0.3111	-0.3977	-0.3591	-0.3162
Soybean	0	1	0.1621	0.4597	-0.0856	-0.0164	-0.0796	-0.2588
Wheat	0	0	1	-0.0744	-0.0797	0.2014	0.2297	-0.3065
Sorghum	0	0	0	1	-0.0062	0.0726	-0.2146	-0.1445
Corn	0	0	0	0	1	0.7274	0.7489	0.9252
Soybean	0	0	0	0	0	1	0.6584	0.4899
Wheat	0	0	0	0	0	0	1	0.5656
Sorghum	0	0	0	0	0	0	0	1

Inter-Temporal Correlation Coefficients

Corn	0.0536
Soybean	-0.1810
Wheat	-0.1187
Sorghum	0.0132
Corn	0.1531
Soybean	0.1426
Wheat	0.4231
Sorghum	-0.1577

Projected Means for Simulation Period

Years	Yield				Prices			
	Corn	Soybean	Wheat	Sorghum	Corn	Soybean	Wheat	Sorghum
	bu.	bu.	bu.	cwt.	\$/bu.	\$/bu.	\$/bu.	\$/cwt.
2000	118.7	37.2	53.3	43.4	1.960	4.520	2.910	3.232
2001	121.1	37.9	54.4	44.3	2.000	4.710	2.990	3.375
2002	123.5	38.7	55.5	45.2	2.060	4.870	3.090	3.482

Assumed Expansion Factors

Years	Yield				Prices			
	Corn	Soybean	Wheat	Sorghum	Corn	Soybean	Wheat	Sorghum
2000	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2001	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2002	1.0	1.0	1.0	1.0	1.4	1.4	1.4	1.4

Table 8.4. Results of Simulating Yields and Prices for Three Years.

	Yields				National Prices			
	Corn	Soybean	Wheat	Sorghum	Corn	Soybean	Wheat	Sorghum
Year 1								
Mean	118.61	37.17	53.29	43.29	1.96	4.52	2.91	3.23
Std Deviation	25.93	7.29	7.83	21.65	0.20	0.51	0.42	0.34
Coef Var.	0.22	0.20	0.15	0.50	0.10	0.11	0.14	0.11
Minimum	80.00	26.00	46.00	5.00	1.64	3.94	2.16	2.71
Maximum	165.00	47.00	72.00	75.00	2.42	5.37	3.62	4.17
Year 2								
Mean	121.02	37.99	54.30	44.31	2.00	4.71	2.99	3.37
Std Deviation	26.38	7.46	7.96	22.03	0.21	0.53	0.43	0.36
Coef Var.	0.22	0.20	0.15	0.50	0.10	0.11	0.14	0.11
Minimum	81.60	26.52	46.92	5.10	1.67	4.11	2.22	2.83
Maximum	168.30	47.94	73.44	76.50	2.47	5.60	3.72	4.35
Year 3								
Mean	123.55	38.74	55.41	45.14	2.06	4.86	3.09	3.48
Std Deviation	27.20	7.55	8.12	22.60	0.30	0.76	0.62	0.51
Coef Var.	0.22	0.19	0.15	0.50	0.14	0.16	0.20	0.15
Minimum	83.23	27.05	47.86	5.20	1.59	4.00	1.98	2.69
Maximum	171.67	48.90	74.91	78.03	2.73	6.15	4.15	4.90

Table 8.5. Correlation Matrix Calculated from Simulation Results for Yields and Prices Over Three Years.

Yields for Year 3				Prices for Year 3				Yields for Year 2				Prices for Year 2				Yields for Year 1				Prices for Year 1			
Corn	Soybean	Wheat	Sorghum	Corn	Soybean	Wheat	Sorghum	Corn	Soybean	Wheat	Sorghum	Corn	Soybean	Wheat	Sorghum	Corn	Soybean	Wheat	Sorghum	Corn	Soybean	Wheat	Sorghum
1.000	0.519	-0.318	0.440	-0.279	-0.342	-0.297	-0.273	0.045	0.026	-0.011	0.019	-0.012	-0.015	-0.018	-0.011	0.003	-0.002	-0.009	0.005	0.008	-0.001	-0.008	0.013
	1.000	0.133	0.403	-0.064	-0.007	-0.058	-0.224	-0.088	-0.166	-0.024	-0.077	0.002	-0.014	-0.002	0.027	0.001	-0.004	-0.017	0.008	0.009	0.005	-0.007	0.013
		1.000	-0.080	-0.061	0.152	0.158	-0.226	0.060	-0.002	-0.087	0.013	-0.010	-0.042	-0.031	0.008	-0.003	0.004	-0.008	0.013	0.006	0.009	0.004	0.004
			1.000	-0.004	0.068	-0.170	-0.130	-0.005	0.000	0.010	-0.002	0.003	-0.004	0.003	0.001	-0.008	-0.021	-0.007	-0.013	0.002	-0.004	-0.007	0.012
				1.000	0.683	0.684	0.827	-0.054	-0.016	-0.020	-0.010	0.158	0.112	0.101	0.148	0.018	-0.007	-0.007	-0.001	-0.012	-0.013	-0.022	-0.010
					1.000	0.576	0.409	-0.054	-0.003	0.007	0.003	0.102	0.120	0.073	0.080	0.010	-0.003	-0.007	0.015	-0.006	-0.007	-0.015	-0.007
						1.000	0.416	-0.151	-0.019	0.088	-0.093	0.325	0.277	0.396	0.239	-0.002	-0.003	0.001	0.010	-0.008	-0.004	-0.013	-0.008
							1.000	0.038	0.026	0.028	0.017	-0.118	-0.050	-0.067	-0.119	0.016	-0.006	0.000	-0.009	-0.005	-0.011	-0.008	-0.003
								1.000	0.517	-0.306	0.443	-0.298	-0.352	-0.308	-0.285	0.057	0.031	-0.030	0.025	-0.003	-0.013	-0.008	-0.005
									1.000	0.139	0.415	-0.071	-0.002	-0.055	-0.224	-0.091	-0.155	-0.032	-0.059	0.025	0.020	0.017	0.050
										1.000	-0.053	-0.054	0.165	0.164	-0.233	0.033	-0.009	-0.077	0.018	0.009	-0.009	-0.026	0.031
											1.000	-0.014	0.060	-0.194	-0.133	0.015	0.017	-0.016	0.013	0.001	0.003	-0.007	-0.007
												1.000	0.677	0.681	0.831	-0.037	-0.005	-0.019	0.010	0.138	0.092	0.081	0.125
													1.000	0.574	0.403	-0.042	-0.001	0.015	0.024	0.090	0.118	0.067	0.062
														1.000	0.415	-0.143	-0.029	0.079	-0.068	0.308	0.259	0.395	0.226
															1.000	0.046	0.039	0.040	0.013	-0.141	-0.082	-0.100	-0.141
																1.000	0.537	-0.304	0.440	-0.279	-0.346	-0.332	-0.275
																	1.000	0.139	0.419	-0.075	-0.014	-0.076	-0.223
																		1.000	-0.066	-0.089	0.140	0.185	-0.252
																			1.000	0.009	0.066	-0.202	-0.115
																				1.000	0.682	0.715	0.880
																					1.000	0.606	0.435
																						1.000	0.513
																							1.000

Table 8.6. Sum of Present Value of Total Revenue Assuming Alternative Levels of Correlation Among the Random Variables.

	No Correlation	Only Inter-Temporal Correlation	Only Intra-Temporal Correlation	Total Correlation
	----- \$1,000 -----			
Mean	199.41	199.43	198.74	198.72
Minimum	135.56	134.88	126.69	126.65
Maximum	264.33	263.01	279.06	274.61
Coefficient of Variation	0.0883	0.0898	0.1101	0.1110

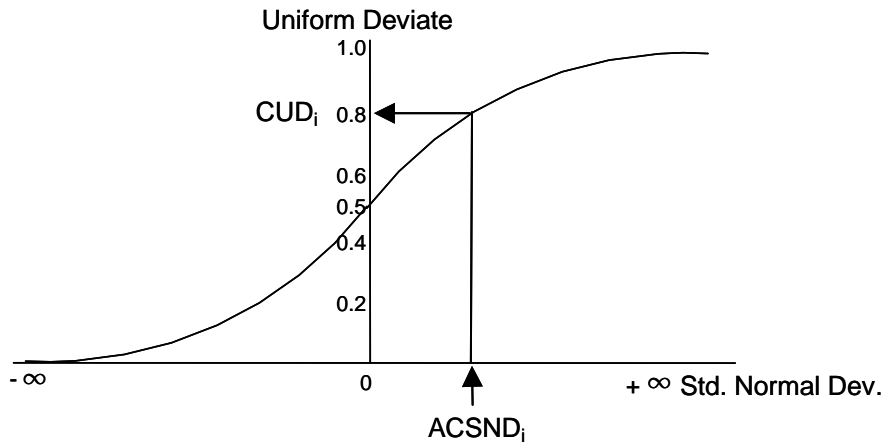


Figure 8.1. Conversion of a Standard Normal Deviate to a Uniform Random Number.

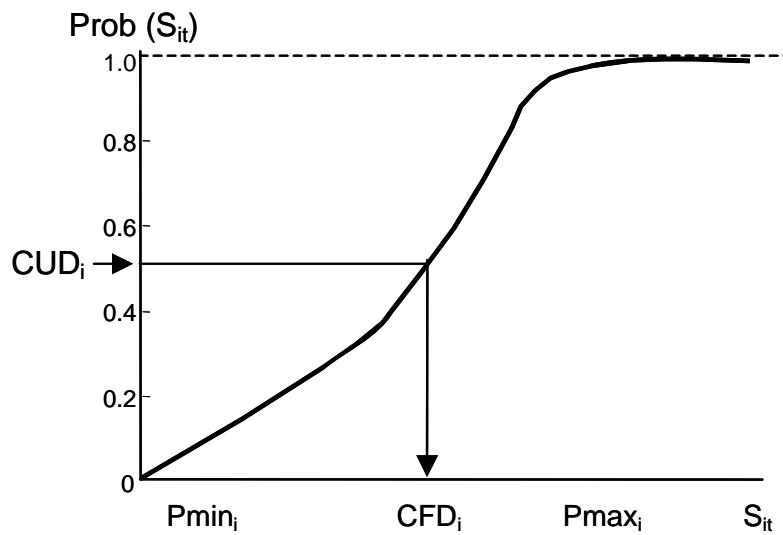


Figure 8.2. Illustration of the Inverse Transform to Simulate on Empirical Distribution.

References

- Agrawal, R.C. and E.O. Heady. Operation Research Methods for Agricultural Decisions. Ames: The Iowa State University Press, 1972.
- Anderson, J.R., J.L. Dillan and J.B. Hardaker. Agricultural Decision Analysis. Ames: The Iowa State University Press, 1977.
- Clements, A.M., Jr., H.P. Mapp, Jr., and V.R. Eidman. "A Procedure for Correlating Events in Farm Firm Simulation Models." Technical Bulletin T-131, Oklahoma Agricultural Experiment Station, August 1971.
- Eidman, V.R. (editor). "Agricultural Production Systems Simulation." Proceedings of a workshop by the Southern Farm Management Research Committee. Stillwater: Oklahoma State University, May 1971.
- Fackler, P.L. "Modeling Interdependence: An Approach to Simulation and Elicitation." American Journal of Agricultural Economics, 73(1991): 1091-1097.
- Food and Agricultural Policy Research Institute (FAPRI). A November 1999 Baseline. University of Missouri-Columbia, November 1999.
- Gray, A.W. "Agribusiness Strategic Planning Under Risk." Department of Agricultural Economics, Texas A&M University, Ph.D. Dissertation, August 1998.
- Hardaker, J.B., R.B.M. Huirne, and J.R. Anderson. Coping With Risk in Agriculture. New York: CAB International, 1997.
- King, R.P. "Operational Techniques for Applied Decision Analysis Under Uncertainty." Ph.D. dissertation, Department of Agricultural Economics, Michigan State University, 1979.
- King, R.P., J.R. Black, F.J. Benson, and P.A. Pavkov. "The Agricultural Risk Management Simulator Microcomputer Program." Southern Journal of Agricultural Economics, 20(Dec. 1988): 165-171.
- Klose, S.L. "A Decision Support for Agricultural Producers." Texas A&M University, Department of Agricultural Economics, Ph.D. Dissertation, 2001.
- Law, A.M. and W.D. Kelton. Simulation Modeling and Analysis. Second edition, New York: McGraw-Hill Book Co., 1991.
- Li, S.T. and J.L. Hammond. "Generation of Pseudorandom Numbers with Specified Univariate Distributions and Correlation Coefficients." IEEE Transactions on Systems, Management and Cybernetics, Sept. 1975: 557-561.

- Ray, D.E., J.W. Richardson, D.G. De La Torre Ugarte, and K.H. Tiller. "A Estimating Price Variability in Agriculture: Implications for Decision Makers." Journal of Agricultural and Applied Economics, 30,1(July 1998): 21-33.
- Richardson, J.W., S.L. Klose, and A.W. Gray. "An Applied Procedure for Estimating and Simulating Multivariate Empirical (MVE) Probability Distributions in Farm Level Risk Assessment and Policy Analysis." Journal of Agricultural and Applied Economics, 32(2000): 299-315.
- Richardson, J.W. and C.J. Nixon. AUAP-Agribusiness Financial Analyzer. College Station, Texas, 1999a.
- Richardson, J.W. and C.J. Nixon. AUAP-Farm Financial Analyzer. College Station, Texas, 1999b.
- Richardson, J.W. and C.J. Nixon. "A Description of FLIPSIM V: A General Firm Level Policy Simulation Model." Bulletin 1528, Texas Agricultural Experiment Station, July 1986.
- Richardson, J.W. and G.D. Condra. "A Farm Size Evaluation in the El Paso Valley: A Survival/Success Approach." American Journal of Agricultural Economics, 63(1981): 430-437.
- Richardson, J.W. and G.D. Condra. "A General Procedure for Correlating Events in Simulation Models." Department of Agricultural Economics, Texas A&M University, 1978.
- Taylor, C.R. "Two Practical Procedures for Estimating Multivariate Nonnormal Probability Density Functions." American Journal Agricultural Economics. 72(1990): 210-17.
- Van Tassell, L.W., J.W. Richardson and J.R. Conner. "A Empirical Distributions and Production Analysis: A Documentation Using Meteorological Data." The University of Tennessee Agricultural Experiment Station, Bulletin 671, September 1989.
- Winston, W.L. Simulation Modeling Using @Risk©. New York: Duxbury Press, 1996.

Chapter 9

Coefficient of Variation Stationarity and Simulating Heteroskedastic Error Terms

When validating a simulation model the developer must verify that the simulated values accurately reproduce the historical deterministic and stochastic components for each random variable, i.e., the simulated values should have the same means and standard deviations as the historical data. When the mean for the planning horizon is assumed to change over time the coefficient of variation (CV) will not be the same as observed over history. As a result the model will not reflect the same relative risk that was observed for the past. By over- or understating the relative risk in the out years, the model will not provide a reasonable approximation of the risk facing the system being studied.

Relative risk of the random variables, as defined by the coefficient of variation (CV), should remain constant over the planning horizon. When probability distribution miss-specification occurs the CV increases or decreases over the planning horizon, thus changing the risk relative to what was observed in the original data. CV stationarity should be checked in the model validation/verification phase. The first two sections of this chapter deal with CV stationarity for the normal distribution and the empirical distribution.

A related topic is that of purposely simulating a random variable so the CV will increase or decrease over the planning horizon. Accommodating this type of problem could occur when simulating a random variable which is assumed to have more or less relative variability in the future than over the historical period. A procedure for simulating random variables with more or less relative variability than was observed for the historical period is described in the third section of this chapter.

CV Stationarity for the Normal Distribution

The normal distribution is easy to use but one of its faults is that the relative variability (CV) changes as the mean changes. This is particularly a problem when the mean of a random variable, such as prices, increases or decreases over the planning horizon. When simulating a normally distributed random variable with an increasing mean and a fixed standard deviation, the CV decreases by definition; thus increasing the relative confidence in the projected values as the planning horizon lengthens. To demonstrate this problem observe the (CVs) for a probability distribution when the mean increases from 3.4 to 4.0 over a 5 year period in Table 9.1.

Table 9.1. Coefficient of Variation for a Distribution with Increasing Means.

	2001	2002	2003	2004	2005
Mean _t	3.40	3.50	3.70	3.80	4.00
Std. Dev.	0.34	0.34	0.34	0.34	0.34
C.V.	10.0%	9.7%	9.1%	8.9%	8.5%

Using the projected mean prices in the example and simulating the variable with a constant standard deviation will actually reduce the relative price risk (CV) five years in the future! This is just the opposite of conventional wisdom regarding future price risk -- risk usually increases over time as we are less and less certain of our forecasts.

If the mean price decreases over the simulation period, the CV for simulated prices actually increases as the standard deviation remains constant. Thus the normal distribution generates non-CV stationary simulated results when the standard deviation is held constant and the mean changes.

A simple adjustment to the standard deviation used to simulate the price for each year will correct the non-stationarity relative risk problem. The adjustment or correction is made by multiplying the standard deviation for each year, t , by the “J-Factor” defined as:

$$J_t = \frac{\bar{X}_t}{\bar{X}_h}$$

where \bar{X}_h is the historical mean of the distribution used to calculate the standard deviation, and \bar{X}_t is the mean for each year $t = 1, 2, 3, \dots T$ in the planning horizon.

Using the example in Table 9.1, the J-Factors are calculated in Table 9.2, assuming the historical mean is 3.4. The standard deviations used for simulation after correction for the J-factor are in the fourth line of Table 9.2. The coefficient of variation equals 10 percent in each year, once the standard deviation is adjusted annually by the appropriate J-Factor. The five normal distributions in Table 9.1 are simulated using the Mean_t and the corrected standard deviations for each year in CV Stationarity Normal Demo.XLS. The Simetar commands used to simulate the five random variables are:

= NORM (3.4, (0.34 * 1.0))
 = NORM (3.5, (0.34 * 1.0294))
 = NORM (3.7, (0.34 * 1.088))
 = NORM (3.8, (0.34 * 1.1176))
 = NORM (4.0, (0.34 * 1.1764))

Table 9.2. Corrected Standard Deviations to Make a Normal Distribution CV Stationary When Means Change Over Time.

	2001	2002	2003	2004	2005
Mean _t	3.40	3.50	3.70	3.80	4.00
Std. Dev.	0.34	0.34	0.34	0.34	0.34
J-Factor	1.00	1.0294	1.088	1.1176	1.1764
Corrected Std. Dev.	0.34	0.35	0.37	0.38	0.40
C.V.	10.0%	10.0%	10.0%	10.0%	10.0%

The CV Stationarity Normal Demo.XLS workbook demonstrates how a CV non-stationarity result differs from one that is CV stationary, assuming the random variable is normally distributed. A five year simulation of the problem presented in Table 9.2 is presented in CV Stationarity Normal Demo.XLS. The spreadsheet equations demonstrate how to calculate the J_t Factors and how to apply them in the Simetar. The Simetar output variables are in cells B30-B34 and B50-B54. Results for simulating the model are summarized in rows 61-65. Note the calculated coefficient of variation values for both the corrected and non-corrected variables. The CVs for the corrected variables are approximately equal to the historical value (10.0) in all years.

CV Stationarity and the Empirical Distribution

An empirical distribution that is expressed in terms of the actual data (Table 9.3) will suffer from a non-stationary CV if the mean changes over the simulation period. Use random variable X for this example.

Table 9.3. Empirical Probability Distribution Expressed as Actual Numbers.

	X_i	$P(X_i)$
Pmin	2.09	0.00
	2.10	0.10
	4.20	0.30
	7.30	0.50
	11.20	0.70
	14.10	0.90
Pmax	14.11	1.00
Historical mean	7.78	

If the distribution in Table 9.3 is simulated using the empirical formula in Simetar for multiple years with an increasing mean, the CV will actually decrease over time. The recommended correction is to re-specify the distribution in terms of fractions of the mean.

Expressing the cumulative distribution values as fractions of the mean and then simulating the distribution as fractions automatically corrects for CV stationarity when means are assumed to increase or decrease over the planning horizon. The simple MVE distribution in Table 9.3 is converted to a CV stationary form in Table 9.4 to demonstrate the concept.

Table 9.4. Empirical Distribution Corrected to be CV Stationary.

	X_t	\bar{X}_t	\hat{e}_{xt}	\hat{e}_{xt} / \bar{X}_t	S_i	$P(S_i)$
					Pmin	0.0
1	4.2	7.78	-3.58	-0.460154	-0.73015	0.1
2	2.1	7.78	-5.68	-0.730077	-0.460154	0.3
3	14.1	7.78	6.32	0.812339	-0.061697	0.5
4	11.2	7.78	3.42	0.439589	0.439589	0.7
5	7.3	7.78	-0.48	-0.061697	0.812339	0.9
					Pmax	1.0

Simulate the corrected empirical distribution for any mean value of \bar{X}_t using the Simetar simulation formula:

$$\tilde{X}_t = \bar{X}_t + [\bar{X}_t * \text{Empirical}(S_i, P(S_i))]$$

Proof that CV stationary holds for CV Stationary empirical distributions, S_i is provided by calculating the statistics for the sorted fractional deviates, as follows:

<u>Sorted X_t</u>	<u>Sorted Deviates S_i</u>	<u>$P(S_i)$</u>
	Pmin S_i	0.0
2.1	-.73015	.1
4.2	-.730077	.3
7.3	-.460154	.5
11.2	-.061697	.7
14.1	.439589	.9
	Pmax S_i	1.0

$$\bar{X} = 7.78$$

$$\hat{\sigma} = 4.931$$

$$CV = .6317$$

$$\bar{X} = 0.0$$

$$\hat{\sigma} = 0.6317$$

$$CV = \text{undefined}$$

Simulate random variable X using the formula:

$$\tilde{X} = \bar{X} * (1 + \tilde{S})$$

$$\text{where } \tilde{S} = \text{Empirical}(S_i, P(S_i))$$

- The statistics for X_t and S_i above show that the expected value of \tilde{S} is zero and the expected standard deviation of \tilde{S} is 63.17% of the mean because of the way \tilde{X} is simulated in the equation.
- Therefore if \bar{X} is 7.78 and \tilde{S} is 63.17% of \bar{X} the standard deviation is 4.91 and the CV equals its historical value of 0.6317. Changing the mean to 10.0, the simulated \tilde{S} will be 6.317 or 63.17% of the mean. The CV remains constant regardless of the mean because the CV formula can be re-written as:

$$\hat{\sigma} = \bar{X} * CV$$

This formula to calculate $\hat{\sigma}$ demonstrates that the $\hat{\sigma}$ will change in proportion to \bar{X} so the CV always remains constant across any \bar{X} values.

Another way to correct an empirical distribution so it is CV stationary is to do the correction during the simulation process. This procedure allows one to use a raw (or historical) data empirical distribution as follows:

$$\tilde{X} = \bar{X}_t * [\text{Empirical}(X_i, P(X_i)) / \bar{X}_h]$$

where \bar{X}_t is the assumed mean for year t in the planning horizon,

X_i is the sorted historical observations for the random variable,

$P(X_i)$ is the probability of the X_i values, and

\bar{X}_h is the mean of the raw or historical data.

During the simulation process Simetar generates an empirically distributed random \tilde{X}_i value and then Excel transforms the value to a fraction of the historical mean prior to multiplying it by the mean for year t. This is actually the same process outlined in Table 9.4, but it is done here during the simulation process rather than ahead of time. For large models with numerous random variables it is recommended that the procedure in Table 9.4 be used to speed up the model by not doing the transformation for each iteration.

Both procedures for simulating CV stationary empirical distributions are presented in CV Stationarity Empirical Demo.XLS.

Controlling Heteroskedasticity for Simulation

Heteroskedasticity occurs when the risk (or variability) of a random variable changes (increases or decreases) over time. It has been suggested that yields for corn and wheat are more variable over the past few years than the previous 10-20 years (Atwood, Baquet and Watts). Also, research by Ray, et. al. has suggested that changing the farm bill in 1996 will lead to more variable crop prices than experienced over the past 15 years.

Heteroskedastic variability may also be a feature that the decision maker wants to impose on the model. Consider the case where the decision maker wants to analyze the effects of switching to a new bio-tech enhanced seed which has 25 percent less yield risk, half way through the planning horizon. Imposing heteroskedastic risk on prices could be considered if the decision maker is more certain about projections in the first two years than in years 3-10.

The recommended procedure for controlling the relative variability of a random variable is to first convert the distribution to be CV stationary and then use the appropriate Expansion Factor to increase or decrease the CV. For the normal distribution, this involves using the J-Factor to correct the standard deviation first and then using the appropriate Expansion Factor E_i . The Expansion Factor E_i is one plus the fractional change assumed for the CV for a particular year. For example, if the CV is to be the same as the historical period for years 1 and 2 but increase 50 percent each year for years three and four, the E_i factors are: 1.0, 1.0, 1.5, and 1.5.

The simulation of a normally distributed random variable for four years, assuming different means and CVs each year is used to demonstrate how to control the CV over time. Let the random variable have a historical mean of 10 and standard deviation of 3, or $X \sim N(10,3)$. Next simulate the variable assuming the mean is 10, 15, 20, and 25, so the equations in the model are:

$$\begin{aligned}\tilde{X}_1 &= 10 + 3 * \text{SND} \\ \tilde{X}_2 &= 15 + 3 * \text{SND} \\ \tilde{X}_3 &= 20 + 3 * \text{SND} \\ \tilde{X}_4 &= 25 + 3 * \text{SND}\end{aligned}$$

Correcting these equations to simulate a normally distributed random variable X so that it is CV stationary requires multiplying the standard deviation by the J_i -Factor where:

$$\begin{aligned}J_i &= \bar{X}_t / \bar{X}_h \\ \tilde{X}_i &= \bar{X}_i + \sigma * \text{SND} * J_i \\ \tilde{X}_1 &= 10 + 3 * \text{SND} * 1.0 \\ \tilde{X}_2 &= 15 + 3 * \text{SND} * 1.5 \\ \tilde{X}_3 &= 20 + 3 * \text{SND} * 2.0 \\ \tilde{X}_4 &= 25 + 3 * \text{SND} * 2.5\end{aligned}$$

To control the degree of heteroskedasticity for the X variable, simply multiply the corrected standard deviation by an Expansion Factor (E_i). The expansion factor is one plus the fractional change in the CV assumed for the simulated values. For example, if the CV is to remain constant at its historical level in years 1 and 2 the E_i is 1.0 and if the CV is to increase 50 percent in years 3 and 4 the E_i is set to 1.5 for these years. Adding the E_i factors to the CV stationary equations above yields:

$$\begin{aligned}\tilde{X}_i &= \bar{X}_i + \sigma * \text{SND} * J_i * E_i \\ \tilde{X}_1 &= 10 + 3 * \text{SND} * 1.0 * 1.0 \\ \tilde{X}_2 &= 15 + 3 * \text{SND} * 1.5 * 1.0 \\ \tilde{X}_3 &= 20 + 3 * \text{SND} * 2.0 * 1.5 \\ \tilde{X}_4 &= 25 + 3 * \text{SND} * 2.5 * 1.5\end{aligned}$$

The Expansion Factor can be calculated as the ratio of the coefficient of variation (CV) for a particular period (years 15-20) and the coefficient of variation for the full period (years 1-20) when simulating a variable with a heteroskedastic historical data series:

$$E_i = CV_{15-20} / CV_{1-20}$$

A warning about using the expansion factor for incorporating heteroskedasticity, is namely: **The equation must be corrected to be CV stationary prior to incorporating the expansion factor.**

The heteroskedasticity adjustment factor can be incorporated into the empirical distribution by multiplying the random fractional deviations by E_i or:

$$\tilde{X}_i = \bar{X}_i * \left\{ 1 + \left[\tilde{S}_i * E_i \right] \right\}$$

where $\tilde{S}_i = \text{EMPIRICAL} (S_i, P(S_i))$

The Excel/Simetar equations for simulating random variables that include an expansion factor are presented here for the normal and empirical distributions.

– Normally distributed random numbers

$$= \text{Norm} \left(\bar{X}_i, (\sigma * J_i * E_i) \right)$$

– Empirically distributed random numbers

$$= \bar{X}_i + \bar{X}_i * \left[\text{EMPIRICAL} (S_i, P(S_i)) * E_i \right]$$

assuming that S_i are sorted fractional deviates and not actual values.

An Excel spreadsheet named Heteroskedasticity Demo.XLS is provided to demonstrate how to simulate random variables that have heteroskedastic relative variability. Three random variables are simulated for five years in two different experiments. In the first experiment the random variables are simulated with the historical relative risk (CV) in years 1-3 and then the relative risk for years 4 and 5 was reduced to by a fraction (56, 18, and 49.9 percent, respectively for the three variables). In the second experiment the relative risk for the random variables is increased by 50 percentage points per year for each year 2-5. Results for both experiments are presented at the bottom of the second page of the Heteroskedasticity Demo.XLS.

References

- Atwood, J.A., A.E. Baquet, and M.J. Watts. "Income Protection." Montana State University, Department of Agricultural Economics and Economics, Staff Paper 97-9, August 1997.
- Hardaker, J.B., Huirne, R.B.M., and J.R. Anderson. Coping with Risk in Agriculture. Wallingford, UK: CAB International, 1997.
- Law, A.M. and W.D. Kelton. Simulation Modeling and Analysis, Third Edition. New York: McGraw-Hill, Inc., 2000.
- Ray, D.E. and J.W. Richardson, D.G. DeLaTorre Ugarte and K.H. Tiller. "Estimating Price Variability in Agriculture: Implications for Decision Makers." Journal of Agricultural and Applied Economics, 30(July 1998): 21-33.

Chapter 10 Simulating Alternative Scenarios and Selecting the “Best” Scenario

The word scenario has been adopted by economists and business analysts to mean an alternative strategy or policy action. A complete analysis of the alternative strategies a business manager or policy maker wants to consider generally consists of simulating multiple scenarios. Each scenario in a simulation analysis is unique because it is based on a different set of assumptions for the exogenous or control variables. Each scenario results in unique distributions for the key output variables.

As a risk analyst, it is not your job to tell the decision maker which scenario to pick. It is, however, your job to educate the decision maker as to the consequences of choosing alternative scenarios or strategies. In simulation and risk analysis this means your role is to help the decision maker choose the best scenario for their situation.

The analyst’s job of ranking alternative risky scenarios in simulation thus becomes one of comparing and ranking the empirical distributions estimated by the simulation model. The purpose of this chapter is to describe, demonstrate, and critique alternative methods that can be used to rank distributions for risky alternatives.

Simulating Multiple Scenarios

The question of how do I generate or simulate multiple scenarios for ranking can be answered with a simple example. The Business Model With Risk Demo.XLS model described in Chapter 3 (Figure 3.2) is restated for a 4 scenario analysis as:

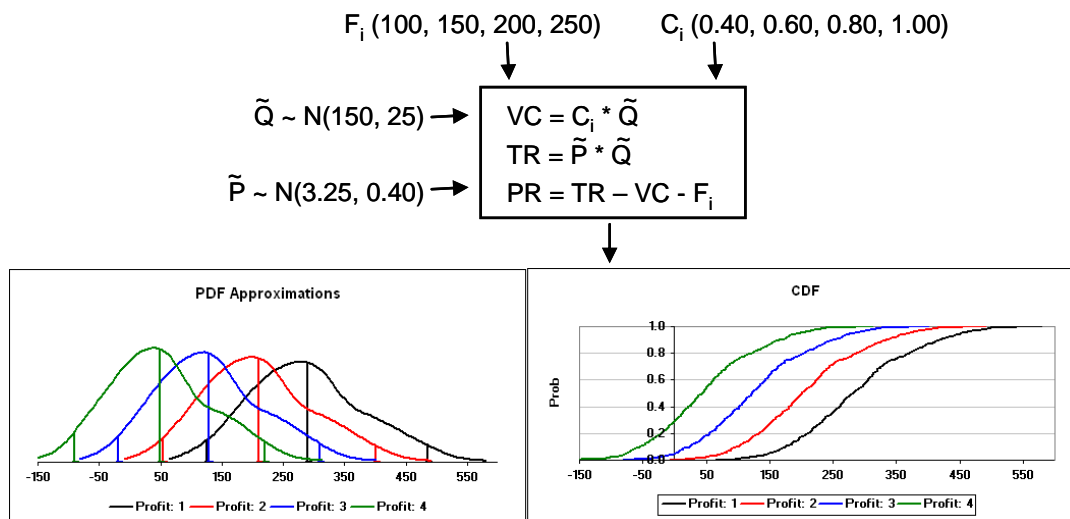


Figure 10.1. Schematic of Business Model With Risk Demo.XLS.

In Chapter 3, the model was simulated for 500 iterations to estimate the parameters for the profit (PR) function assuming that $C_i = 0.90$ and $F_i = 100.0$. The 500 iteration simulation for this pair of C_i and F_i values constituted one scenario for the model and generated one estimate of the pdf for the profit (PR) variable.

The model can also be run for alternative assumptions for C and F, as indicated in Figure 10.1. The model is depicted in Figure 10.1 as having four scenarios based on the values in Table 10.1. Each scenario results in a unique estimate of the pdf for PR that the analyst must compare and help the decision maker rank. For a large scale analysis with N scenarios, the simulation results from Simetar might resemble the information in Table 10.2.

Table 10.1. Scenario Table of Four Scenarios with Two Control Variables, for the Model in Figure 10.1.

Scenario No.	\bar{C}_i	\bar{F}_i
1	0.70	120
2	0.85	100
3	0.90	90
4	1.00	80

Table 10.2. Simulation Results for Simulating N Scenarios.

Scenarios 1 - N		
PR for S_1	PR for S_2 . . .	PR for S_N
\bar{X}	\bar{X}	\bar{X}
S.D.	S.D.	S.D.
C.V.	C.V.	C.V.
Min	Min	Min
Max	Max	Max
PR:1	PR:2	PR:N
1	1	1
2	2	2
3	3	3
.	.	.
.	.	.
.	.	.
500	500	500

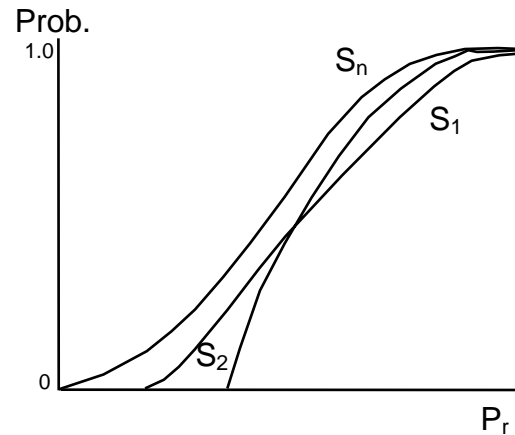


Figure 10.2. CDF's for Simulating a Model for N Scenarios.

The number of scenarios for the Profit Model included in Scenario Analysis Demo.XLS is set at 4 to demonstrate a point, however, in actuality the decision maker may want to compare dozens of scenarios, as depicted in Figure 10.2 and Table 10.2. In addition to changing the control variables C and F, the decision maker may want to see the effects of different parameters for \tilde{Q} and \tilde{P} .

To facilitate this type of scenario analysis the Simetar simulation engine includes a Scenario option. (See Chapter 16 for further details about using the Scenario option in Simetar.) In the simulation dialog box the user may specify the number of scenarios to simulate the model. When the number of scenarios is greater than 1, Simetar reruns the model with the alternative values for the control variables, using the same random values for every random variable, to insure the results between scenarios differ only by the scenario values specified by the analyst. The analyst specifies the alternative control values for C_i and F_i (and all other scenario control variables) using the =SCENARIO() function. An example of the =SCENARIO() function for two control variables and a 4 scenario analysis of the model depicted in Figure 10.1 is:

= SCENARIO (0.75, 0.85, 0.90, 1.00)

= SCENARIO (120, 100, 90, 80)

Instead of typing numbers in the =SCENARIO() function, cell references [e.g., =SCENARIO(A1:A4)] should be entered so the C_i and F_i values can be changed easily. When Simetar simulates the model in this fashion (see the SimData worksheet in Scenario Analysis Demo.XLS) the results are presented in scenario order for each output variable in SimData. In other words, if there are three output variables (Q, P, and PR) for two scenarios the results would be presented as:

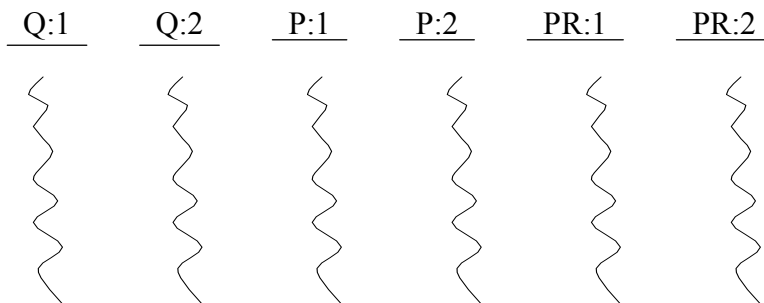


Figure 10.3. Results for Three Output Variables Simulated for Two Scenarios.

Ranking Risky Alternatives

After using a simulation model to simulate multiple scenarios (risky alternatives) the decision maker is faced with the problem of which scenario is best. Much has been written in the economics and business literature about this problem, but it is still a mystery.

Alternative procedures for selecting the “best” strategy are discussed here; but realize the preferred method may change from one situation to the next, depending on the decision maker. The scenario ranking procedures presented here go from the easiest to the most difficult. The problem with using a cookbook procedure for ranking strategies is that you as an analyst cannot make the decision for another person because you do not face their risk/income preferences, you do not have their assets/liabilities, and you do not have their age/life time experiences that go into making a decision. Thus use caution in implementing these scenario comparison procedures. Run each of the procedures and present them to the decision maker so they can make an informed decision.

The Excel spreadsheet, Best Demo.XLS, performs the scenario rankings described in this section for a five scenario problem. The spreadsheet is programmed to rank 5 scenarios that have 50 observations and are entered into columns A-E starting in row 9 (Table 1 of the spreadsheet). All but two of the scenario ranking procedures are completely programmed to operate on the raw data while the remaining ranking procedures require that the analyst type in the rankings after checking the charts and tables in the spreadsheet.

Mean Only

Risky alternatives can be ranked from best to worst based on the means for the key output variable, say NPV. A problem with using the Means Only procedure is that the benefits of stochastic simulation are lost because the risk for each scenario is ignored. This procedure ranks the five scenarios in Table 10.3. The rankings suggest selecting Strategy E, then Strategy B, and so on. This criteria provides a unambiguous ranking and is based on the economic principal that more is preferred to less, regardless of the risk for each scenario. The procedure also assumes the decision maker is risk neutral. (See Table 2 in Best Demo.XLS for this ranking procedure.)

Table 10.3. Rankings of Five Risky Alternatives Based on Means Only Procedure.

<u>Scenario</u>	<u>\bar{X}</u>	<u>Rank</u>
A	19.90	3
B	29.73	2
C	14.86	4
D	9.97	5
E	34.83	1

Standard Deviation

Scenarios are ranked strictly based on their absolute risk, i.e., their standard deviations. The ranking ignores the level of income generated by the alternative scenarios. The five scenarios are ranked in Table 10.4 based on the simulated standard deviations. Using this criteria the decision maker should select scenario D. (See Table 3 of the Best Demo.XLS spreadsheet.)

Table 10.4 Rankings Based on Standard Deviation Procedure.

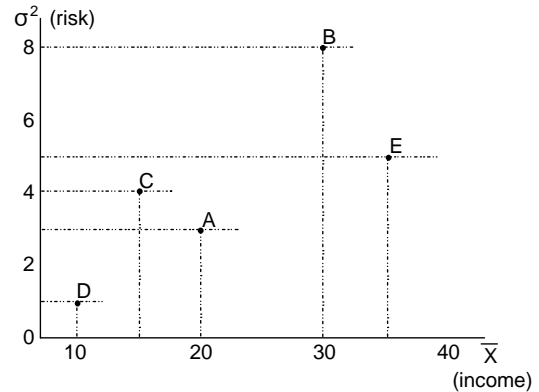
<u>Rank</u>	<u>Scenario</u>	<u>$\hat{\sigma}$</u>
2	A	2.90
5	B	7.74
3	C	3.87
1	D	0.55
4	E	4.84

Mean Variance (MV)

Strategies can be ranked based on how much income they produce relative to the risk associated with earning that income. The measure of risk is the standard deviation (or variance). The mean variance method can be displayed in a graph showing the means and variances (or standard deviations) for all scenarios. In the case of a five scenario decision you can display the results in a format such as Table 10.5.

Table 10.5. Ranking Based on Mean Variance Procedure.

Scenario	\bar{X}	σ^2	Rank by MV
A	19.90	8.42	1
B	29.73	59.89	3
C	14.86	14.97	2
D	9.97	0.29	1
E	34.83	23.39	1



The rule for selecting strategies under the mean variance criteria is to get as far to the right on the \bar{X} (income) axis and as low as possible on the σ^2 (risk) axis. This can be restated as “always select the strategy which has no other strategy in its southeast quadrant.” Strategy A is preferred to C because it provides a higher return for less risk. Strategy E satisfies this criteria as no other strategy is in its southeast quadrant. Strategy B is probably least preferred because it is associated with the largest risk and less income than E. However, strategy D may be preferred over the other three for someone who is very risk averse because it offers the least risk. As you see the MV ranking is dependent upon the decision maker’s preference for the trade off between income and risk. Additionally, mean variance often results in more than one alternative in the efficient set of preferred alternatives, i.e., it results in ties like in Table 10.5.

Minimum and Maximum (or Worst and Best Case)

Several procedures have been suggested that utilize the minimum and maximum values for the key output variable. A distinct disadvantage to these procedures is that they focus the decision on the worst or the best cases and ignore the probabilities of observing these extreme outcomes. When decision makers base their choices on the best or the worst outcomes, then results of a stochastic simulation are largely ignored as these procedures ignore 98 percent, or more, of the iterations. Additionally, these procedures put all of the weight for ranking a strategy on a single iteration, which had a 1 percent (or less) chance of being observed.

– **Minimum Only (Worst Case)**

Rank the strategies based only on their simulated minimums. The strategy with the smallest minimum (C) is least preferred and the one with the largest minimum (E) is the most preferred (Table 10.6). This procedure ignores the average level of income for the scenarios and the dispersion of income (risk) about the mean. Strategy E is preferred over all the rest in our example (Table 10.6). Figure 10.4 depicts the problem with using the worst case method for ranking risky alternatives.

– **Maximum Only (Best Case)**

The Maximum Only Strategy is the reverse of Minimum Only. For the five scenarios being ranked in this Chapter, scenario B is preferred over all the others, because the maximum for B is the largest of the five (Table 10.6). See Figure 10.5 for an example where the Best Case ranking procedure may make the wrong ranking.

– **Mini - Max**

Rank the strategies based on minimizing the chance of a maximum error. Minimize the chance of selecting the wrong strategy (regret) by choosing a strategy that has the smallest minimum range between the mean and the minimum, in our case it is scenario D (Table 10.4). See Figure 10.6 for an example of how this procedure could give an incorrect ranking of two risky alternatives.

Table 10.6. Ranking Based on Minimums, Maximums, and Min-Max.

Scenario	Mean	Minimum	Maximum	Range of Mean-Min	Rank by Minimum	Rank by Maximum	Rank by Mini-Max
A	19.90	14.3	27.7	5.57	3	3	2
B	29.73	14.9	50.5	14.85	2	1	5
C	14.86	7.4	25.3	7.43	5	4	3
D	9.97	9.1	10.9	0.90	4	5	1
E	34.83	25.6	47.8	9.28	1	2	4

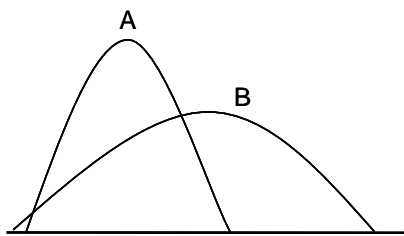


Figure 10.4. Worst Case Scenario Method Ranks A over B.

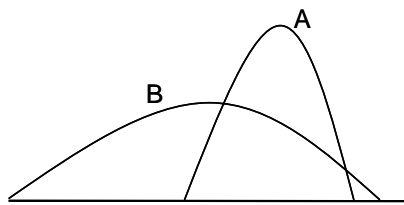


Figure 10.5. Best Case Scenario Method Ranks B over A.

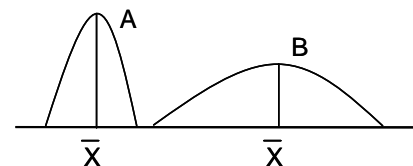


Figure 10.6. Min-Max Ranks A over B to Minimize Distance from Mean to Minimum.

Relative Risk (CV)

The coefficient of variation (CV) is defined here as the absolute ratio of the standard deviation and the mean, or the relative risk associated with a scenario. Ranking risky alternatives based on CV calls for selecting the scenario with the lowest absolute CV. The advantage to this procedure over Mean Only is that it considers the average risk for each scenario. Its advantage over the Mean Variance procedure is that it simplifies the criteria to one value (CV) for each scenario and it eliminates the ambiguity of multiple alternatives in the efficient set. Table 10.7 shows that the strategies receive a unambiguous ranking unless there is a tie between two or more CV's. Break ties by assigning a higher rank to the scenario with the larger mean. The Relative Risk procedure works well if the means of all the alternatives are similar and not close to zero. If the range of the variables includes zero the CV is not a reliable procedure for ranking scenarios. A disadvantage to the Relative Risk procedure is that it ignores the skewness and extreme downside risks associated with some strategies.

Table 10.7. Rankings Based on Absolute Relative Risk.

Rank	Scenario	CV	\bar{X}	$\hat{\sigma}$
3	A	14.6	19.9	2.90
4	B	26.0	29.7	7.74
5	C	26.0	14.9	3.87
1	D	5.4	9.9	0.55
2	E	13.9	34.8	4.84

Probabilities of Target Values

This procedure was suggested by Richardson and Mapp when they demonstrated the use of probabilistic cash flows and the probability of economic success for deciding among risky scenarios. The probability of success was defined as the probability that NPV is positive so the business earns a rate of return greater than the discount rate. Probabilistic cash flows examine the probability that ending cash reserves in each year will be positive or will exceed the decision maker’s reservation level.

Other probabilities can be simulated and calculated for each scenario, such as the probability of remaining solvent, and the probability of increasing real net worth. For the example problem we can use the EDF function in Simetar to calculate probabilities of interest to the decision maker. (See Table 7 in Best Demo.XLS for an example of calculating a Target Value table.) The probability results can be summarized in a table like Table 10.8.

Table 10.8. Rankings Based on Probabilities of Target Values for Rates of Return.

Rate of Return	Alternative Scenarios				
	A	B	C	D	E
			(percent)		
Prob (X > 20%)	49.6	89.5	6.5	0	100
Rank wrt P(X>20%)	3	2	4	5	1
Prob (X>30%)	0	49.6	0	0	83.2
Rank wrt P(X>30%)	3	2	4	5	1
Prob (X>35%)	0	27.6	0	0	49.6
Rank wrt P(X>35%)	2	3	4	5	1

The results in a probability setting such as this communicate to the decision maker the chance of earning a return that is greater than his/her minimum rate of return, say 20 percent. The decision maker will most likely select the strategy which yields the largest probability that the rate of return will exceed his/her minimum rate of return. Strategy D has a 100% chance of a return greater than 20%. Strategy E provides the highest probability (zero) of earning a rate of return greater than 35%.

The =EDF() function in Simetar can be used to calculate the probability that the X_i values will be less than the target (20.0 in this case). Because the maximum value for Scenario E is greater than 20.0 there is a 100 percent chance of exceeding the target. The reverse situation exists for Scenario D.

The StopLight function in Simetar can also be used to develop probability ranking tables (see the StopLight1 worksheet in Best Demo.XLS). The StopLight table summarizes the probabilities that the scenarios will be less than the lower target of 15 (in red) and the probabilities that the risky alternatives will exceed a maximum target of 25 (in green). The probability of each scenario falling between the two targets is reported in the table in amber. StopLight facilitates the ranking of scenarios by providing a table of the probabilities as well as graphically showing the probabilities of each range. The graphical display of probabilities of a risky alternative exceeding an upper target and falling below a lower target have proven a very powerful tool for helping decision makers rank risky alternatives. The target values in Simetar's StopLight worksheet can be changed to accommodate different decision makers and the probabilities will be updated instantly.

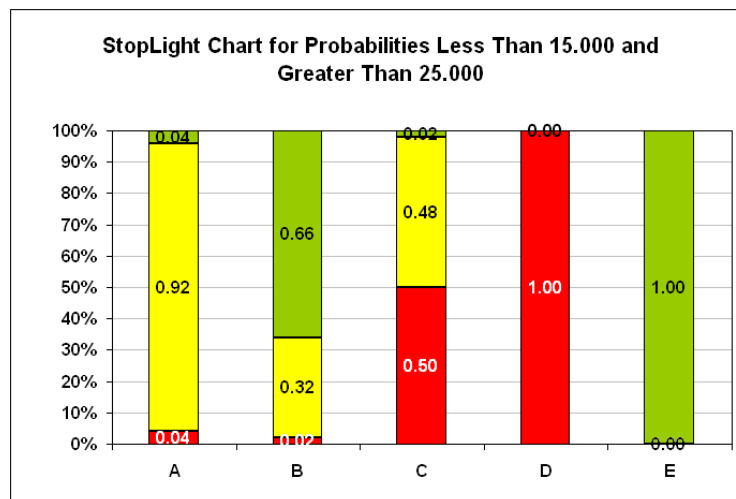


Figure 10.7. Example of a StopLight Chart.

Complete Distribution - CDF Chart

The first seven risk ranking procedures rely on summary statistics for the output variable. A superior procedure is one which utilizes all of the simulated outcomes; in other words, one which considers the full range of possible outcomes rather than just the mean or standard deviation. A graph of all simulated values drawn with a probability scale (0 - 1.0) on the Y axis and the output variable on the X axis (a cumulative distribution function or CDF chart) facilitates full distribution comparison. The CDF chart for the five scenarios are displayed in rows 125-148 of Best Demo.XLS and in Figure 10.8. The CDF graph shows that the E scenario lies more to the right than the other four scenarios. This result suggests that scenario E should be preferred over the others because at each probability level scenario E is associated with higher rates of return (or values for the KOV). Scenario D lies further to the left than the others so it is the least preferred for the same reason. Although the CDF Graph procedure is superior to the first seven strategies it does not always result in an unambiguous ranking of the strategies. Generally, when

the CDF lines cross there is no clear ranking. When this occurs the strategies need to be ranked based on expected utility. The graphical display of the CDFs would rank the strategies as indicated in Table 10.9.

Table 10.9. Ranking Based on CDF Charts.

Rank by CDF Chart	Scenarios
3	A
2	B
4	C
5	D
1	E

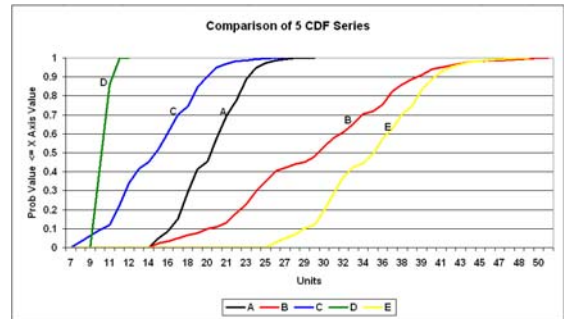


Figure 10.8. Example of Comparing CDFs for Risky Scenarios.

Expected Utility (EU)

Von Neumann and Morgenstern (1944) put forth the idea of using expected utility to rank risky alternatives. Their hypothesis was that individuals maximize their expected utility. Expected utility refers to the analysis of an economic model under the assumption that utility is maximized. Arrow (1965) demonstrated that EU could be used to predict risky portfolio decisions. Pratt (1965) proposed measures of absolute and relative risk aversion similar to Arrow. Hadar and Russell (1969) and Hanock and Levy (1969) added to the EU literature and by the early 70’s EU was an accepted decision analysis tool.

The EU decision analysis paradigm has three components (Meyer). The utility function used to calculate EU depends on a vector of variables or

$$U(\tilde{X}, \alpha)$$

where \tilde{X} is a random variable and α is a choice variable for decision makers.

The utility function is also written as:

$$U(Z)$$

where Z depends on \tilde{X} and α .

So the utility function can be re-written as:

$$U(Z) = U(Z(\tilde{X}, \alpha))$$

For stochastic simulation models Z is wealth or net income for the economic decision and \tilde{X} are the random variables while α is (are) the decision makers control variable(s).

In a simulation modeling context we re-write the utility function as:

$$U(Z_i) = U(Z_i(\tilde{X}, \alpha_i))$$

with α_i represents the alternative scenarios (i) for the control variables. The Z_i variables are thus the empirical probability distributions (CDFs) derived from the stochastic simulation model. Expected utility theory holds that the decision maker will pick the α_i scenario which maximizes expected utility. The $Z(\tilde{X}, \alpha)$ function is often restricted. The outcome variable (KOV) is commonly assumed to be monotonic in the random parameter (\tilde{X}) and concave in the decision variable (α) for all X and α . Meyer suggests that this assumption holds for stochastic rates of return and prices but may not hold for weather.

The restrictions on the utility function are:

$$U'(Z) \geq 0 \text{ and } U''(Z) \leq 0.$$

The first restriction indicates that decision makers prefer more to less, which is consistent with Z being income or wealth. The second assumption is imposed to explain (insure) the commonly observed behavior of risk aversion.

The negative exponential utility function is the most commonly used utility function for EU analysis.

$$U(Z) = 1 - \exp(-r_a Z)$$

where r_a is the absolute risk aversion coefficient (RAC).

Arrow (1965) proposed that the relative risk aversion coefficient is $R(Z) \leq 1$, where Z is wealth. The utility function is expressed in terms of absolute risk aversion which is calculated as $r_a = r_r/Z$ where r_r is Arrow's relative risk aversion coefficient.

Bernoulli proposed an "everyman's" utility function had a r_r of 1. Anderson and Hardaker (1992) proposed a classification of r_r levels about the normal value of 1.0:

- 0.5 hardly risk averse,
- 1.0 normal or somewhat risk averse,
- 2.0 rather risk averse,
- 3.0 very risk averse, and
- 4.0 extremely risk averse.

McCarl and Bessler (1989) proposed the following rules for assigning ranges for r_r values:

- $2 * (\text{coef var}) / \text{std. dev.}$
- $5 / \text{std. dev.}$
- $5.14 / \text{std. dev.}$

- 28 / std. dev.

The remainder of this section discusses several risk ranking procedures based on the expected utility principle.

- stochastic dominance
 - first degree stochastic dominance,
 - second degrees stochastic dominance,
 - generalized stochastic dominance
- confidence premiums,
- certainty equivalents,
- breakeven risk aversion coefficients,
- stochastic efficiency with respect to a function, and
- risk premiums.

– **First Degree Stochastic Dominance (FSD)**

Hadar and Russell (1969) proposed the concept FSD for ranking risky alternatives. The problem is set up in terms of ranking two risky alternatives $F(z)$ and $G(z)$. (In simulation modeling, $F(z)$ and $G(z)$ are CDFs of $Z(\tilde{X}, \alpha)$ for two assumed α 's.) For FSD, $F(z)$ is preferred to $G(z)$ if $[G(z) - F(z)] \geq 0$ for all z in $[a, b]$. In Figure 10.9 the $F(z)$ distribution is always to the right of $G(z)$ at all z_i values which satisfies the FSD condition. All classes of decision makers prefer F to G , whether risk averse, risk neutral, or risk loving.

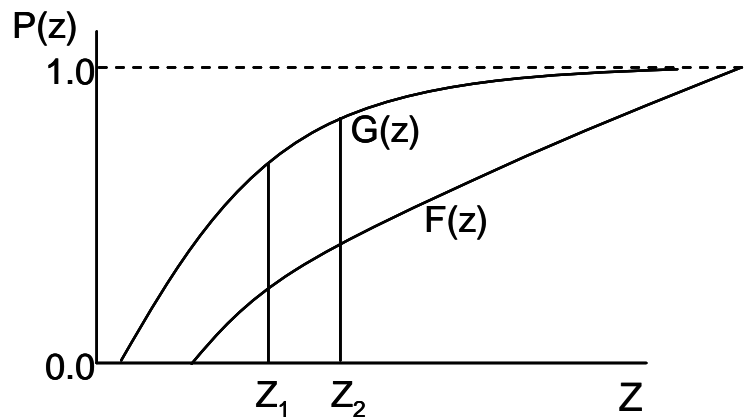


Figure 10.9. FSD Ranking of Risky Alternatives.

– **Second Degree Stochastic Dominance (SSD)**

Hadar and Russell (1969) proposed SSD for ranking risky alternatives for risk averse decision makers. They proposed that $F(z)$ is preferred to $G(z)$ if

$$\int_a^b [G(s) - F(s)] ds \geq 0 \text{ for all } z \text{ in } [a, b].$$

The condition is that $F(s)$ lies to the right more than the $G(s)$ if we sum the differences over all values of s . The SSD ranking can be calculated in Excel by calculating the sum of the differences between the distributions, s , over all iterations for the two CDFs.

$$s = \sum_1^N (G(z) - F(z))$$

If s is positive G is preferred to F , if s is zero G and F are indifferent and if s is negative F is preferred to G .

Simetar calculates FSD and SSD for risky alternatives. The results of the stochastic dominance menu calculations are presented in the StochSum worksheet. An example of using SSD and FSD to rank five risky alternatives is provided in Figure 10.10. The SSD table is read by row. If an alternative for a row is SSD over another alternative, the dominated alternative's name appears in its respective column. The five risky alternatives example for this chapter in Figure 10.10 indicates that E is SSD over A , B , C , and D and E is FSD over A , C , and D .

	A	B	C	D	E	F	G	H	I	J
1	Estimates of First and Second Degree Stochastic Dominance									
2										
3	First Degree Dominance Table									
4			A	B	C	D	E			
5	A FDD:				FDD	FDD				A is FDD over C and D
6	B FDD:		FDD		FDD	FDD				B is FDD over A, C and D
7	C FDD:									
8	D FDD:									
9	E FDD:		FDD		FDD	FDD				E is FDD over A, C, and D
10										
11	Second Degree Dominance Table									
12			A	B	C	D	E			
13	A SDD:				C	D				A is SSD over C and D
14	B SDD:		A		C	D				B is SSD over A, C, and D
15	C SDD:					D				C is SSD over D
16	D SDD:									
17	E SDD:		A	B	C	D				E is SSD over A, B, C, and D

Figure 10.10. Example of SSD and FSD for Ranking Five Risky Alternatives.

– Generalized Stochastic Dominance (SDRF)

Generalized stochastic dominance was introduced by Meyer (1977) and is generally referred to as stochastic dominance with respect to a function (SDRF). Meyer proposed ranking risky alternatives for a class of decision makers, i.e., for decision makers who's utility function is defined by a lower risk aversion coefficient (LRAC or r_1) and an upper risk aversion coefficient (URAC or r_2) which is denoted as $U(r_1(z), r_2(z))$. The condition for F preferred to G under SDRF is :

$$\int_{r_1}^{r_2} u(z) dF(z) \geq \int_{r_1}^{r_2} U(z) dG(z)$$

which is often expressed as:

$$\int_{r_1}^{r_2} [G(z) - F(z)] u^1(z) dz \geq 0$$

The SDRF criteria indicates that utility is calculated for each z value and the sum of the weighted utilities is used to rank F and G. The preferred risky alternative is calculated for the LRAC and for the URAC. If the same risky alternative is preferred for both RACs it is considered to be in the risk efficient set which is generally shortened to the “efficient set.” In the event that the SDRF ranking is different for the LRAC and the URAC, the decision makers with these RACs are said to be indifferent between the two alternatives or the efficient set contains both alternatives. The SDRF criteria is useful for ranking risky alternatives who’s CDFs cross. A limitation of SDRF is that it is a pairwise ranking of risky alternatives not a simultaneous ranking of all alternatives. Another limitation is that if the LRAC and URAC are set to far apart the procedure will not result in a consistent ranking at both RACs and only one alternative in the efficient set. However, the incentive in setting RACs for SDRF is to set them as far apart as possible to include a larger class of decision makers.

Simetar includes SDRF as a tool for ranking risky alternatives. An example of Simetar’s SDRF output is presented in Figure 10.11. In the example, alternatives B and E are in the efficient set because they are ranked 1 and 2 in reverse order for the two RACs. The least preferred alternative is D for both RACs.

The scenario rankings for the upper RAC are presented in the right side of the stochastic dominance output table. The degree of risk aversion is reported at the top of each section of output. The larger the risk aversion coefficient (RAC) the greater the degree of risk aversion and thus the

	A	B	C	D	E	F	G
1	Analysis of Stochastic Dominance with Respect to a Function (SDRF)						
2	© 2006						
3							
4	Efficient Set Based on SDRF at			Efficient Set Based on SDRF at			
5	Lower RAC			Upper RAC			
6	-1			1			
7	Name Level of Preference			Name Level of Preference			
8	1 B	Most Preferred		1 E	Most Preferred		
9	2 E	2nd Most Preferred		2 B	2nd Most Preferred		
10	3 A	3rd Most Preferred		3 A	3rd Most Preferred		
11	4 C	4th Most Preferred		4 C	4th Most Preferred		
12	5 D	Least Preferred		5 D	Least Preferred		
13	*The efficient sets are not the same for both RAC values. This result suggests that the efficient set						
14	changes between the two RACs. Use SERF analysis to determine the RAC(s) where the efficient set						
15	changes.						
16							
17	Summary Statistics						
18	Name	Mean	Std Dev	Coef Var	Skewness	Minimum	
19	1 A	19.90	2.90	14.59	0.26	14.33	
20	2 B	29.73	7.74	26.03	0.26	14.87	
21	3 C	14.86	3.87	26.03	0.26	7.44	
22	4 D	9.97	0.55	5.47	0.08	9.08	
23	5 E	34.83	4.84	13.89	0.26	25.55	
24							
25	Certainty Equivalents Under Alternative Utility Functions						
26	Name	CE Under Exponential Utility			CE Under Power Utility		
27		Lower	Upper	Lower	Upper		
28	A	24.00	17.12	23.34	16.85		
29	B	46.63	18.51	44.14	17.17		
30	C	21.43	10.55	19.81	9.75		
31	D	10.12	9.83	10.11	9.83		
32	E	43.95	28.86	42.89	28.37		
33							
34	Confidence Premiums Between Probability Distributions						
35		Lower Bound			Upper Bound		
36	Dominant Series	Absolute Amount	Percent of Dominant Mean	Absolute Amount	Percent of Dominant Mean		
37							
38	1 B	The Most Dominant Distribution					
39	2 E	2.68	9.02	-10.36	-34.84		
40	3 A	22.63	76.13	1.38	4.66		
41	4 C	25.20	84.78	7.96	26.77		
42	5 D	36.51	122.83	8.67	29.18		

Figure 10.11. Example of SDRF to Rank Five Alternatives.

more weight stochastic dominance places on the risk between scenarios when preparing the ranking. Experiment with alternative RAC levels by changing the value (in Cell C5 or G5) and observing the results in the StochSum1 worksheet in Best Demo.XLS.

– **Confidence Premiums (CP)**

Mjelde and Cochran (1988) took advantage of constant absolute risk aversion properties to extend the stochastic dominance literature by introducing confidence premiums. If SDRF results imply that $F(x)$ is preferred to $G(x)$ because the expected utility for F is greater than for G , Mjelde and Cochran proposed subtracting a constant value, π , from each value of $F(x)$ until the decision maker is indifferent between F and G at the LRAC or: expected utility for $F(x - \pi) =$ expected utility for $G(x)$. The value of π where indifference occurs is the lower confidence premium and indicates the minimum amount a decision maker would have to be paid to switch from the preferred strategy (F) to the inferior strategy (G). The maximum premium the decision maker places on F relative to G is found by evaluating $F(x - \pi) = G(x)$ using the URAC. Mjelde and Cochran's confidence premiums are interpreted as follows:

- both premiums are positive occurs when F is initially preferred to G at both the LRAC and the URAC,
- negative lower bound premium and a positive upper bound premium occurs when G is initially preferred at the LRAC and F is initially preferred at the URAC. (In this case increase the LRAC or lower the URAC until both RACs yield the same ranking.)
- positive lower bound premium and a negative upper bound premium occurs when F is initially preferred at the LRAC and G is preferred at the URAC, and
- both premiums are negative implies unreliable results so change the RACs.

The lower and upper bound confidence premiums are provided for all pair-wise combinations of the risky strategies in Simetar's stochastic dominance output table (Figure 10.11). Changing the RACs causes Excel to recalculate the confidence premiums as it updates the rankings. Mjelde and Cochran suggested using the premiums to value information used in a risky decision. In keeping with this type of application, you can use the premiums to indicate how much a risk averse decision maker values $F(x)$ over $G(x)$ or how much to pay decision makers to get them to switch from the preferred strategy. If the confidence premium is small relative to the mean value for $F(x)$, then the stochastic dominance ranking (preference) is not strongly held or not very important for the type of decision maker represented by the RAC's. The confidence premiums will change (increase or decrease) as you change the RACs so use caution in setting these values.

– **Certainty Equivalence (CE)**

Hardaker (2000) proposed using certainty equivalence to rank risky alternatives. The basic principle of ranking with CE is the same as ranking with SDRF, more is preferred to less. Hardaker proposed that the expected utility of any risky alternative can be expressed through the inverse utility function as a CE. Freund (1956) defined the CE for a risky alternative as:

$$CE = \bar{Z} - 0.5 r_a V$$

where \bar{Z} is expected income or wealth, r_a is absolute risk aversion, and V is the variance of income or wealth. The CE is calculated for the negative exponential utility function and for the power utility function by Simetar in the SDRF worksheet (Figure 10.11). Based on CE rankings for the risky alternatives in the example would put B and E in the efficient set for the exponential utility function and E in the efficient set for the power utility function.

The stochastic dominance ranking for the five scenarios is presented in the SDRF1 worksheet of Best Demo.XLS and in Table 10.10 assuming a LRAC of zero (risk neutral) and a URAC of 1.0 or risk averse. The stochastic dominance results were generated using the Stochastic Dominance option in Simetar. Simetar’s results of the stochastic dominance analysis are presented in four sections: Lower Risk Aversion ranking, Upper Risk aversion ranking, confidence premiums, and certainty equivalents (Figure 10.11). The scenario ranking for a decision maker with a risk aversion coefficient (RAC) equal to the lower RAC is provided in the left side of the Simetar table of stochastic dominance results (Figure 10.11).

Table 10.10. Rankings of Five Alternative Scenarios Based on Stochastic Dominance.

Scenarios	LRAC = -1 Risk Neutral	URAC = +1 Risk Averse	Confidence Premium B to Alter.	Exp. Utility Certainty Equivalence
A	3	3	22.63 – 1.38	24.00 – 17.12
B	1	2	--- – ---	46.63 – 18.51
C	C	4	25.20 – 7.63	21.43 – 10.55
D	5	5	36.51 – 8.67	10.12 – 9.83
E	2	1	2.68 – -10.36	43.95 – 28.86

The confidence premiums in the stochastic dominance output table (Figure 10.10) indicate the relative conviction that the decision maker has to a particular scenario ranking. In other words a confidence premium of 2.68 to -10.36 between scenario B (the dominate one) and E (the second ranked one) is how much the decision maker would have to be paid to accept E over B (Table 10.10 and Figure 10.11). The value of 2.68 may not seem very large but it amounts to 9.02 percent of the mean for scenario B. As we move down the table of confidence premiums we find that scenario B is even more highly valued relative to scenario C and D.

Break Even Risk Aversion Coefficients (BRACs)

McCarl (1988) proposed a procedure for ranking risky alternatives by finding the RAC where the decision maker would be indifferent between two risky alternatives, i.e., the BRAC. Like SDRF, McCarl’s procedure is a pairwise ranking of risky alternatives. All decision makers less risk averse than the BRAC prefer one alternative and all decision makers more risk averse than the BRAC will prefer the other alternative. McCarl’s BRAC work showed that if two CDFs cross once there will be one BRAC. If the CDFs for two alternatives cross more than once, there will be multiple BRACs.

McCarl’s research explains why the efficient set for a SDRF analysis may contain more than one risky alternative. When two alternatives are in an efficient set it means that there is a BRAC between the two RACs or:

$$LRAC \leq BRAC \leq URAC$$

This is the reason that increasing LRAC or decreasing URAC to make the range smaller will frequently change the SDRF ranking so only one alternative is in the efficient set.

Stochastic Efficiency Respect to a Function (SERF)

Hardaker, Richardson, Lien and Schumann (2004) merged the use of CEs and Meyer’s range of risk aversion coefficients to create stochastic efficiency with respect to a function (SERF). SERF assumes a utility function with a risk aversion range of, $U(r_1(z), r_2(z))$, but instead of evaluating CEs at the two extreme RACs, it evaluates CEs for many RACs between the LRAC and the URAC. The SERF ranking is performed on many risky alternatives simultaneously, however, in keeping with the notation thus far it is described in terms of ranking two alternatives. Two risky alternatives, F and G, can be compared and ranked at each RAC_i as follows:

- F(z) preferred to G(z) at RAC_i if $CE_{Fi} > CE_{Gi}$
- F(z) indifferent to G(z) at RAC_i if $CE_{Fi} = CE_{Gi}$
- G(z) preferred to F(z) at RAC_i if $CE_{Fi} < CE_{Gi}$

SERF extends the lower RAC and upper RAC case to use a large number of RAC’s uniformly distributed between two extreme RACs. In other words, define the lower RAC and the upper RAC and then divide the range of the RAC’s into 23 equal intervals and evaluate the CEs for all risky alternatives at each interval. This series of calculations produces 25 CEs for each alternative so one can check the ranking of all alternatives at 25 RACs (Figure 10.12). An advantage of SERF over SDRF is that SERF simultaneously compares several risky alternatives while SDRF is a pairwise comparison. The values for the resulting SERF table can be converted to a SERF chart with the RACs on the horizontal axis and the CEs on the vertical axis (Figure 10.13.) Read the SERF chart as follows for two risky alternatives F(x) and G(x):

1	A	B	C	D	E	F	G
2	Stochastic Efficiency with Respect to A Function (SERF) Under a Neg. Exponenti						
3	© 2005						
4	Min RAC	-1	1= Neg. Exponential				
5	Max RAC	2	1				
6	ARAC	TRUE	U(x) = -exp(-ARAC*x)				
7	Initial Weal	0					
8	a. Paramet	Null	Overall:				
9	b. Paramet	Null	Average 21.77				
10	c. Paramet	Null	Std. Dev. 10.378258				
11		ARAC	A	B	C	D	E
12	1	-1	24.00	46.63	21.43	10.12	43.95
13	2	-0.8750	23.60	46.07	20.94	10.10	43.42
14	3	-0.7500	23.14	45.34	20.33	10.08	42.74
15	4	-0.6250	22.63	44.32	19.60	10.06	41.87
16	5	-0.5000	22.09	42.85	18.74	10.05	40.75
17	6	-0.3750	21.54	40.66	17.79	10.03	39.39
18	7	-0.2500	20.98	37.49	16.80	10.01	37.87
19	8	-0.1250	20.43	33.60	15.81	9.99	36.32
20	9	0.0000	19.90	29.73	14.86	9.97	34.83
21	10	0.1250	19.40	26.44	13.99	9.95	33.48
22	11	0.2500	18.94	24.03	13.22	9.94	32.35
23	12	0.3750	18.53	22.33	12.56	9.92	31.43
24	13	0.5000	18.17	21.10	12.02	9.90	30.69
25	14	0.6250	17.86	20.18	11.55	9.88	30.10
26	15	0.7500	17.58	19.48	11.17	9.87	29.61
27	16	0.8750	17.34	18.94	10.83	9.85	29.20
28	17	1.0000	17.12	18.51	10.55	9.83	28.86
29	18	1.1250	16.93	18.16	10.30	9.82	28.58
30	19	1.2500	16.76	17.86	10.09	9.80	28.34
31	20	1.3750	16.61	17.62	9.91	9.79	28.13
32	21	1.5000	16.48	17.41	9.74	9.77	27.95
33	22	1.6250	16.36	17.23	9.60	9.76	27.79
34	23	1.7500	16.25	17.07	9.47	9.74	27.65
35	24	1.8750	16.15	16.93	9.36	9.73	27.52
36	25	2.0000	16.06	16.81	9.25	9.72	27.41

Figure 10.12. Example of a SERF Table for Comparing Five Scenarios.

- $F(x)$ is preferred to $G(x)$ over the range of RACs where the CE_F line is above the CE_G line,
- $G(x)$ is preferred to $F(x)$ over the range of RACs where the CE_G line is above the CE_F line, and
- decision makers are indifferent between G and F at the RAC where the CE lines intersect.

If a CE line in the SERF chart remains positive then rational decision makers will prefer the risky alternative over a risk free alternative. However, if the CE line goes negative at RAC r_j , then decision makers with RACs greater than r_j would prefer a risk free alternative.

The SERF chart displays the rankings of risky alternatives that are consistent with particular RAC values for different types (classes) of decision makers. Rankings based on SERF charts are easy to explain because one can identify ranges of RAC levels over which a scenario is preferred to other contenders. For example, if the CE_F line remains above the CE_G line for the RAC range of 0 to 4.0 then it is highly likely that all risk neutral and risk averse decision makers will prefer scenario F over this range of risk aversion.

The SERF procedure uses the absolute risk aversion coefficient or ARAC or r_a . The discussion of risk aversion coefficients earlier in this chapter indicated that the relative risk aversion coefficients for risk neutral to extremely risk averse range from zero to 4.0 and that $r_a = r_r/Z$. Using this formula the analyst can run the SERF analysis for the range of r_a (0.0, $4/\bar{Z}$), where \bar{Z} is the average wealth for the decision maker. This type of range tests the rankings for decision makers who are risk neutral, moderately risk averse, to extremely risk averse.

The SERF table and SERF chart method for ranking risky scenarios is analogous to performing 25 separate SDRF analyses using very small RAC intervals. After each stochastic dominance analysis, one would record the F vs. G ranking and then report the results in a table and a chart before proceeding with the next comparison of F to H.

Simetar's SERF option takes advantage of Excel's power to update calculations and allows the user to interactively experiment with alternative RACs and utility functions. The SERF Chart provides a visual depiction of how decision makers with different levels of risk aversion will likely rank risky scenarios. Change the min RAC and max RAC to test the range of the rankings and to add precision to identifying the RACs where the rankings switch. Use the SERF Table to determine the BRAC where decision makers become indifferent between two scenarios, i.e., where the differences between the CE's is zero.

An example of a SERF analysis is provided in Figure 10.13 and in SERFTb1 worksheet of Best Demo.XLS. The SERF table is calculated for 25 RAC values ranging uniformly between the LRAC (-1) and the URAC (2) (Figure 10.13). Comparing the CEs across the scenarios for a particular RAC show the absolute preference or ranking. For example, when the RAC is 0.875 the E scenario is ranked first with a CE of 29.20 while B is ranked second with a CE of 18.94 (Figure 10.13).

Using a RAC range of -1 to $+2$ in Figure 10.12 reveals that scenario B is ranked higher than (preferred to) E when the RAC is less than -0.375 , but E is preferred over B once the RAC increases beyond -0.25 . This result indicates a ranking switch between -0.375 and -0.25 . To make the rankings easier to see, a SERF chart is calculated by Simetar in SERFTb11 worksheet (Figure 10.13). At each RAC value the scenario that has the largest CE line in the SERF chart is ranked first, as it has the highest CE. At the 0.50 RAC we can see that the order of ranking for the five scenarios is E, B, A, C, and D in the SERF Chart (Figure 10.13). From Figure 10.13 we can see that there is a BRAC between scenarios C and D at about 1.25 . The vertical distance between the CE lines is the degree of conviction or confidence premium of the dominate strategy over the other scenarios.

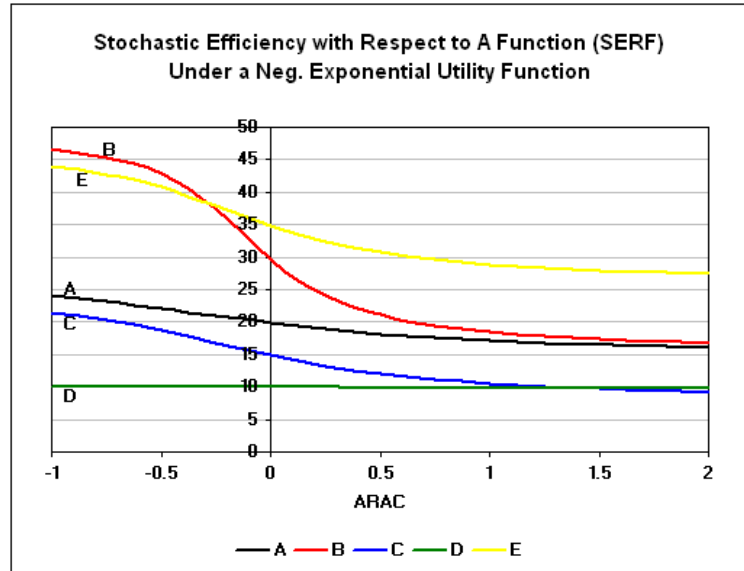


Figure 10.13. Example of a SERF Chart.

An advantage of using the SERF ranking strategy is that it is dynamic and updates itself each time the minimum and maximum RACs or the utility function are changed. Changing the utility function from the Exponential to the Power Utility function may change the rankings. Simetar’s default utility function for SERF is the negative exponential. In the SERF output worksheet an option is provided for changing the utility function in cell D4. The alternative utility functions available and their code number(s) are:

- negative exponential, 1,
- power, 2,
- expo-power, 3,
- quadratic, 4,
- log, 5,
- exponent, 6,
- HARA, 7.

The functional forms for these seven utility functions are summarized in Table 10.11. Also in Table 10.11 is a list of the

	I	J	K	L	M	N
1	Neg. Exponential Utility Weighted Risk Premiums Relative to A					
2	Base Scenario					
3	A Base Alternative for Risk Premiums					
4	ARAC	A	B	C	D	E
5	-1	-	22.63	(2.57)	(13.88)	19.95
6	-0.8750	-	22.48	(2.66)	(13.50)	19.82
7	-0.7500	-	22.20	(2.81)	(13.06)	19.60
8	-0.6250	-	21.69	(3.03)	(12.57)	19.23
9	-0.5000	-	20.76	(3.35)	(12.05)	18.66
10	-0.3750	-	19.13	(3.74)	(11.51)	17.85
11	-0.2500	-	16.51	(4.18)	(10.97)	16.89
12	-0.1250	-	13.17	(4.62)	(10.44)	15.89
13	0.0000	-	9.83	(5.03)	(9.92)	14.93
14	0.1250	-	7.04	(5.41)	(9.44)	14.08
15	0.2500	-	5.09	(5.72)	(9.01)	13.41
16	0.3750	-	3.80	(5.97)	(8.62)	12.90
17	0.5000	-	2.92	(6.16)	(8.27)	12.52
18	0.6250	-	2.32	(6.30)	(7.97)	12.24
19	0.7500	-	1.90	(6.42)	(7.71)	12.03
20	0.8750	-	1.60	(6.50)	(7.49)	11.86
21	1.0000	-	1.38	(6.57)	(7.29)	11.74
22	1.1250	-	1.22	(6.63)	(7.12)	11.64
23	1.2500	-	1.10	(6.67)	(6.96)	11.57
24	1.3750	-	1.00	(6.71)	(6.83)	11.51
25	1.5000	-	0.93	(6.74)	(6.71)	11.47
26	1.6250	-	0.87	(6.76)	(6.60)	11.43
27	1.7500	-	0.82	(6.78)	(6.50)	11.40
28	1.8750	-	0.78	(6.79)	(6.42)	11.38
29	2.0000	-	0.75	(6.80)	(6.34)	11.36

Figure 10.14. Example of a Risk Premium Table.

parameters required for each utility function. The negative exponential requires the user provide a lower and upper RAC, and they can be either an absolute or relative RACs. The remaining utility functions require additional parameters. For example, the power function requires a “w” parameter which reflects wealth; but w minus any stochastic income value, x_i , has to be positive. Constraints for specifying the parameters on the remaining utility functions are indicated in Table 10.11.

Table 10.11. Utility Functions Available for SERF and Their Parameters.

	Neg. Exponential $\Theta \cdot \exp(-ARAC \cdot x)$	Power $(1/(1-RRAC)) \cdot x^{(1-RRAC)}$	Expo-Power $\Theta \cdot \exp(-b \cdot x^a)$	Quadratic $a \cdot x - b/2 \cdot x^2$
ARAC ¹	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$
RRAC ²	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$
W ³	-	$w + x > 0$	$w + x > 0$	$w + x > 0$
a ⁴	-	-	$ab > 0$	$a > 0$
B	-	-	$ab > 0$	$b > 0$
C	-	-	-	-
Θ	-	-	-	-

	Log $\ln(x+a)$	Exponent $(x+a)^b$	HARA $c/(1-c) \cdot (a+b/c \cdot x)^{(1-c)}$
ARAC ¹	$(-\infty, \infty)$	$(0, \infty)$	$(-\infty, \infty)$
RRAC ²	$(-\infty, \infty)$	$(0, \infty)$	$(-\infty, \infty)$
W ³	$w + x > 0$	$w + x > 0$	$w + x > 0$
a ⁴	$x + a > 0$	$0 < a < 1$	$(-\infty, \infty)$
B	-	$b \geq 0$	$b > 0$
c	-	-	$(-\infty, \infty)$
Θ	-	-	$a + b/c > 0$

¹ARAC is the absolute risk aversion coefficient.
²RRAC is the relative risk aversion coefficient.
³w reflects wealth and w plus x_i must be positive for all random values.
⁴a, b, c, Θ are parameters required for the particular utility functions.

- Risk Premiums (RP)

The SERF option in Simetar also produces a risk premium table and chart. For this table, one of the scenarios is selected as the base by the user and a table of RPs is calculated as well as an RP chart (Figures 10.14 and 10.15). The table and chart show the perceived premium that each risky scenario provides relative to the base scenario at 25 alternative RAC levels. The user can experiment with alternative base scenarios to calculate the relative risk premiums across scenarios. A drop down menu is provided so the user can change base scenarios.

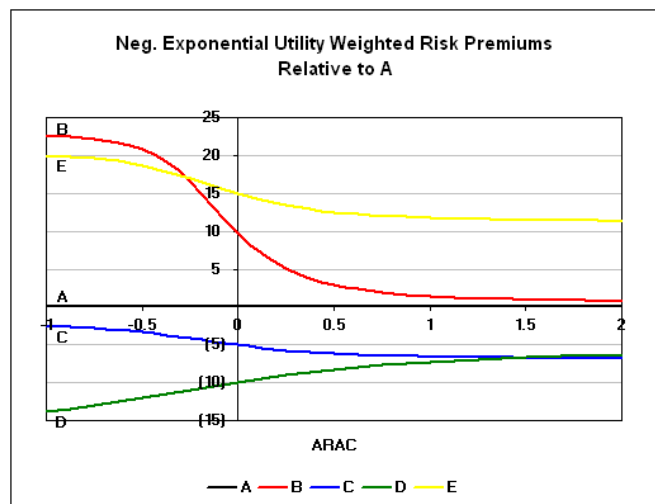


Figure 10.15. Example of a Risk Premium Chart.

The RP values in Figure 10.14 demonstrate how scenarios B, C, D, and E rank relative to the base scenario A at alternative RACs. The results show that B is preferred over A at all RACs because of positive RP's over this range. The negative RP values for scenario C indicate that A is preferred over the range of RACs evaluated. In the RP chart (Figure 10.15) the preferred scenario is the line on top, i.e., has the highest RP. The base scenario, in this case A, is always the zero axis because of the RP formula:

$$RP_i = CE_{\text{Scenario } i} - CE_{\text{Base } i} \text{ for } RAC_i.$$

Thus, if an RP line is positive it shows the value an alternative scenario has over the base. If the RP line is negative it shows the value of the base scenario over another scenario.

Summary of Risk Ranking Procedures

Once you have completed the scenario rankings using the 11 procedures described in this section prepare a summary of the rankings (Table 10.12). A sample summary rankings table below indicates that scenario E is ranked first or second by 9 procedures, all except the standard deviation ($\hat{\sigma}$) and the Mini-Max procedures. Scenario D is ranked first or second 3 times. On the other hand scenario D is ranked fourth or fifth by 7 procedures and scenario C is ranked fourth or fifth by 7 of the 11 procedures.

Table 10.12. Summary of Scenario Rankings Across Alternative Procedures.

Scenario	Mean	$\hat{\sigma}$	MV	Min	Max	Mini-Max	CV	P(X > 20%)	CDF	SDRF*	SERF	ARAC >
A	3	2	2	3	3	2	3	3	3	3	3	3
B	2	5	5	2	1	5	4	2	2	1	2	2
C	4	3	3	5	4	3	5	4	4	4	4	4
D	5	1	4	4	5	1	1	5	5	4	5	5
E	1	4	1	1	2	4	2	1	1	1	1	1

*SDRF results in a tie between B and E in the most efficient set given ARAC's of -1 and +1. Alternatives C and D are ranked last in the least efficient set.

References

- Anderson, J.R. and J.L. Dillon. Risk Analysis in Dryland Farming Systems. Farming Systems Management Series No. 2, FAO, Rome, 1992.
- Arrow, K.J. Essays in the Theory of Risk Bearing. Chicago: Markham, 1971.
- Freund, R.J. "The Introduction of Risk into a Programming Model." Econometrica, 24(1956):253-261.
- Hadar, J. and W.R. Russell. "Rules for Obtaining Uncertain Prospects." American Economic Review, 59(1969):25-34.
- Hardaker, J.B. "Some Issues in Dealing with Risk in Agriculture." University of New England, Graduate School of Agricultural and Resource Economics, Working Paper Series in Agricultural and Resource Economics, ISSN 14421909, No. 2000-3-March 2000.
www.une.edu.au/feb1/EconStud/wps.htm
- Hardaker, J.B., Huirne, R.B.M., and J.R. Anderson. Coping with Risk in Agriculture. New York: CAB International, 1997.
- Hardaker, J.B., J.W. Richardson, G. Lien, and K.D. Schumann. "Stochastic Efficiency Analysis with Risk Aversion Bound: A Simplified Approach." Australian Agricultural and Resource Economics Journal, forthcoming, June 2004.
- McCarl, B. "Notes on Stochastic Dominance." Department of Agricultural Economics, Texas A&M University, 1988.
- McCarl, B. "Preference Among Risky Prospects Under Constant Risk Aversion." Southern J. Agricultural Economics, Dec. 1988.
- McCarl, B.A. and D. Bessler. "Estimating an Upper Bound on the Pratt Risk Aversion Coefficient When the Utility Function is Unknown." Australian Journal of Agricultural Economics, 33(1989):56-63.
- Meyer, J. "Choice Among Distributions." Journal of Economic Theory, 14(1977):326-336.
- Mjelde, J.W. and M.J. Cochran. "Obtaining Lower and Upper Bounds on the Value of Seasonal Climate Forecasts as a Function of Risk Preferences." Western Journal Agricultural Economics, Vol. 13, Dec. 1988.
- Pratt, J.W. "Risk Aversion in the Small and the Large." Econometrica, 32(1964):122-136.
- Richardson, J.W. and H.P. Mapp, Jr. "Use of Probabilistic Cash Flows in Analyzing Investment Under Conditions of Risk and Uncertainty." Southern Journal Agricultural Economics. Dec. 1976.

Chapter 11 Bootstrap Simulation

In forecasting it is useful to indicate how much confidence you have in the forecast. One means of doing so is to report confidence intervals for each forecast. Thus giving the user an impression that you did not generate the perfect forecast and there is some “risk” about the forecast. Confidence intervals about the forecast itself are pretty standard, but confidence intervals for the parameters in a sample or in a model are more difficult to develop. For example, the confidence interval about the standard deviation in a sample or for an elasticity maybe hard to estimate due to small sample size. Bootstrap simulation offers an economical way to estimate confidence intervals for parameters in a sample or a model.

Bootstrap Simulation

A standard deviation is required to develop confidence intervals about parameter estimates such as the mean, standard deviation or elasticity. For example, the confidence interval about the mean is defined as:

$$\bar{X} \pm Z_{1-\sigma/2} \left(\frac{\hat{s}}{\sqrt{N}} \right)$$

At the $\alpha = 5\%$ level this formula becomes:

$$\bar{X} \pm 1.96 * \left(\frac{\hat{s}}{\sqrt{N}} \right)$$

The estimate of \bar{X} is simple and given an adequate size sample, N, the \hat{s} can be calculated. The problem usually occurs in calculating the variance (or S^2) for the sample. Variances for other parameters that describe distributions similarly can be difficult to impossible to develop in a theoretical manner. The simulation technique, bootstrap, reported in 1979 by Elfon and thoroughly described by Vose offers a procedure for estimating variances for distribution parameters.

The need for bootstrap simulation comes from small samples and extremely high costs of increasing sample size. Small sample size leads to large variances on population parameters. When sample size is small the bootstrap method provides a means of increasing the sample size by re-sampling the original sample with replacement, many times. Simulation thus offers an inexpensive way to expand sample size and reduce the variance on the population parameters.

Conover provides a simple description of bootstrap simulation that is summarized as follows. The bootstrap simulation procedure draws M samples of N random values with replacement from an actual random sample of size N. The parameters of interest for the distribution (say, \bar{X} , S^2 , skewness, kurtosis, elasticity, etc.) are calculated and recorded for each of M samples. The resulting empirical distribution of M observations for each parameter of interest constitutes a “distribution of the population.” In other words, the bootstrap sample of M means is the empirical distribution of the true mean and can be used to estimate the \bar{X} and S^2 for the true mean.

The confidence intervals for the empirical distribution of M means can be observed directly from the simulated distribution population. From the sorted empirical distribution the $\alpha/2$ and $1-\alpha/2$ sample quantiles define the confidence interval about the mean of the true population. For an $\alpha=0.05$ level and $M = 1000$ the confidence intervals are at the 25th and 975th sorted sample values.

The problem with the bootstrap technique then rests on the validity and randomness of the original sample and the size of M . Elfon and Tibshirane recommend a minimum of 250 iterations or M of 250 and more if necessary. The bootstrap is such a simple simulation problem that 10,000 or more iterations can be run without much effort so M can be quite large.

To demonstrate the bootstrap simulation procedure, an original sample of size 20 random values is used. The number of iterations, M , is set at 10,000 and the KOVs for the simulation experiment are the mean and standard deviation of the sample. In the original sample the mean is 10.5 and the standard deviation is 5.916 (see Bootstrap Demo.XLS). The =BOOTSTRAPPER() function in Simetar is programmed 20 times (E8:E27 in Figure 11.1) and the mean of this 20 observation sample (cell E29) is specified as the Output Variable for Simetar. The results of the bootstrap simulation are summarized in Figure 11.2: the sample mean is 10.5 and its standard deviation is 1.289. The minimum simulated mean is 6.85 and the maximum mean is 14.7. A pdf graph of the empirical distribution simulated for the mean is provided in Figure 11.2. The mean, and $\alpha = 5\%$ upper and lower confidence intervals for the mean distribution are indicated in the pdf graph (Figure 11.2).

	A	B	C	D	E	F	G	H
7	Random Sample of Values				Simulated Sample			
8		1			16	=bootstrapper(\$B\$8:\$B\$27)		
9		2			2	=bootstrapper(\$B\$8:\$B\$27)		
10		3			1	=bootstrapper(\$B\$8:\$B\$27)		
11		4			12	=bootstrapper(\$B\$8:\$B\$27)		
12		5			2	=bootstrapper(\$B\$8:\$B\$27)		
13		6			1	=bootstrapper(\$B\$8:\$B\$27)		
14		7			6	=bootstrapper(\$B\$8:\$B\$27)		
15		8			20	=bootstrapper(\$B\$8:\$B\$27)		
16		9			9	=bootstrapper(\$B\$8:\$B\$27)		
17		10			13	=bootstrapper(\$B\$8:\$B\$27)		
18		11			6	=bootstrapper(\$B\$8:\$B\$27)		
19		12			4	=bootstrapper(\$B\$8:\$B\$27)		
20		13			16	=bootstrapper(\$B\$8:\$B\$27)		
21		14			14	=bootstrapper(\$B\$8:\$B\$27)		
22		15			19	=bootstrapper(\$B\$8:\$B\$27)		
23		16			3	=bootstrapper(\$B\$8:\$B\$27)		
24		17			2	=bootstrapper(\$B\$8:\$B\$27)		
25		18			7	=bootstrapper(\$B\$8:\$B\$27)		
26		19			9	=bootstrapper(\$B\$8:\$B\$27)		
27		20			1	=bootstrapper(\$B\$8:\$B\$27)		
28					KOVs			
29	Mean	10.5		Mean	8.15	=AVERAGE(E8:E27)		
30	StDev	5.91608		StDev	6.393047	=STDEV(E8:E27)		
31	CV	56.34362		CV	78.4423	=(E30/E29)*100		
32	Min	1		Min	1	=MIN(E8:E27)		
33	Max	20		Max	20	=MAX(E8:E27)		

Figure 11.1. Example of a Bootstrap Simulation Model.

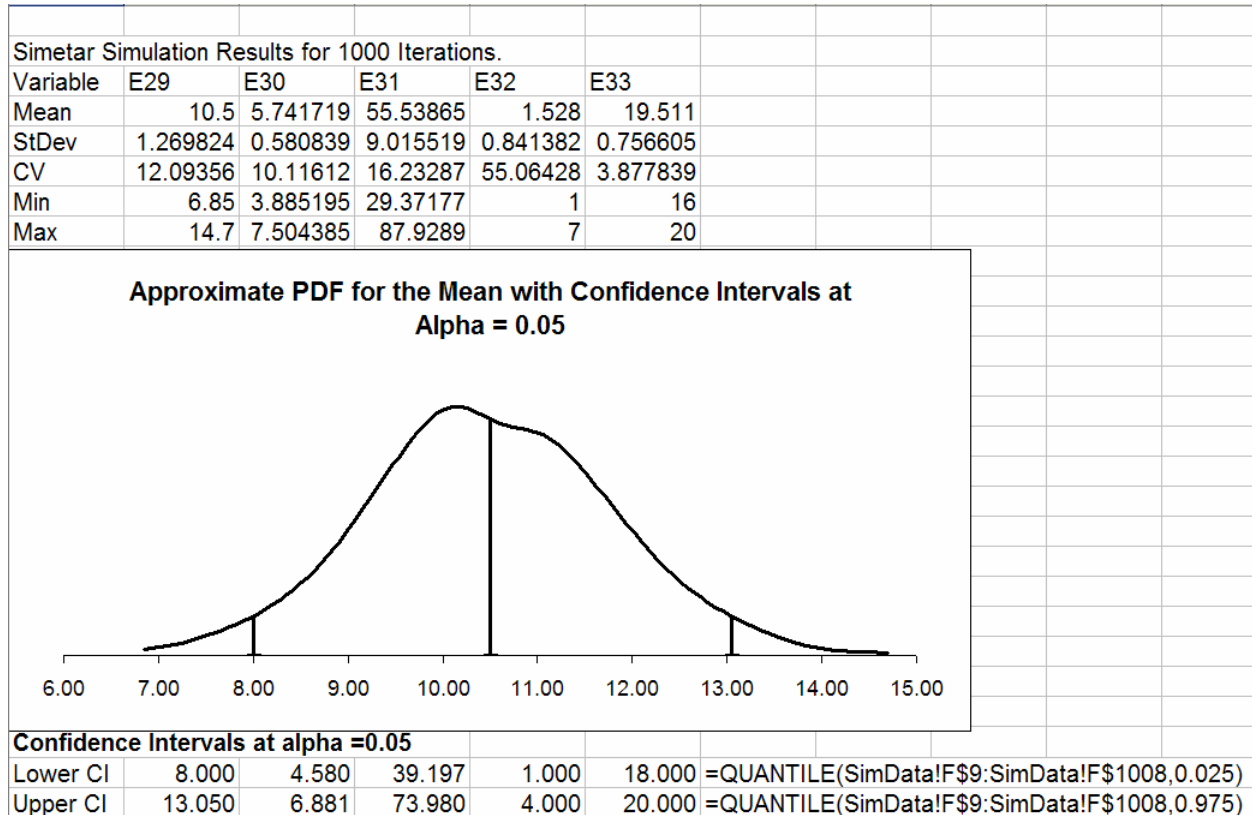


Figure 11.2. Results of a Bootstrap Simulation for the Distribution Parameters of a Population.

The demo bootstrap simulation also generate the empirical distributions for the standard deviation, coefficient of variation, minimum, and maximum. It is interesting to note that the average standard deviation, 5.741, has a simulated standard deviation of 0.5808 and the $\alpha = 5\%$ confidence intervals are 4.571 and 6.878 (Figure 11.2). The lower and upper confidence intervals at the $\alpha = 5\%$ level were calculated using the Simetar function =QUANTILE() as follows:

=QUANTILE (range of M simulated values, 0.025)
 and
 =QUANTILE (range of M simulated values, 0.975)

The bootstrap simulation procedure demonstrated here is nonparametric. It is possible to assume/enforce your own assumed distribution on the bootstrap sample, thus creating a parametric bootstrap. This is done by estimating the best distributional shape for the observed random sample of N values and then sampling from that distribution. See Vose for further discussion of this topic.

Bootstrap Simulation of Multivariate Distributions

Bootstrap simulation can be applied to simulating large multivariate empirical distributions. For a complete description of this application see Chapter 7.

Bootstrap Simulation and Regression Analysis

When a regression model of X on Y is calculated we use paired observations x_i and y_i . The parameters for the OLS model are \hat{a} and \hat{b} or:

$$\hat{Y} = \hat{a} + \hat{b}X$$

Bootstrap simulation can be used to estimate the uncertainty about the regression coefficients, the \hat{b} 's. Vose simply frames the regression bootstrap simulation problem as falling in two types: A and B, based on the type of data in the problem.

- Type A is the case where the x_i and y_i are paired observations and come from a bivariate normal distribution (cross sectional data).
- Type B is the case where x_i is determined and the resulting paired value for y_i is a random variable from a normal distribution or $Y \sim N(\hat{a} + \hat{b}X_i, \sigma)$ or time dependent economic data.

Vose argues that the experimental design is Type A if x_i and y_i are generated randomly together such as measuring people's height and weight or input and output response. Type B is where alternative x_i values are fixed and their associated y_i values are observed. The problem of estimating a demand function falls in Type B because quantity supplied is determined and one random price is observed each year (Figure 11.3). In a sector model the quantity supplied is fixed once harvest is completed (Figure 11.3). During the marketing year price is determined as a stochastic variable given the quantity supplied and the factors of demand. On average the forces of demand result in an average demand function D in Figure 11.3 which is the center of the price pdf. In any one year price is a stochastic value as depicted by a pdf of prices for the given quantity supplied.

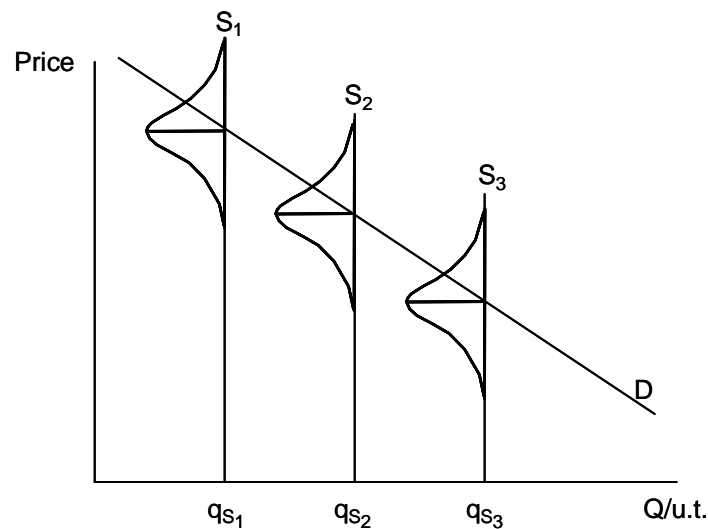


Figure 11.3. Random Price Observed for a Fixed Quantity Supplied Defines a Demand Function.

– **Type A Bootstrap**

The bootstrap simulation for a Type A data set is demonstrated in Figure 11.4 for a multiple regression problem. The bootstrap calls for resampling the original sample of Ys N times with replacement to get a new sample called Y". Next the Y" _i values are used in a table lookup function to find their paired x_i values to fill the X" matrix. The complete sample of N X"s and Y"s is used in a regression to estimate $\hat{\alpha}$ and $\hat{\beta}$ s. The $\hat{\alpha}$ and $\hat{\beta}$ s are the KOVs for the system and after M iterations the uncertainty and confidence intervals for the parameters are calculated.

	A	B	C	D	E	F	G	H	I	J	K	L
24	Generate the Paired Random Y"s and X"s					Formulas Used to Generate the Bootstrap Random Paired Values						
25	Obs	Y"	X1"	X2"		Column B			Column C			
26	1	5.394074	2.547105	8.021146		=bootstrapper(\$A\$7:\$A\$22)			=VLOOKUP(\$B26,\$A\$7:\$C\$22,2)			
27	2	5.36772	2.335552	8.136348		=bootstrapper(\$A\$7:\$A\$22)			=VLOOKUP(\$B27,\$A\$7:\$C\$22,2)			
28	3	5.349434	2.338206	8.156177		=bootstrapper(\$A\$7:\$A\$22)			=VLOOKUP(\$B28,\$A\$7:\$C\$22,2)			
29	4	5.397776	2.340115	8.097324		=bootstrapper(\$A\$7:\$A\$22)			=VLOOKUP(\$B29,\$A\$7:\$C\$22,2)			
30	5	5.36772	2.335552	8.136348		=bootstrapper(\$A\$7:\$A\$22)			=VLOOKUP(\$B30,\$A\$7:\$C\$22,2)			
31	6	5.366653	2.662592	7.930083		=bootstrapper(\$A\$7:\$A\$22)			=VLOOKUP(\$B31,\$A\$7:\$C\$22,2)			
32	7	5.349434	2.338206	8.156177		=bootstrapper(\$A\$7:\$A\$22)			=VLOOKUP(\$B32,\$A\$7:\$C\$22,2)			
33	8	5.388031	2.558678	7.998506		=bootstrapper(\$A\$7:\$A\$22)			=VLOOKUP(\$B33,\$A\$7:\$C\$22,2)			
34	9	5.409495	2.512356	8.083274		=bootstrapper(\$A\$7:\$A\$22)			=VLOOKUP(\$B34,\$A\$7:\$C\$22,2)			
35	10	5.366653	2.662592	7.930083		=bootstrapper(\$A\$7:\$A\$22)			=VLOOKUP(\$B35,\$A\$7:\$C\$22,2)			
36	11	5.36772	2.335552	8.136348		=bootstrapper(\$A\$7:\$A\$22)			=VLOOKUP(\$B36,\$A\$7:\$C\$22,2)			
37	12	5.318948	2.328787	8.289946		=bootstrapper(\$A\$7:\$A\$22)			=VLOOKUP(\$B37,\$A\$7:\$C\$22,2)			
38	13	5.349434	2.338206	8.156177		=bootstrapper(\$A\$7:\$A\$22)			=VLOOKUP(\$B38,\$A\$7:\$C\$22,2)			
39	14	5.355041	2.25004	8.223416		=bootstrapper(\$A\$7:\$A\$22)			=VLOOKUP(\$B39,\$A\$7:\$C\$22,2)			
40	15	5.394074	2.547105	8.021146		=bootstrapper(\$A\$7:\$A\$22)			=VLOOKUP(\$B40,\$A\$7:\$C\$22,2)			
41	16	5.377923	2.395636	8.128136		=bootstrapper(\$A\$7:\$A\$22)			=VLOOKUP(\$B41,\$A\$7:\$C\$22,2)			
42												
43	Multiple Regression Esimated from the Y" and X"s data								Simulation Results for the Regression Analysis			
44	OLS Regression Statistics for Y", 8/9/2001 8:44:36 AM								Simetar Simulation Results for 1000 Iterations. 8:4			
45	F-test	4.237	Prob(F)	0.038	Unrestricted Model			Mean	7.305008	-0.0433	-0.22501	
46	S.E. Regr	0.019	CV Regr	0.360	F-test	4.237	StDev	1.384124	0.101638	0.143946		
47	R²	0.395	Durbin-W	2.092	R²	0.395	CV	18.9476	-234.714	-63.9731		
48	RBar²	0.301	Rho	-0.088	RBar²	0.301	Min	0.194878	-0.30957	-0.46223		
49	Akaike Inf	-7.851	Goldfeld-C	0.392			Max	9.874683	0.409924	0.516594		
50	95%	Intercept	X1"	X2"								
51	Beta	7.640	-0.093	-0.253	KOVs for Simulation			Confidence Intervals				
52	S.E.	1.243	0.095	0.127			Lower CI	3.599868	-0.21145	-0.40768		
53	t-test	6.146	-0.979	-1.988			Upper CI	9.142967	0.195451	0.155943		

Figure 11.4. Example of Bootstrap Simulation for Regression Analysis – Type A Data.

An example of a Type A bootstrap simulation for regression is demonstrated in the first part of Bootstrap Regression Demo.XLS. The Y, X₁ and X₂ data were pairwise sorted on the Y values in ascending order in rows 6-22. The =BOOTSTRAPPER() function was used to generate random Ys in rows 26-41 by sampling with replacement (Figure 11.4). The X₁" and X₂" values in rows 25-42 are generated using Excel's VLOOKUP so the observed pairwise relationships are maintained (Figure 11.4). The Multiple Regression option in Simetar was used to estimate the betas ($\hat{\alpha}$ and $\hat{\beta}$ s) in row 51 for the OLS regression using the Y" and X"s in rows 26-41. The betas are specified as the KOV values in the Simetar Simulation Engine which was run for M = 1000 iterations. Results from the simulation are summarized to the right of the regression output (Figure 11.4). The uncertainty for the betas is displayed in terms of their standard deviations and confidence intervals (rows 52 and 53) in Figure 11.4 and as pdf graphs with the means drawn in as vertical lines in the demo (not shown in Figure 11.4). The lower and

upper confidence intervals for the betas are calculated using the =PTARGET() function at the $\alpha = 5\%$ level.

– Type B Bootstrap

For temporally dependent data, the Type B bootstrap simulation procedure must be used. In this case, assume that the known X's predict the unknown Ys and the error lies about the regression line. In other words, the uncertainty is the variation about the line which is measured by the residuals. Generally it is assumed the residuals are normally distributed so this becomes a parametric bootstrap simulation problem.

An example of a Type B bootstrap simulation is provided in the second half of Bootstrap Regression Demo.XLS. The problem is a demand function (Y) with four explanatory variables ($X_1 - X_4$). An initial regression is fit to the data (rows 89-106) to estimate the standard deviation of the residuals about the regression line. In the example, the S.E. Residuals is 0.009457. The S.E. Residuals is used to simulate random deviations from the regression line for the 16 observed Ys in rows 111-126. The procedure assumes constant Xs and a fixed set of betas for the original or average regression model, so use the predicted (Y-hat) from the first regression. Assuming normality for the residuals the bootstrap simulation equation for the Y"s becomes:

$$\tilde{Y}'' = \hat{Y} + \text{S.E. Residual for Base Regression} * \text{SND}$$

A multiple regression is then estimated using the bootstrap \tilde{Y}'' s or:

$$\tilde{Y}'' = \hat{a} + \hat{b}_i X_i$$

in rows 129-143 of the demo (Figure 11.5). The KOVs for the $M = 1000$ iteration simulation are the betas, the standard deviation of the residuals (S.E. Residuals is row 145), and the elasticities.

Results of the simulation are presented to the right of the bootstrap regression. In addition to the estimated standard deviation for the betas this procedure estimated the standard deviation for the S.E. of Residuals and the elasticities. The $\alpha = 5\%$ confidence intervals are developed for all of these parameters using the empirical distributions generated by the regression. Graphs of the pdfs for the elasticities are developed using Simetar to demonstrate the risk associated with their estimates.

The results of the bootstrap simulation are summarized in Figure 11.5. The standard errors for the elasticities are summarized in line 131 and their confidence intervals at the $\alpha = 5\%$ level are in lines 137 and 138. Charts of the pdfs for the elasticities are calculated from their empirical distributions. The pdf charts are presented in the demo program and the first two are included in Figure 11.5.

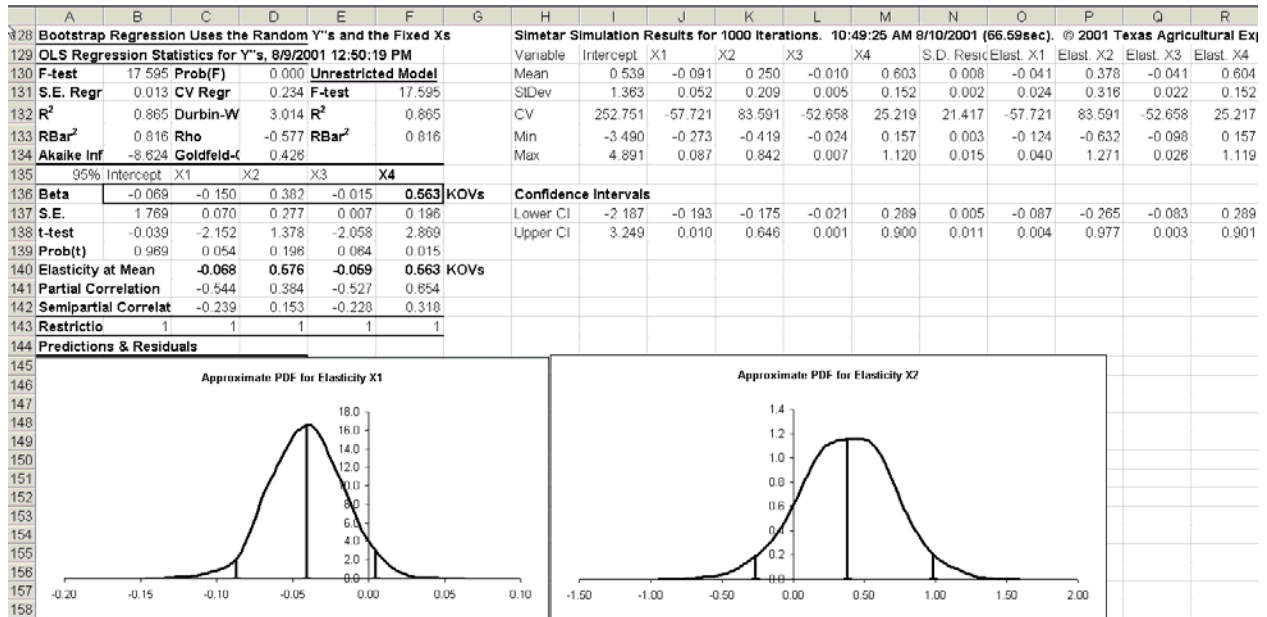


Figure 11.5. Example of Bootstrap Simulation for Regression Analysis to Estimate Confidence Intervals for Elasticities.

References

- Conover, W. J. Practical Nonparametric Statistics. New York: John Wiley & Sons, Inc. 1999.
- Elfon, B. "Bootstrap Methods: Another Look at the Jackknife." Annals of Statistics 7(1979): 1-26.
- Elfon, B. and R.J. Tibshirane. An Introduction to the Bootstrap. New York: Chapman and Hall, 1993.
- Vose, D. Risk Analysis: A Quantitative Guide. New York: John Wiley & Sons, Ltd. Sec. Ed., 2000.

Chapter 12

Optimization of a Simulation Model

Generally simulation models are not optimized. However, special techniques are available to optimize deterministic simulation models. Optimal control theory techniques can be used to calculate “approximate” optimal solutions for simulation models. The following discussion draws directly from a bulletin based in part on my dissertation on optimal control theory.

Optimal control theory is a mathematical technique for analyzing systems under alternative sets of controls. Specifically optimal control theory is a technique to determine the optimal values for particular control variables in a model. The technique has been used primarily by engineers and mathematicians in dealing with control problems in physical systems.

Optimal control theory can be readily applied to many agricultural economics problems. Agricultural economists, like engineers, are dealing with complex systems that emit reactions and signals which require management responses. Optimal control analysis can assist in designing information systems and managerial decision procedures that will create desired economic results.

For discrete time models, or continuous models for which discrete numerical approximations can be found, the optimal control problem can be viewed as the problem of choosing variables to maximize an objective function. From this perspective, optimal control becomes the process of maximizing (or minimizing) an objective function. The maximization process may be either static or dynamic, depending on the nature of the model, but is generally thought of in control theory as being dynamic.

Principles of Control Theory

The objective of optimal control theory is to determine the values of control variables that cause a particular system to maximize (or minimize) a given performance measure subject to a set of constraints. Formulation of a control problem involves three steps; development of: (1) a simulation model of the system to be controlled, (2) constraints on the controls, and on input and output variables, and, (3) a performance measure for the system.

In control theory literature, the endogenous variables in the model are referred to as the state variables and are denoted as: $x_1(t)$, $x_2(t)$, ..., $x_n(t)$ for time period t (e.g., production, profits, net worth). The subset of state variables used in the performance measure are referred to as the output variables, and are designated as: $y_1(t)$, $y_2(t)$, ... $y_k(t)$ (e.g., IROR, NPV, profit, ending net worth). Uncontrollable exogenous variables, (e.g., weather, prices, interest rates) are denoted as $z_1(t)$, $z_2(t)$, ..., $z_q(t)$. The exogenous variables that can be controlled by the decision maker, such as fertilizer use or the crop mix are referred to as control inputs (controls). Controls for period t are represented by: $u_1(t)$, $u_2(t)$, ..., $u_m(t)$.

The model equations that describe the endogenous variables can be a function of the controls, other state variables, time and the noncontrollable exogenous variables. For the system to be controlled, one or more of the equations describing the state variables must contain a control variable. In turn, controls are normally a function of one or more of the state variables and/or time and other variables. When controls are a function of state variables, dynamic feed back from the system can be used to throttle successive control values. This circular causal flow

which relates control values to state values and then back to the controls is called a closed loop control problem. When controls are not a function of the state variables the system is an open-loop control problem.

Constraints are usually imposed on the control variables, and can be imposed on the state variables. The constraints limit the controls within boundaries (minimum and maximum) established by the user in light of physical, economic, and political limits of the system. The constraints reduce the number of alternative control paths that must be investigated. Realistic constraints on the controls allows more accurate modeling of the system while reducing the number of feasible trajectories.

A single valued performance measure, the criterion for evaluating the alternative control paths, must be developed for the particular problem being investigated. The performance measure (F) is defined by a mathematical equation that sums weighted values of the output variables or consists of the single output variable of interest. In application, values for the controls are selected by a control procedure in an iterative process that ultimately leads to the set of controls (or control path) that cause the performance measure to be optimized.

An illustration of a dynamic control system is presented in Figure 12.1. The model is simulated to obtain values of the state variables, using as input the following variables: the controls (u_j), initial or lagged values of the states (x_j), and values for any uncontrollable exogenous variables (z_j). The results from the model are used to estimate the values for the state variables (x_i). The estimated values for a subset of the state variables, which have been referred to as output variables (y_j), are used in conjunction with user provided weights (r_j) to compute the value of the performance measure (F). If the stopping criterion for reaching a maximum (or minimum) is not reached the control process continues. The control mechanism (or numerical optimization routine) computes new values for the control variables (u_j) for each iteration, based upon previous values of the performance measure and controls until the objective function value is optimized.

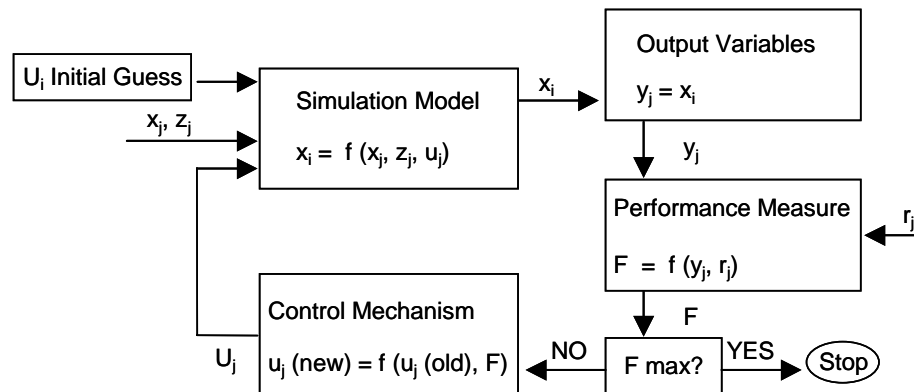


Figure 12.1. Flowchart for an Optimal Control Problem.

In general, the functional form of the performance measure should formalize assumptions regarding the rate of substitution among the output variables. In application, the functional form needs to be as simple as possible in its assignment of a unique real number to each set of output variables. The nature of the functional form for the performance measure depends upon the type of problem being analyzed.

– Terminal Control Problems

A terminal control problem attempts to minimize the system's deviations from some desired level for the output variables in the final year (t_f) or:

$$\text{Minimize: } F = \sum_{i=1}^n r_i [y_i(t_f) - s_i(t_f)]^2$$

where t_f is the final year or stage of the system, s_i is the target value for output variable y_i , and r_i is the parameter weight assigned to the i^{th} output variable measure.

– Tracking Control Problem

Tracking problems where the objective is to keep the output variable, $y_i(t)$, as close as possible to a series of target value, $s_i(t)$, over the interval t_0 to t_f :

$$\text{Minimize: } F = \sum_{j=t_0}^{t_f} \left(\sum_{i=1}^n r_{ij} [y_i(t_j) - s_i(t_j)]^2 \right)$$

where r_{ij} is the weight assigned to the deviation for output variable y_i in time period j from the target value s_{ij} . (This type of objective function has been used to minimize the sum of squared deviates in parameter estimation for non-linear regression models, e.g., Outlaw.)

– Max (or Min) Control Problems

A simple maximization of the key output variables, such as net present value, real net worth, or internal rate of return takes on the form of:

$$\text{Maximize: } F = y_i$$

A weight (r_i) need not be applied to the objective function if there is only one output variable in F .

Numerical Solution of Optimal Control Problems

An alternative to using direct-solution techniques to optimize a set of equations (or a model) is to use direct-search or numerical techniques. Numerical techniques do not require the model be in the state form and can obtain the final (optimal) solution without solving derivatives. In general, the direct-search techniques are hill climbing procedures that utilize alternative methods

of searching the surface of the performance measure for its global maximum (or minimum). In application, the control mechanism selects values for the control variables, determines their impacts on the system's output variables and evaluates the performance measure based on the values of the relevant output variables. This process is repeated in an iterative fashion until any change in the control variables results in a reduction in the value of the performance measure.

Numerical Optimization in Excel

Excel provides an excellent numerical optimization program in its Solver option under the Tools menu. (If Solver does not appear in the Tools dialog box, use Add-Ins option to activate the option.) The Solver provides a pop up menu (Figure 12.2) so the user can specify the three types of values that activate Solver:

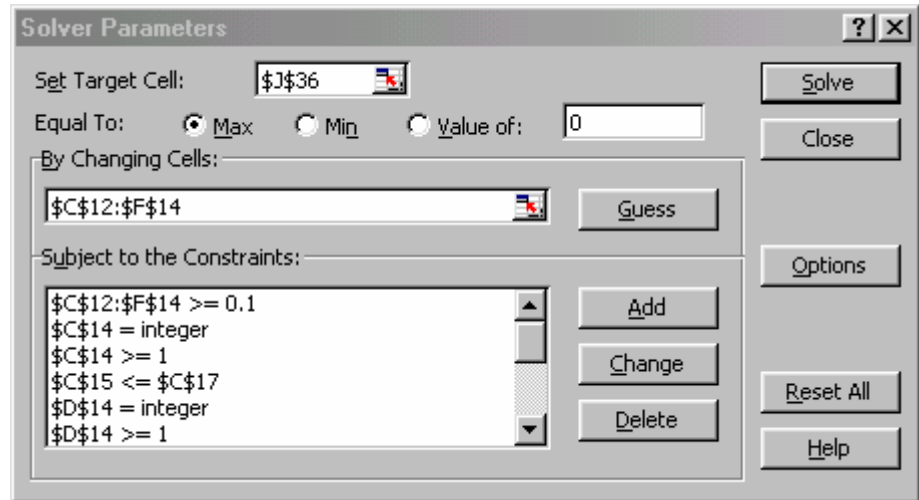


Figure 12.2. Dialog Box for Excel's Solver.

– Target Variable

The user must specify the cell which contains the objective function to be maximized or minimized. The formula in the cell that is indicated in the “Set Target Cell” box must be a function, either indirectly or directly, of the control variables.

– Controls

The control variables which Excel can change to optimize the simulation model are indicated by entering their cell locations into the “By Changing Cells” box (Figure 12.2).

– Constraints

All constraints on the controls and output variables in the model are specified using the constraints editor in the “Subject to the Constraints” box (Figure

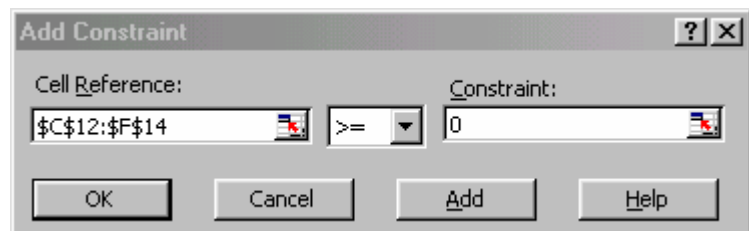


Figure 12.3. Dialog Box for Entering Constraints in Excel's Solver.

12.2). Selecting the “Add” a constraint button brings up a menu (Figure 12.3) which allows the user to specify the cell for the variable to constrain (“Cell Reference” in the left hand box), the type of constraint ($>=$, $=$, $<=$, Integer), and the constraint value (“Constraint” in the right hand box). Caution must be exercised when creating constraints, in that: (1) they must be consistent with each other, (2)

the initial guesses for the controls must be feasible, and (3) the constraint value must be a fixed number. An example of the constraints set to bound the integer control variable in cell C1 between zero and 2000 and is:

C1 \geq 0.0
 C1 \leq 2000.0
 C1 = integer

– Options

The user can control the precision of the optimization by selecting the Options button (Figure 12.2) and changing the settings in the Options menu.

Once the values for Solver have been specified, click the Solve button to make Solver optimize the simulation model. Experience suggests that you should first optimize the simulation model with no constraints. After a satisfactory solution is obtained then add the constraints, one or two at a time. In this way you can locate and avoid inconsistent constraints and observe the impact of the constraints on the objective function.

Once you have obtained a final solution from Solver, change the starting guesses for the controls and rerun Solver. This should be done several times to insure that the optimal solution is robust, i.e., the answer is about the same regardless of your starting values. Numerical optimization is a heuristic search procedure and is thus an approximation of the optimal solution. Thus if the objective function has a single peak you will always get the same answer, while if the function has numerous peaks you will get a different answer each time. Adding terms to make the objective function more non-linear can help in the search for a global optimum.

Two Excel spreadsheets are provided to demonstrate Solver. The first is a simple profit maximizer, Optimal Control Demo.XLS, where a firm has three products, using four inputs, and faces demand functions for its outputs and supply functions for pricing inputs. The printout shows the initial guesses for the 12 controls and a table on the right side with the optimal values. When optimized the model has a maximum profit of \$548,646.33. Excel optimizes the model in a few seconds. The constraints force all controls to be greater than zero, the sum of the controls (x_1 , x_2 , x_3 , and x_4) across all products are less than or equal to 2000, 3000, 2100, and 1200, respectively. Also the x_i values for producing Y_3 must be integers. This last constraint adds to the optimization time; try it without this constraint.

The optimization problem in Optimal Control Demo.XLS is summarized as a firm with three outputs (y_1 , y_2 , and y_3), four inputs (x_1 , x_2 , x_3 , and x_4) and constraints on the maximum amount of each input that can be used. The problem can be stated as:

Maximize: profits for outputs y_1 , y_2 , and y_3
 Subject to: production functions –

$$y_1 = x_{11}^{.33} x_{12}^{.17} x_{13}^{.20} x_{14}^{.30}$$

$$y_2 = x_{21}^{.10} x_{22}^{.08} x_{23}^{.25} x_{24}^{.40}$$

$$y_3 = x_{31}^{.09} x_{32}^{.19} x_{33}^{.15} x_{34}^{.20}$$

Output demand function (prices) –

$$Py_1 = 1050.0 - 0.5y_1$$

$$Py_2 = 1000.0 - 0.25(y_2)^2$$

$$Py_3 = 100.0 - 0.15(y_3)^2$$

Input constraints –

$$2000.0 \geq x_{11} + x_{21} + x_{31} = \text{sum } x_1$$

$$3000.0 \geq x_{12} + x_{22} + x_{32} = \text{sum } x_2$$

$$2100.0 \geq x_{13} + x_{23} + x_{33} = \text{sum } x_3$$

$$1000.0 \geq x_{14} + x_{24} + x_{34} = \text{sum } x_4$$

Input marginal costs –

$$Px_1 = 3.0 + 0.0009 \text{ sum } x_1$$

$$Px_2 = 6.0 + 0.00011 \text{ sum } x_2$$

$$Px_3 = 9.0 + 0.0003 \text{ sum } x_3$$

$$Px_4 = 7.0 + 0.000199 \text{ sum } x_4$$

And: $x_{ij} > 0.0$ for $i = 1, 2, 3$ and $j = 1, 2, 3, 4$,
 x_{3i} 's are integers.

The second spreadsheet, *Deterministic Optimal Control Demo.XLS*, is a crop mix optimizer for a crop farm. The model is taken from the *Deterministic Demo.XLS* introduced in Chapter 2. Modification of the model for optimal control required no changes; the net present value was set as the target cell and the controls were cotton acres for each of 5 years. Sorghum acres equal total acres minus cotton acres. Constraints that prevent sorghum acres from falling below 200 acres were added via the constraint editor and total acres were constrained to be less than or equal to 1000.

If you have problems using Solver consult Excel's Help Menu, Index, Solver. There is extensive on-line help under the Solver title.

Optimizing a Deterministic Econometric Model

A deterministic simulation model can be optimized using the Solver in Excel and the optimal control techniques presented in this chapter. To demonstrate, an econometric model of the US wheat economy is optimized for two different objective functions in *Wheat Model Demo.XLS*. The model is a simple recursive model of the farm sector for wheat and can be summarized as:

$$\text{Yield}_t = f(\text{Price}_{t-1} \text{ or } \text{Loan}_t)$$

$$\text{Acres}_t = f(\text{Price}_{t-1} \text{ or } \text{Loan}_t \text{ and } \text{CRP}_t)$$

$$\text{Supply}_t = \text{Carry Over}_{t-1} + \text{Yield}_t * \text{Acres}_t$$

$$\text{Price}_t = f(\text{Supply}_t)$$

$$\text{Demand}_t = f(\text{Price}_t \text{ and } \text{Income}_t)$$

$$\text{Carry Over}_t = \text{Supply}_t - \text{Demand}_t$$

Policy variables are loan rates and conservation reserve program acres (CRP). Deterministically the model solves for all of the endogenous variables, given alternative values for loan rates and CRP acres. The model is simulated for 5 years to demonstrate the technique in the demo program.

Optimal control theory can be applied to the model to find the optimal combination and levels of loan rates and CRPs for wheat, given any objective function. More complex models can be controlled, as this technique has been used to optimize a multiple crop model by Richardson and Ray.

The first step in applying optimal control to an econometric model is to validate it thoroughly. Signs on all coefficients must conform to theory, otherwise the optimizer will wonder off into a black hole of irrelevant answers. Test the model's stability under alternative values for the exogenous variables to insure stability and reasonableness of the answers.

The next step for optimizing a simulation model is to specify the objective function to optimize. One possible objective function is to maximize the sum of consumer and producer surplus (CS and PS) minus total government payments (GP) over the six year planning horizon. Such a function can be written as:

$$\text{Max } J = \sum_{i=1}^6 r_i \text{CS}_i + r_i \text{PS}_i - r_i \text{GP}_i$$

where r_i equals the present value ratio of $1/(1 + \text{discount rate}_i)^i$

This particular function is specified in rows 7-12 of the demo wheat model. An alternative objective function could be a trajectory of annual prices that the decision makers want to observe. A price trajectory objective function for a six year planning horizon could be simulated as:

$$\text{Min } J = \sum_{i=1}^6 r_i (P_i - \text{PT}_i)^2$$

where r_i is the present value ratio or any non-zero weight to be applied to the squared difference between price (P_i) and the price trajectory (PT_i).

This particular objective function was specified for the wheat demo model, as indicated in the Solver insert. The objective value is calculated in cell C20 and then cell referenced to cell B7 for the Solver in the demo program.

The last step to optimizing a simulation model is to specify the control variables. The control variables, for the model can be any exogenous variable in the system. In the case of the wheat model, annual loan rates and CRP acreage levels for 2002-2006 are the control variables, thus creating 10 actual controls. The control variables are programmed in Solver by specifying the controls as the Changing Cells values in the Solver Parameters dialog box.

Restrictions on the control variables can be implemented through the constraints in the Excel Solver. For example, minimum and maximum ranges can be applied to the annual loan rate and CRP levels. A minimum CRP level could be set equal to the acres that are under contract for each year. The maximum could be set at an upper limit expected to be politically acceptable. For a five year model, the constraints on CRP acres could be:

$$\text{Min } U_i = [20, 15, 15, 15, 15]$$

$$\text{Max } U_i = [25, 26, 26, 26, 26]$$

Similar constraints can be applied to each of the control variables.

The demo wheat model is solved for a five year price trajectory by changing CRP levels and loan rate levels. The trajectory for prices is 2.5, 2.6, 2.7, 2.8, 2.9 for 2002-2006 (row 15) and initial loan rates and CRP values are set in rows 23 and 24 at their respective minimums. The Solver optimized the objective function and returns the answer of \$2.00/bu loan rate in all years and CRP levels equal to 22.8, 17.0, 18.0, 19.2, and 20.5 million acres. Alternative starting values were tried with the same answer on the controls being returned each time.

Optimizing a Stochastic Econometric Model

A stochastic simulation model can be optimized using Excel's Solver. This can best be demonstrated using a simultaneous equation model specified as:

$$Q_{St} = a + b P_t + cX_t + e$$

$$Q_{Dt} = a + b P_t + cY_t + e$$

$$Q_{St} = Q_{Dt}$$

Price is solved so that the quantity supplied equals the quantity demanded in each year. The SimSolver option in the Simetar Simulation Engine allows you to solve the model stochastically. (The demonstration program Wheat Sim Solve Demo.XLS is provided to show how this works.) To use the SimSolve option, first use Tools > Solver and specify the objective value to minimize, the prices as the change variables and any constraints desired. Once the Solver runs, the parameters are set and can be used by Simetar. Next open the Simetar Simulation Engine, select the Output Variables (prices and other variables), specify the number of iterations, select the Incorporate Solver option, and select SIMULATE. Excel will take some time to simulate and solve the model. The way that SimSolver works is that for each iteration, the random values are

drawn and added to the equations (thus affecting the intercepts), then the Solver takes over and optimizes the objective function by solving for the optimal prices, next the results are fed to Simetar as output variables and the next iteration begins.

For a simultaneous equation model annual prices are the change variables or the controls. The Solver systematically tries alternative prices until the price that causes quantity supplied to equal quantity demanded is discovered. In Figure 12.4 the initial guess could be P_1 so $q_1 - q_2$ is the excess demand. The Solver guesses at a second price P_2 so excess demand is $q_3 - q_4$. Based on these control values and objective function values, P_1 is rejected and a new price is calculated at P_3 . The process is continued until the equilibrium price is discovered.

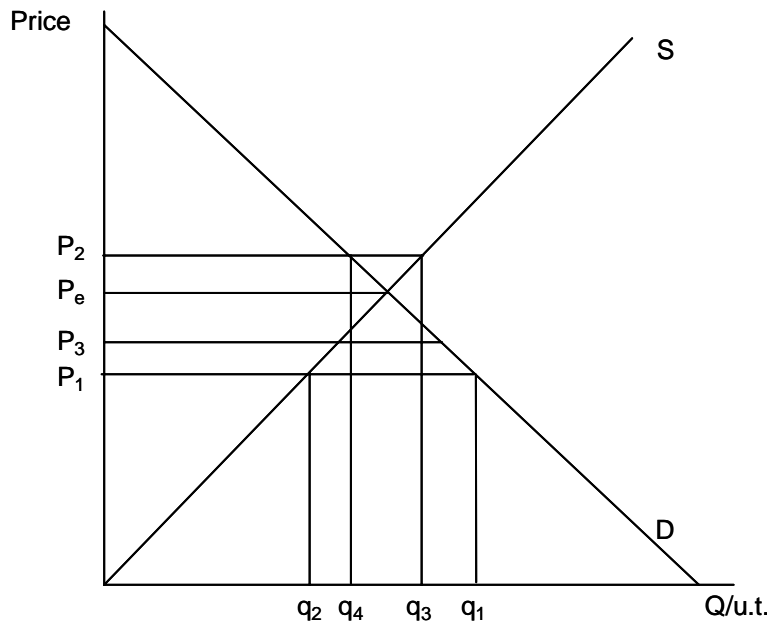


Figure 12.4. Example of an Optimal Control Problem Solving for Equilibrium Price.

An example of optimizing a stochastic simultaneous model is provided in Wheat Sim Solve Demo.XLS. The objective function is:

$$J = \sum_{t=1}^{10} (Q_{St} - Q_{Dt})^2$$

Solver is programmed to make the objective value equal zero by changing the annual prices as the change variables or cells. Initial starting values for prices must be provided and alternative starting values should be used to insure the solver is finding a global optimum. In the demo program, a table summarizing mean prices based on alternative starting values is provided. Note that mean simulated prices are the same whether the start values were \$1, \$2, \$5, or \$4 per bu.

References

- Judd, K.L. Numerical Methods in Economics. Cambridge, MA: The MIT Press, 1999.
- Miranda, M.J. and P.L. Fackler. Applied Computational Economics and Finance. Cambridge, MA: The MIT Press, 2002.
- Outlaw, Joe L. "The Impacts of Regional Structural Changes on the Supply Response of Milk in the United States." Department of Agricultural Economics, Texas A&M University, Ph.D. Dissertation, 1992.
- Richardson, J.W., D.E. Ray, J.N. Trapp. "Illustrative Applications of Optimal Control Theory Techniques to Problems in Agricultural Economics." Oklahoma State University, Agricultural Experiment Station, Bulletin B-739, January 1979.
- Richardson, J.W. "An Application of Optimal Control Theory to Agricultural Policy Analysis." Oklahoma State University, Department of Agricultural, Ph.D. Dissertation, 1978.
- Richardson, J.W. and D.E. Ray. "Commodity Programs and Control Theory." American Journal Agricultural Economics, 64(1982): 28-38.

Chapter 13

Special Problems in Modeling Risk

This chapter provides examples of how to handle particular problems in simulation models. Business analysis models must deal with risky cash flows, which present particular problems for simulating pro forma financial statements. The first topic in this chapter deals with simulating risky cash flows. In subsequent sections, solutions are provided for simulating income tax, calculations, debt amortization, and calculating net present value (NPV) and internal rate of return (IRR). The problems of calculating a fair insurance premium and replacement of machinery are also covered in the chapter. A comprehensive farm simulation model is presented to demonstrate how to apply most of these features into a firm level risk model. The last section of this chapter demonstrates how an econometric model of a commodity can be developed and simulated.

Simulating Risky Cash Flows

Businesses faced with risk can observe negative cash flows, even though they may not be observed in the deterministic solution of a simulation model. Failure to incorporate the possibility of a negative cash flow into the pro forma financial tables will cause the model to fail when used for stochastic simulation. Additions to the financial statements necessary to handle negative cash flows are outlined in this section.

An abbreviated set of financial statements would look like the example on the next page with the assigned row and column values for referencing. The cells with “-” indicate a non-zero value normally appears in the cell. The formulas in the cells show how to calculate cash flow deficits and how to handle deficits in financial statements.

The first year of simulation has no interest on carryin short-term debt (cell B9), unless the business is allowed to have carryin debt. The cell C9 is for calculating interest on carryover debt from the first year. The formula in cell C9 calculates interest if there is a balance for carryover debt for year 1 (cell B33) using the appropriate interest rate. The formula for calculating interest for a carryover loan must consider the length of the loan, usually less than a year, i.e.,

$$\text{interest due} = \text{cash flow deficit for previous year} * (\text{annual operating interest rate} / 365 \text{ days}) * \text{number of days carryover money is borrowed}$$

The beginning cash each year in the Cash Flow Statement (line 15) is the positive ending cash from the year before. This value is found in the first line of Assets in the Balance Sheet (line 27). Thus the beginning cash reserves in year 2 (C15) equals cash on hand Dec. 31 in cell B27 and so on. The Cash Flow Statement is where the model must show repayment of the short-term loan to meet cash flow deficits (line 20). Principal payments and family living withdrawals appear in the Cash Flow as outflows; these are not expenses in the Income statement because they are not tax deductions. In cell C20, enter the value of the cash flow deficit, if any existed, in the previous year. The value of the previous year’s cash flow deficit is in row 33.

	<u>A</u>	<u>B</u>	<u>C</u>	<u>D</u>
	<u>Income Statement</u>	<u>2000</u>	<u>2001</u>	<u>2002</u>
1	<u>Income Statement</u>			
2	Receipts			
3	Total Receipts	= B2	= C2	= D2
4	Expenses			
5	All Non-Interest Expenses	-	-	-
6	Interest for Land Loans	-	-	-
7	Interest for Machinery Loans	-	-	-
8	Interest for Operating Loans	-	-	-
9	Interest for Carry over Loans	= 0.0	= B33 * iRate	= C33 * iRate
10	Total Expenses	= SUM (B5:B9)	= SUM (C5:C9)	= SUM (D5:D9)
11				
12	Net Cash Income	= B3 - B10	= C3 - C10	= D3 - D10
13				
14	<u>Cash Flow Statement</u>			
15	Beginning Cash Jan. 1	= Initial Value	= B27	= C27
16				
17	Net Cash Income	= B12	= C12	= D12
18	Other Inflows	-	-	-
19	Total Inflows	= B17 + B18 + B15	= C17 + C18 + C15	= D15 + D17 + D18
20	Repay cash flow deficits	= 0.0	= B33	= C33
21	Family Withdrawals	-	-	-
22	Total Outflows	= B20 + B21	= C20 + C21	= D20 + D21
23	Ending Cash Balance Dec. 31	= B19 - B22	= C19 - C22	= D19 - D22
24				
25	<u>Balance Sheet</u>			
26	Assets Dec. 31			
27	Cash Reserves	= IF (B23 >= 0, B23, 0)	= IF (C23 >= 0, C23, 0)	= IF (D23 >= 0, D23, 0)
28	Land Value	-	-	-
29	Machinery	-	-	-
30	Other Assets	-	-	-
31	Total Assets	= SUM (B27:B30)	= SUM (C27:C30)	= SUM (D27:D30)
32	Liabilities Dec. 31			
33	Cash Flow Deficits	= IF (B23 < 0, (-1 * B23), 0)	= IF (C23 < 0, (-1 * C23), 0)	= IF (D23 < 0, (-1 * D23), 0)
34	Land	-	-	-
35	Machinery	-	-	-
36	Total Debts	= SUM (B33:B35)	= SUM (C33:C35)	= SUM (D33:D35)
37	Net Worth	= B31 - B36	= C31 - C36	= D31 - D36

The Balance Sheet is the next place where the financial statements are augmented to simulate negative cash flows. In the assets side of the Balance Sheet only positive cash reserves may appear. This is done by using an IF() statement as indicated in cell B27. The IF() statement only allows positive cash balances to be treated as an asset, and zeros appear in row 27 when there is a negative cash balance. Negative cash balances enter the Balance Sheet in row 33 as liabilities. The IF() statements in row 33 insure that the row has either zeros or positive liabilities that equal the cash flow deficits for the current year. The short-term loans to meet cash flow deficits in line 33 are the values used in the next year to: (a) calculate interest due for carryover loans in line 9 and (b) calculate the principal to repay cash flow deficits in line 20.

The pro forma financial statements require only the changes outlined here to insure the proper handling of cash flow deficits. An added feature to using this procedure for handling cash flow deficits is that the ending cash balance can be positive or negative. As a result this variable can be used as an output variable for simulation and one can calculate the probability of having negative ending cash reserves in each year.

An example Excel model which uses this procedure for handling negative cash balances is the Feedlot Demo.XLS. The DEMOPROFORMA provides one realization for the feedlot model to demonstrate how the feedlot's financial statements appear when there is a cash flow deficit. In the example, a cash flow deficit occurs in 2003 and in 2005. The deficit in 2003 (line 33) appears also in the liabilities (line 40) as a positive value. In the next year the interest for the loan to cover this cash flow deficit appears in the Income Statement (line 20). The principal payment in 2004 to repay this loan appears in the Cash Flow Statement (line 31).

Income Taxes

Federal income tax schedules for both a corporation and an individual are used in Income Tax Demo.XLS to demonstrate how to simulate income taxes. (Also refer to Business Demo.XLS for an example of tax calculation in a business simulation model.) To simulate the annual activities of a business it is generally assumed that the income tax provisions remain constant. Income tax provisions are usually not projected to change over time. As a result, the actual IRS code for the most recent year is what you will have to use for each year of the planning horizon even though the tax rates could change in the future. An actual income tax schedule is provided as an example of an IRS tax schedule in Table 2 of Income Tax Demo.XLS.

Two steps to simulate federal income taxes for a corporation are:

- Calculate taxable income, such as the values in rows 11-14 of Income Tax Demo.XLS. This value is net cash income minus deductions (such as depreciation) and standard deduction (if sole proprietor).
- Calculate income tax, such as the values in rows 21-37 of Income Tax Demo.XLS. Use the taxable income and the income tax brackets to calculate the federal income tax due for each year.

The formula used to calculate taxes is:

$$\text{Taxes Due} = \text{base tax for the bracket} + \text{marginal tax rate} * \left(\text{taxable income} - \text{minimum income for bracket} \right)$$

Given the income tax formula and the income tax schedule (Table 2 in Income Tax Demo.XLS) the Excel command =VLOOKUP should be used to obtain the three unknowns for the above formula. Assume taxable income is \$88,000, then the =VLOOKUP function is used as follows:

- =VLOOKUP (\$88,000.0, A24:D31, 3) returns the value in column 3 of \$13,750 for taxes due on income earned up to \$75,000.
- =VLOOKUP (\$88,000.0, A24:D31, 4) gives the marginal tax rate for income earned between \$75,000 and \$88,000 or 0.340.
- =VLOOKUP (\$88,000.0, A24:D31, 1) returns the value in column 1 which matches the minimum income for the tax bracket or \$75,000.
- These three values are then used to calculate taxes using the above formula. See rows 33-37 of Income Tax Demo.XLS for an example of how this is programmed.

When simulating a sole proprietor business include tables and equations to simulate self-employment, medicare, and federal income taxes. An example of calculating these taxes is presented in Income Tax Demo.XLS, with detailed description of the variables used to simulate each tax (rows 41-93).

Debt Amortization

The Excel command PMT is essential in a business analysis simulation model. An example of the Excel PMT function is available in Annual Payment Demo.XLS and Monthly Payment Demo.XLS. The example spreadsheet demonstrates how to use Excel's PMT function to calculate how much interest and principal is paid each year and how to update the remaining balance of a loan.

An easily overlooked feature in simulation models is where to account for principal and interest. Interest payments for all loans appear as a cash expense in the income statement. Principal payments are not cash costs but are cash outflows so they must appear in the cash flow statement. The remaining debt for each loan appears in the liability side of the balance sheet so it is important to calculate the remaining debt on the loan. The Excel spreadsheet Annual Payment Demo.XLS shows how to use the PMT formula to calculate all of these values that are essential for a business simulation model.

Other examples of debt repayment in simulation models included on the CD are: Investment Management Demo.XLS, Business Demo.XLS, and Bank Demo.XLS.

Net Present Value (NPV) and Internal Rate of Return (IRR)

Net Present Value is generally a key output variable which is used to summarize the net returns for a multi-year investment into a single variable. This means that the annual net return earned over several years is condensed to one number. Additionally, the net change in real net worth is incorporated into the NPV. In a stochastic model, the NPV is calculated many times, once for each iteration or set of random values. The final summary of output variables is thus a sample of 100 or more NPV's suitable for a CDF graph and for comparing alternative scenarios.

There are many formulas for calculating NPV, as demonstrated by Robison and Barry. Variations arise from the different purposes NPV is put to and different definitions for the components, as well as differences in the discount rates. Two different NPV formulas are presented and demonstrated here.

– NPV for Low Cash Outlay Investments

Purchases of stocks and bonds fit into this category of investments, as the per share cash outlay may be low because the investment is infinitely divisible. A second component of this NPV formula is that it explicitly accounts for when the changes in net worth occur. The NPV formula is:

$$NPV_1 = \sum_{t=1}^T \left(\frac{NR_t}{(1+i)^t} \right) + \sum_{t=1}^T \frac{CNW_t}{(1+i)^t}$$

where CNW represents the nominal change in net worth from one period to the next (annual), NR_t represents the net return extracted from the investment (say, dividends paid) each period t , and i is the discount rate for one period (year).

- NR_t is subject to risk caused by stochastic variables in the model.
- CNW_t is a function of retained (or reinvested) earnings and changes in market value and debt repayment. Annual changes in market value can be due to stochastic forces.

– NPV for Large Capital Outlay Investments

Purchases of large businesses, such as farms and ranches, are lumpy with huge capital outlays. Annual changes in net worth for these types of investments come from debt repayment (which is on a constant schedule) and fairly constant annual inflation rates for land. Annual net returns vary widely from year-to-year due to stochastic forces and thus "when" these net returns occur is important to the NPV formula. The NPV formula in this case is:

$$NPV_2 = -B + \sum_{t=1}^T \left(\frac{NR_t}{(1+i)^t} \right) + \frac{NW_T}{(1+i)^T}$$

where B is the initial cash outlay or net worth after purchasing the business, NR_t is the annual net return withdrawn from the business, NW_T is nominal net worth in the last year (T) of the planning horizon, and i is the annual discount rate.

Both NPV formulas are demonstrated in the Net Present Value Demo.XLS spreadsheet. Simulate the spreadsheet with Simetar specifying the two NPV's as the key output variables to see how these formulas work.

– Internal Rate of Return (IRR)

IRR is the interest rate (i) which causes the net present value of an income stream to equal zero. For each income stream there could be multiple i values thus IRR is not a perfect summary statistic. Multiple i values are certain to occur if the income stream changes signs over the planning horizon.

The income stream used for IRR in a simulation model is the same as the one used for NPV_2 . We again calculate NPV and IRR at the end of each iteration so we get an "estimate of the empirical pdf for IRR." That is the purpose of simulation.

IRR: solve for the value i which causes the following income stream to equal zero:

$$0 = -B_0 + \sum_{t=1}^T \left(\frac{NR_t}{(1+i)^t} \right) + \frac{NW_T}{(1+i)^T}$$

Excel will not calculate the IRR reliably if the $-B_0$ is ignored or left out of the equation. Excel calculates IRR using =IRR (guess, range of values) and is demonstrated in the Net Present Value Demo.XLS spreadsheet. NPV and IRR calculations are also integral calculations in Investment Management Demo.XLS, Bank Demo.XLS, Feedlot Demo.XLS,

Business Demo.XLS, and Deterministic Demo.XLS.

– Capital Investment Analyzer

An after-tax NPV and IRR calculator is provided in Net Present Value Internal Rate of Return Demo.XLS. This Excel spreadsheet was developed to teach business managers how to use NPV and IRR for comparing investments. As a result the program is in finished format with colors, protected cells, conditional colors for cells, graphics, and some on-line documentation (cells with comments). The cells for the actual calculation of IRR would be hidden from the user in black cells, but for this example they are gray.

The Net Present Value Internal Rate of Return Demo.XLS spreadsheet is deterministic so the user must specify the average cash receipts and cash expenses, for each year (or month) of the investment analysis. These values are entered in a table provided for this purpose, but using a general format that allows the user to enter the data in any order. The calculated NPV and IRR values remain in view at all times so the user can see how a change in the discount rate, the number of years in the analysis, the tax rate, the depreciation life, the sale price, or the financial assumptions affects the output variables.

Take note of how Excel was programmed to calculate the IRR. Even with all of the precautions in programming the IRR section of the program, the IRR sometimes fails to find a unique solution. This type of result is usually due to an unrealistic investment cost relative to the annual inflows and outflows or the user does not enter a down payment for the initial cash outlay.

Estimating Insurance Premiums

Insurance premium rates can be estimated using simulation. Simulation may be the only way to do this when insufficient information is available for direct calculation of losses. An example of how to estimate the insurance premium rates for an average producer is provided in Insurance Premium Demo.XLS.

To estimate insurance premium rates first determine the probable payout (indemnities). The premium is set so that it fully covers all payouts plus the cost of doing business and the profits the insurance firm requires. In the example it is assumed that the historical yields for one producer are representative of the population of producers to be insured.

The stochastic yield in Step 4 is used to calculate the lost yield for seven different insurance programs in Step 6. Lost yield equals the difference between simulated yield and insured yield if the stochastic yield is less than the insured yield or in Excel terms it is:

$$= \text{IF} \left(\begin{array}{c} \text{simulated} \\ \text{yield} \end{array} < \begin{array}{c} \text{insured} \\ \text{yield} \end{array}, \left(\begin{array}{c} \text{insured} \\ \text{yield} \end{array} - \begin{array}{c} \text{simulated} \\ \text{yield} \end{array} \right), 0 \right)$$

Seven different levels of yield loss fractions were analyzed, making up seven different insurance programs. The first insured yield loss fraction equals 50 percent of the historical average yield. The last insured yield loss fraction equals 90 percent of the historical average yield.

The indemnity paid out for each insurance program equals the lost yield times a fixed indemnity price (\$0.60/lb. in this case). The seven indemnity payments (one for each program) are in cells F75-F81 and change as the stochastic yield changes when F9 is pressed. This procedure for simulating alternative insurance options uses the same random yield across all seven insurance programs so the results can be directly compared.

The key output variables (KOV) for the insurance simulation model are the indemnity payments (F75-F81). The statistical summary results of a 500 iteration analysis (lines 98-103) show that the average indemnity is \$9.91/acre for the 50 percent yield coverage option. At the 90 percent yield coverage level, the average indemnity is \$25.44/acre. These average indemnities represent a minimum insurance premium rate before adding in the profit and the cost of administering the insurance program. The higher the rate of coverage (greater the percent of average yield covered) the higher the premium.

Additionally, counter variables are used as KOVs to indicate the frequency of insurance payments (E75-E81). A counter is created by using an IF statement to have a 1 if an indemnity is paid and a 0 if no payment is made or

=IF(D75 > 0, 1, 0)

The statistical summary for a counter indicates the probability of indemnities. For example, the results (lines 98-103) indicate there is a 23.8 percent chance of an indemnity at the 50 percent coverage level and a 45.4 percent indemnity at the 90 percent yield coverage level.

Machinery Replacement

One of the more difficult problems in simulating a business is the replacement of machinery. There are several ways not to replace machinery, such as: assume ten percent of the inventory is replaced each year or assume the inventory is all new at the outset of the planning horizon and all of it is replaced N years later. Machinery replacement is a lumpy expense for businesses because each machine has a different cost and expected life in the business. Also, businesses generally do not replace everything at once, because the economic life of machinery on a farm varies by machine.

The best solution I have found for simulating machinery replacement is to itemize the machinery complement and replace each item at the end of its “economic life in the business.” The information required to use this method is listed below. This information must be obtained for each machine at the start of the simulation period (at t=0):

- machine name
- year placed into use
- original purchase price (excluding trade-in)
- current market value of the machine
- current price of a replacement machine
- number of years machine will be used on the farm

In a simulation model, a machine is replaced only if it has passed its economic life on the farm. The value of a machine at the time it is replaced (traded-off) after N years is the current value at $t=0$ discounted to the year the machine is replaced. The cost of a new machine is the $t=0$ cost of a replacement inflated for N years. The net cost of the replacement is the updated cost less the trade-in value.

Financing machinery replacements can be handled on a machine-by-machine basis with a loan for each purchase. Alternatively the model can sum the new loan requirements and finance one new machinery loan for each year machinery is replaced. I prefer the later myself to reduce the number of loans to simulate. Additionally, with one loan per year your model can easily repay the principal early if surplus cash is generated.

To finance machinery replacements the model must have a value for the minimum down payment on machinery loans. First compare the trade-in value to the down payment requirement; if the trade-in is less than the minimum down payment, the deficit must come from cash reserves. Finance the remainder (cost less minimum down payment) at the prevailing terms for machinery loans (number of years and interest rate). If the trade-in value exceeds the minimum down payment, then finance the difference and do not pay a cash difference. Annual interest payments appear in the Income Statement and principal payments appear in the Cash Flow. The remaining debt and updated value of machinery appear in the Balance Sheet.

Refer to Machinery Demo.XLS for an example of how to simulate machinery replacement for a business. This demo program is set to finance the purchase of replacement machinery for 5 years using fixed interest rate mortgages. Ten different machinery items are included in the demo. Change the information for any or all of the machines to observe the impacts on cash outflow requirements, interest costs and the firm's balance sheet. Adding more machines to the spreadsheet can be done by inserting columns before column M, entering the new machines and copying the equations for the new machines. All of the equations should work, but it is your responsibility to check them.

Using this method for replacing machinery will produce an uneven cash flow requirement for machinery replacement. The uneven cash flow requirements need to be evaluated after solving for the replacement of each machine to see if “too much” machinery is to be replaced in any given year. Spread out these cash flow requirements by changing the economic life of individual machines until the cash flow requirements are reasonable for the business being analyzed.

Farm Level Simulation Model

A 10 year, whole farm simulation model is demonstrated in Farm Simulator Demo.XLS. The model is developed using the type of calculations in FLIPSIM (Richardson and Nixon). The model is capable of simulating one, two, or three crops. Input data required of the user are in bold in the Model and the Stochastic worksheets. Historical prices and yields for the three crops on the farm are entered in the Stochastic worksheet. The historical prices and yields are used to develop and simulate a MVE distribution for prices and yields. Charts showing the historical and stochastic prices and yields are provided at the bottom of the Stochastic worksheet.

Projected inflation rates for variable inputs are also entered in the Stochastic worksheet. Separate inflation rates for fuel, labor, and other inputs are entered so the model can simulate the effects of alternative inflation rates. Annual interest rates for financing operating costs are also included as input for the whole farm simulation model.

All other inputs necessary to simulate a crop farm are entered at the top (rows 8-71) of the Model worksheet. Assets and liabilities, family living, fixed costs, and other income are required for a farm risk analysis. Budgets for each crop are included along with base acres, payment yields, and price wedges. The price wedges are used to localize stochastic national prices for each crop. Annual values must be provided by crop for: planted acres, mean crop yields, mean crop prices, loan rates, direct payment rates, and target prices. Planted acres and mean yields can come from the producer or the farm plan being analyzed. Average annual crop prices can be obtained from FAPRI or developed from other sources. Policy values for 2002-2007 are specified in the 2002 farm bill. Policy values for 2008-2011 are assumed to remain at their 2007 levels.

The whole farm simulation model is made up of several sections where the calculations are done. A detailed income statement is provided in rows 75-104. The values reported in the statement are calculated in specific sections for each crop. A cash flow statement is provided in rows 106-123 for the express purpose of tracking ending cash reserves. The farm's balance sheet is in rows 125-136 to calculate net worth.

Values found in the cash flow and balance sheet are used to calculate net present value, one of the KOVs for the farm (rows 140-147). Three more KOVs useful for evaluating a farm with risky cash flows are: probability of a cash flow deficit, probability of negative cash flows, and probability of losing real net worth. Definitions for these variables are:

- Cash flow deficits occur when cash outlays exceed net cash farm income.
- Negative cash flows occur when beginning cash plus net cash income and other cash earnings are less than total cash outflows. This is the amount of the operating loan, which must be refinanced or carried over.
- Firms lose real net worth when the present value of ending net worth is less than beginning net worth.

All three of the probability KOVs are simulated using counter variables in rows 149-151. When the condition is true the counter is 1, else it is zero. The probability of a cash flow deficit can be simulated for each year using 10 counter variables. This type of KOV facilitates pinpointing years that the firm will likely need to refinance its operating loan.

The section of the model where crop receipts, expenses, and government payments to crop are calculated, is in rows 154-189. Stochastic production equals planted acres times stochastic yields. National prices are adjusted by the local wedge (or basis) to derive stochastic local prices. Market receipts equals the product of production and local prices. LDP, direct and CCP government payments are calculated using annual loan rates, target prices and direct payment rates. Base acres and DP and CCP payment yields are used for these calculations. An AWP fraction is used to convert the stochastic national price to a stochastic adjusted world price used to calculate the LDP rate.

Costs of production for each crop are calculated using annual inflation rates and the crop budgets. For crop 1 the costs of production are calculated in rows 272-285. The same steps are repeated for crops 2 and 3. Results from this section are used in the Income Statement. Other costs for the farm are calculated next. Of particular note are interest costs for the operating loan and the operating loan carryover, i.e., for financing cash flow deficits.

Interest on cash reserves are calculated so they can enter the cash flow statement as earned interest. Family living costs or withdrawals are calculated so these cash outflows can appear in the cash flow statement. Land value is calculated using the user's assumed annual inflation rates for land. Land value shows up in the asset side of the Balance Sheet.

The calculations necessary to amortize the original land loan and a beginning farmer loan are included in the model in rows 369-442. Annual interest costs in these schedules are used in the Income Statement to calculate total costs. Annual principal payments for the two loans appear in the Cash Flow statement as these are not production expenses.

The final section of the model includes the calculations to compute annual federal income taxes for the farm. A corporate income tax schedule is used to compute federal income taxes.

At the top of the model worksheet is a KOV table, rows 10-31. The table is cell referenced to each of the KOVs found in the model. When the farm model is simulated, the user can highlight the same KOVs every time and they are always in the same order. Given that the KOVs are always in the same order, you can develop summary tables for the outputs and permanent charts to interpret the results. An example of such a table is provided in columns AB – AK in SimData1.

Simulating an Econometric Model

Recursive supply demand models for crops and livestock can be estimated and simulated using tools provided in Simetar. A recursive supply demand model for the US soybean sector is estimated and simulated in this section. The Excel demo program is named Soybean Model Demo.XLS.

The objective of the model is to simulate annual prices, ending stocks, and the present value of total government payments, 2002-2004, for the 1996 and the 2002 farm programs. The objective requires that the model incorporate different types of farm programs and endogenously solve for price.

The hypothesized equations for the model are:

$$\text{Planted Acres}_t = f(\text{E Price}_t, \text{Planted Acres}_{t-1})$$

$$\text{Harvested Acres}_t = f(\text{E Price}_t, \text{Planted Acres}_t)$$

$$\text{Yield}_t = f(\text{E Price}_t, \text{Yield}_{t-1}, \text{Trend}_t)$$

$$\text{Domestic Use}_t = f(\text{Price}_t, \text{Domestic Use}_{t-1}, \text{Income/Pop}_t)$$

$$\text{Exports}_t = f(\text{Price}_t, \text{Exports}_{t-1}, \text{GDP EU}_t)$$

$$\text{Ending Stocks}_t = f(\text{Price}_t, \text{Stocks}_{t-1})$$

$$\text{Define Eprice}_t = \text{Max} [\text{Price}_{t-1} \text{ or Loan Rate}_t]$$

Based on these equations historical data for the endogenous and exogenous variables, 1980-2001, were obtained and added to the Data worksheet in the demo program.

A significant time saving step is to create a worksheet for each equation to be estimated. The variables hypothesized for each equation are cell referenced from the Data worksheet to be equation's worksheet. For example the PltAcres worksheet has the values for Planted Acres_t, Eprice_t, and Planted Acres_{t-1}. The multiple regression for each equation is estimated in the equation's own worksheet so it is easy to locate.

The residuals and the standard deviation of residuals from each of the regression models are cell referenced to the Stoch worksheet. The residuals are used to define and simulate a MVN distribution of shocks to the econometric equations. The stochastic shocks for each year in the planning horizon are tested to insure they are appropriately correlated as part of the validation.

The intercept and slope parameters for the six econometric equations are cell referenced from their worksheets to the Model worksheet where the completed model is simulated. The coefficients for each of the equations are assembled in rows 24-40 of the Model worksheet. Solving for price in a recursive model is accomplished by first estimating the total demand equation, re-writing the equation so that price is a function of quantity, and then solving for price using quantity supplied.

Total demand is the sum of domestic use, exports and ending stocks. A separate total demand equation must be estimated for each year 2002-2004 as the exogenous variables in these equations change each year. The total demand is estimated as:

$$\begin{aligned} DD_t &= a + bP_t + c DP_{t-1} + d I_t + \tilde{e}_1 \\ ED_t &= k + fP_t + g ED_{t-1} + h G_t + \tilde{e}_2 \\ ES_t &= i + jP_t + \tilde{e}_3 \\ TD_t &= a + k + i + (b+f+j) P_t + cDP_{t-1} + d I_t + gED_{t-1} + h G_t + (\tilde{e}_1 + \tilde{e}_2 + \tilde{e}_3) \end{aligned}$$

$$\begin{aligned} \text{let } A &= a + k + i \\ B &= b + f + j \\ C &= cDP_{t-1} + d I_t + g ED_{t-1} + h G_t \\ \tilde{e} &= \tilde{e}_1 + \tilde{e}_2 + \tilde{e}_3 \end{aligned}$$

Then total demand can be written as:

$$\begin{aligned} TD &= A + C + B P_t + \tilde{e} \\ \text{or} \\ P_t &= \frac{A}{B} + \frac{C}{B} + \frac{1}{B} TD + \frac{\tilde{e}}{B} \end{aligned}$$

which simplifies to

$$\tilde{P}_t = \alpha + \beta Q + \beta \tilde{\epsilon}$$

$$\text{let } \alpha = \frac{A}{B} + \frac{C}{B}$$

$$\beta = \frac{1}{B}$$

The total demand is thus specified as price is a function of quantity and an error term which makes the intercept stochastic (Figure 13.1).

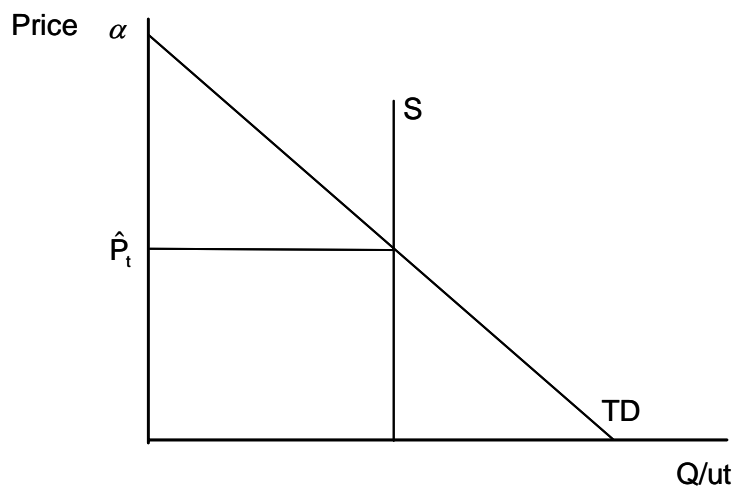


Figure 13.1. Price Dependent Demand Function and Supply Determine Price.

Given a quantity supplied (Q) in the equation, equilibrium price is found directly by solving the demand function. This is permitted because of the assumption that supply equals total demand (Figure 13.1).

In Soybean Model Demo.XLS the process of deriving total demand for each year is summarized in rows 42-52. The intercepts and slopes on price are summed in rows 44-47. The shift effects for domestic demand ($d I_t + c DP_{t-1}$) are calculated in row 44 and added to the domestic demand intercept in row 50. Similar steps are done for exports. The total demand intercept changes each year (cells E54:H54) because the exogenous variables (I , G , DP_{t-1} and ED_{t-1}) and the stochastic shocks change annually.

Because quantity supplied is a function of expected price it is determined in row 73 and then used in the price equation (row 56) to display price in the current year. This specification is consistent with the fact that the season average price is largely determined after harvest and supply is known with certainty (Figure 13.1). Supply is calculated as a sum of fixed imports (3.0 each year) and production. Stochastic shocks on yields, planted acres and harvested acre equations cause supply to be a stochastic variable in addition to responding to lagged prices and

current policy variables. It is assumed that for the simulation period the expected price is defined as:

- 1996 farm bill
E Price_t = Max [Price_{t-1}, Loan Rate_t)

- 2002 farm bill
E Price_t = Max (Price_{t-1}, Target Price_t)

Once price is calculated the value is used to simulate domestic demand and exports. Domestic use and exports (rows 76 and 77) are calculated using the stochastic price and the appropriate exogenous variables and stochastic shocks used in rows 44 and 45 to estimate total demand. Ending stocks are solved as an identity:

$$\text{Ending Stocks}_t = \text{Supply}_t - \text{Domestic Use}_t - \text{Exports}_t$$

One of the KOVs is the present value of total receipts plus government payments. Direct and counter cyclical payments are calculated under the 2002 farm bill using base acres, direct payment yields and counter cyclical payment yields. A CDF chart of the PV of total revenue under each policy is provided in the SimData worksheet. The simulated annual soybean prices are summarized using a fan graph in the SimData worksheet.

References

- Adams, Gary M. "Impact Multipliers of the U.S. Crops Sector: A Focus on the Effects of Commodity Interaction." Vol. I, II, University of Missouri-Columbia, Department of Agricultural Economics, Ph.D. Dissertation, 1994.
- Bell, D.E. and A. Schleifer, Jr. Decision Making Under Uncertainty. Cambridge, MA: Course Technology, Inc., 1995.
- Bernstein, P.L. Against the Gods: The Remarkable Story of Risk. New York: John Wiley & Sons, Inc. 1996.
- Brown, D. Scott. "A Structural Approach to Modeling the U.S. Livestock and Dairy Industries: With Emphasis on the Dynamics Within the System." University of Missouri-Columbia, Department of Agricultural Economics, Ph.D. Dissertation, 1994.
- Catlett, L.B. and J.D. Libbin. Inventory in Futures and Options Market. New York: Delmar Publishers, 1999.
- Fleisher, B. Agricultural Risk Management. Boulder, CO: Lynne Rienner Publishers, Inc., 1990.
- Gempesaw, C.M., II, and J.W. Richardson. "The Use of Simulation in Management Decision Making: The Case of Poultry and Aquaculture Production." In Simulation Systems, G.W. Zobrist and J.V. Leonard, editors. Amsterdam: Gordon and Breach Science Publishers, 2000.
- Gill, R.C., II. "A Stochastic Feasibility Study of Texas Ethanol Production: Analysis of Texas State Legislature Ethanol Subsidy Proposal." Texas A&M University, Department of Agricultural Economics, M.S. Thesis, December 2002.
- Gray, A.W. "Agribusiness Strategic Planning Under Risk." Texas A&M University, Department of Agricultural Economics, Ph.D. Dissertation, August 1998.
- Naylor, T.H. "Corporate Simulation Models." In Business Simulation for Decision Making. W.C. House, editor, New York: PBI, 1977.
- Ragsdale, C.T. Spreadsheet Modeling and Decision Analysis. Cincinnati, OH: South-Western College Publishing, 1998.
- Ray, D.E. and J.W. Richardson. "Detailed Description of POLYSIM." Agricultural Experiment Station, Oklahoma State University and USDA, Technical Bulletin, T-151, December 1978.
- Richardson, J.W. and H.P. Mapp, Jr. "Use of Probabilistic Cash Flows in Analyzing Investment Under Conditions of Risk and Uncertainty." Southern Journal Agric. Econ. Dec. 1976.

Richardson, J.W. and C.J. Nixon. "A Description of FLIPSIM V: A General Firm Level Policy Simulation Model." Bulletin 1528, Texas Agricultural Experiment Station, July 1986.

Robison, L.J. and P. Barry. Present Value Models and Investment Analysis. New Port, AL: The Academic Press, 1996.

Winston, W.L. Simulating Modeling Using @Risk©. New York: Duxbury Press, 1996.

Chapter 13 Special Problems in Modeling Risk

This chapter provides examples of how to handle particular problems in simulation models. Business analysis models must deal with risky cash flows, which present particular problems for simulating pro forma financial statements. The first topic in this chapter deals with simulating risky cash flows. In subsequent sections, solutions are provided for simulating income tax, calculations, debt amortization, and calculating net present value (NPV) and internal rate of return (IRR). The problems of calculating a fair insurance premium and replacement of machinery are also covered in the chapter. A comprehensive farm simulation model is presented to demonstrate how to apply most of these features into a firm level risk model. The last section of this chapter demonstrates how an econometric model of a commodity can be developed and simulated.

Simulating Risky Cash Flows

Businesses faced with risk can observe negative cash flows, even though they may not be observed in the deterministic solution of a simulation model. Failure to incorporate the possibility of a negative cash flow into the pro forma financial tables will cause the model to fail when used for stochastic simulation. Additions to the financial statements necessary to handle negative cash flows are outlined in this section.

An abbreviated set of financial statements would look like the example on the next page with the assigned row and column values for referencing. The cells with “-” indicate a non-zero value normally appears in the cell. The formulas in the cells show how to calculate cash flow deficits and how to handle deficits in financial statements.

The first year of simulation has no interest on carryin short-term debt (cell B9), unless the business is allowed to have carryin debt. The cell C9 is for calculating interest on carryover debt from the first year. The formula in cell C9 calculates interest if there is a balance for carryover debt for year 1 (cell B33) using the appropriate interest rate. The formula for calculating interest for a carryover loan must consider the length of the loan, usually less than a year, i.e.,

$$\text{interest due} = \text{cash flow deficit for previous year} * (\text{annual operating interest rate} / 365 \text{ days}) * \text{number of days carryover money is borrowed}$$

The beginning cash each year in the Cash Flow Statement (line 15) is the positive ending cash from the year before. This value is found in the first line of Assets in the Balance Sheet (line 27). Thus the beginning cash reserves in year 2 (C15) equals cash on hand Dec. 31 in cell B27 and so on. The Cash Flow Statement is where the model must show repayment of the short-term loan to meet cash flow deficits (line 20). Principal payments and family living withdrawals appear in the Cash Flow as outflows; these are not expenses in the Income statement because they are not tax deductions. In cell C20, enter the value of the cash flow deficit, if any existed, in the previous year. The value of the previous year’s cash flow deficit is in row 33.

	<u>A</u>	<u>B</u>	<u>C</u>	<u>D</u>
	<u>Income Statement</u>	<u>2000</u>	<u>2001</u>	<u>2002</u>
1	<u>Income Statement</u>			
2	Receipts			
3	Total Receipts	= B2	= C2	= D2
4	Expenses			
5	All Non-Interest Expenses	-	-	-
6	Interest for Land Loans	-	-	-
7	Interest for Machinery Loans	-	-	-
8	Interest for Operating Loans	-	-	-
9	Interest for Carry over Loans	= 0.0	= B33 * iRate	= C33 * iRate
10	Total Expenses	= SUM (B5:B9)	= SUM (C5:C9)	= SUM (D5:D9)
11				
12	Net Cash Income	= B3 - B10	= C3 - C10	= D3 - D10
13				
14	<u>Cash Flow Statement</u>			
15	Beginning Cash Jan. 1	= Initial Value	= B27	= C27
16				
17	Net Cash Income	= B12	= C12	= D12
18	Other Inflows	-	-	-
19	Total Inflows	= B17 + B18 + B15	= C17 + C18 + C15	= D15 + D17 + D18
20	Repay cash flow deficits	= 0.0	= B33	= C33
21	Family Withdrawals	-	-	-
22	Total Outflows	= B20 + B21	= C20 + C21	= D20 + D21
23	Ending Cash Balance Dec. 31	= B19 - B22	= C19 - C22	= D19 - D22
24				
25	<u>Balance Sheet</u>			
26	Assets Dec. 31			
27	Cash Reserves	= IF (B23 >= 0, B23, 0)	= IF (C23 >= 0, C23, 0)	= IF (D23 >= 0, D23, 0)
28	Land Value	-	-	-
29	Machinery	-	-	-
30	Other Assets	-	-	-
31	Total Assets	= SUM (B27:B30)	= SUM (C27:C30)	= SUM (D27:D30)
32	Liabilities Dec. 31			
33	Cash Flow Deficits	= IF (B23 < 0, (-1 * B23), 0)	= IF (C23 < 0, (-1 * C23), 0)	= IF (D23 < 0, (-1 * D23), 0)
34	Land	-	-	-
35	Machinery	-	-	-
36	Total Debts	= SUM (B33:B35)	= SUM (C33:C35)	= SUM (D33:D35)
37	Net Worth	= B31 - B36	= C31 - C36	= D31 - D36

The Balance Sheet is the next place where the financial statements are augmented to simulate negative cash flows. In the assets side of the Balance Sheet only positive cash reserves may appear. This is done by using an IF() statement as indicated in cell B27. The IF() statement only allows positive cash balances to be treated as an asset, and zeros appear in row 27 when there is a negative cash balance. Negative cash balances enter the Balance Sheet in row 33 as liabilities. The IF() statements in row 33 insure that the row has either zeros or positive liabilities that equal the cash flow deficits for the current year. The short-term loans to meet cash flow deficits in line 33 are the values used in the next year to: (a) calculate interest due for carryover loans in line 9 and (b) calculate the principal to repay cash flow deficits in line 20.

The pro forma financial statements require only the changes outlined here to insure the proper handling of cash flow deficits. An added feature to using this procedure for handling cash flow deficits is that the ending cash balance can be positive or negative. As a result this variable can be used as an output variable for simulation and one can calculate the probability of having negative ending cash reserves in each year.

An example Excel model which uses this procedure for handling negative cash balances is the Feedlot Demo.XLS. The DEMOPROFORMA provides one realization for the feedlot model to demonstrate how the feedlot's financial statements appear when there is a cash flow deficit. In the example, a cash flow deficit occurs in 2003 and in 2005. The deficit in 2003 (line 33) appears also in the liabilities (line 40) as a positive value. In the next year the interest for the loan to cover this cash flow deficit appears in the Income Statement (line 20). The principal payment in 2004 to repay this loan appears in the Cash Flow Statement (line 31).

Income Taxes

Federal income tax schedules for both a corporation and an individual are used in Income Tax Demo.XLS to demonstrate how to simulate income taxes. (Also refer to Business Demo.XLS for an example of tax calculation in a business simulation model.) To simulate the annual activities of a business it is generally assumed that the income tax provisions remain constant. Income tax provisions are usually not projected to change over time. As a result, the actual IRS code for the most recent year is what you will have to use for each year of the planning horizon even though the tax rates could change in the future. An actual income tax schedule is provided as an example of an IRS tax schedule in Table 2 of Income Tax Demo.XLS.

Two steps to simulate federal income taxes for a corporation are:

- Calculate taxable income, such as the values in rows 11-14 of Income Tax Demo.XLS. This value is net cash income minus deductions (such as depreciation) and standard deduction (if sole proprietor).
- Calculate income tax, such as the values in rows 21-37 of Income Tax Demo.XLS. Use the taxable income and the income tax brackets to calculate the federal income tax due for each year.

The formula used to calculate taxes is:

$$\text{Taxes Due} = \text{base tax for the bracket} + \text{marginal tax rate} * \left(\text{taxable income} - \text{minimum income for bracket} \right)$$

Given the income tax formula and the income tax schedule (Table 2 in Income Tax Demo.XLS) the Excel command =VLOOKUP should be used to obtain the three unknowns for the above formula. Assume taxable income is \$88,000, then the =VLOOKUP function is used as follows:

- =VLOOKUP (\$88,000.0, A24:D31, 3) returns the value in column 3 of \$13,750 for taxes due on income earned up to \$75,000.
- =VLOOKUP (\$88,000.0, A24:D31, 4) gives the marginal tax rate for income earned between \$75,000 and \$88,000 or 0.340.
- =VLOOKUP (\$88,000.0, A24:D31, 1) returns the value in column 1 which matches the minimum income for the tax bracket or \$75,000.
- These three values are then used to calculate taxes using the above formula. See rows 33-37 of Income Tax Demo.XLS for an example of how this is programmed.

When simulating a sole proprietor business include tables and equations to simulate self-employment, medicare, and federal income taxes. An example of calculating these taxes is presented in Income Tax Demo.XLS, with detailed description of the variables used to simulate each tax (rows 41-93).

Debt Amortization

The Excel command PMT is essential in a business analysis simulation model. An example of the Excel PMT function is available in Annual Payment Demo.XLS and Monthly Payment Demo.XLS. The example spreadsheet demonstrates how to use Excel's PMT function to calculate how much interest and principal is paid each year and how to update the remaining balance of a loan.

An easily overlooked feature in simulation models is where to account for principal and interest. Interest payments for all loans appear as a cash expense in the income statement. Principal payments are not cash costs but are cash outflows so they must appear in the cash flow statement. The remaining debt for each loan appears in the liability side of the balance sheet so it is important to calculate the remaining debt on the loan. The Excel spreadsheet Annual Payment Demo.XLS shows how to use the PMT formula to calculate all of these values that are essential for a business simulation model.

Other examples of debt repayment in simulation models included on the CD are: Investment Management Demo.XLS, Business Demo.XLS, and Bank Demo.XLS.

Net Present Value (NPV) and Internal Rate of Return (IRR)

Net Present Value is generally a key output variable which is used to summarize the net returns for a multi-year investment into a single variable. This means that the annual net return earned over several years is condensed to one number. Additionally, the net change in real net worth is incorporated into the NPV. In a stochastic model, the NPV is calculated many times, once for each iteration or set of random values. The final summary of output variables is thus a sample of 100 or more NPV's suitable for a CDF graph and for comparing alternative scenarios.

There are many formulas for calculating NPV, as demonstrated by Robison and Barry. Variations arise from the different purposes NPV is put to and different definitions for the components, as well as differences in the discount rates. Two different NPV formulas are presented and demonstrated here.

– NPV for Low Cash Outlay Investments

Purchases of stocks and bonds fit into this category of investments, as the per share cash outlay may be low because the investment is infinitely divisible. A second component of this NPV formula is that it explicitly accounts for when the changes in net worth occur. The NPV formula is:

$$NPV_1 = \sum_{t=1}^T \left(\frac{NR_t}{(1+i)^t} \right) + \sum_{t=1}^T \frac{CNW_t}{(1+i)^t}$$

where CNW represents the nominal change in net worth from one period to the next (annual), NR_t represents the net return extracted from the investment (say, dividends paid) each period t , and i is the discount rate for one period (year).

- NR_t is subject to risk caused by stochastic variables in the model.
- CNW_t is a function of retained (or reinvested) earnings and changes in market value and debt repayment. Annual changes in market value can be due to stochastic forces.

– NPV for Large Capital Outlay Investments

Purchases of large businesses, such as farms and ranches, are lumpy with huge capital outlays. Annual changes in net worth for these types of investments come from debt repayment (which is on a constant schedule) and fairly constant annual inflation rates for land. Annual net returns vary widely from year-to-year due to stochastic forces and thus "when" these net returns occur is important to the NPV formula. The NPV formula in this case is:

$$NPV_2 = -B + \sum_{t=1}^T \left(\frac{NR_t}{(1+i)^t} \right) + \frac{NW_T}{(1+i)^T}$$

where B is the initial cash outlay or net worth after purchasing the business, NR_t is the annual net return withdrawn from the business, NW_T is nominal net worth in the last year (T) of the planning horizon, and i is the annual discount rate.

Both NPV formulas are demonstrated in the Net Present Value Demo.XLS spreadsheet. Simulate the spreadsheet with Simetar specifying the two NPV's as the key output variables to see how these formulas work.

– Internal Rate of Return (IRR)

IRR is the interest rate (i) which causes the net present value of an income stream to equal zero. For each income stream there could be multiple i values thus IRR is not a perfect summary statistic. Multiple i values are certain to occur if the income stream changes signs over the planning horizon.

The income stream used for IRR in a simulation model is the same as the one used for NPV_2 . We again calculate NPV and IRR at the end of each iteration so we get an "estimate of the empirical pdf for IRR." That is the purpose of simulation.

IRR: solve for the value i which causes the following income stream to equal zero:

$$0 = -B_0 + \sum_{t=1}^T \left(\frac{NR_t}{(1+i)^t} \right) + \frac{NW_T}{(1+i)^T}$$

Excel will not calculate the IRR reliably if the $-B_0$ is ignored or left out of the equation. Excel calculates IRR using =IRR (guess, range of values) and is demonstrated in the Net Present Value Demo.XLS spreadsheet. NPV and IRR calculations are also integral calculations in Investment Management Demo.XLS, Bank Demo.XLS, Feedlot Demo.XLS,

Business Demo.XLS, and Deterministic Demo.XLS.

– Capital Investment Analyzer

An after-tax NPV and IRR calculator is provided in Net Present Value Internal Rate of Return Demo.XLS. This Excel spreadsheet was developed to teach business managers how to use NPV and IRR for comparing investments. As a result the program is in finished format with colors, protected cells, conditional colors for cells, graphics, and some on-line documentation (cells with comments). The cells for the actual calculation of IRR would be hidden from the user in black cells, but for this example they are gray.

The Net Present Value Internal Rate of Return Demo.XLS spreadsheet is deterministic so the user must specify the average cash receipts and cash expenses, for each year (or month) of the investment analysis. These values are entered in a table provided for this purpose, but using a general format that allows the user to enter the data in any order. The calculated NPV and IRR values remain in view at all times so the user can see how a change in the discount rate, the number of years in the analysis, the tax rate, the depreciation life, the sale price, or the financial assumptions affects the output variables.

Take note of how Excel was programmed to calculate the IRR. Even with all of the precautions in programming the IRR section of the program, the IRR sometimes fails to find a unique solution. This type of result is usually due to an unrealistic investment cost relative to the annual inflows and outflows or the user does not enter a down payment for the initial cash outlay.

Estimating Insurance Premiums

Insurance premium rates can be estimated using simulation. Simulation may be the only way to do this when insufficient information is available for direct calculation of losses. An example of how to estimate the insurance premium rates for an average producer is provided in Insurance Premium Demo.XLS.

To estimate insurance premium rates first determine the probable payout (indemnities). The premium is set so that it fully covers all payouts plus the cost of doing business and the profits the insurance firm requires. In the example it is assumed that the historical yields for one producer are representative of the population of producers to be insured.

The stochastic yield in Step 4 is used to calculate the lost yield for seven different insurance programs in Step 6. Lost yield equals the difference between simulated yield and insured yield if the stochastic yield is less than the insured yield or in Excel terms it is:

$$= \text{IF} \left(\begin{array}{c} \text{simulated} \\ \text{yield} \end{array} < \begin{array}{c} \text{insured} \\ \text{yield} \end{array}, \left(\begin{array}{c} \text{insured} \\ \text{yield} \end{array} - \begin{array}{c} \text{simulated} \\ \text{yield} \end{array} \right), 0 \right)$$

Seven different levels of yield loss fractions were analyzed, making up seven different insurance programs. The first insured yield loss fraction equals 50 percent of the historical average yield. The last insured yield loss fraction equals 90 percent of the historical average yield.

The indemnity paid out for each insurance program equals the lost yield times a fixed indemnity price (\$0.60/lb. in this case). The seven indemnity payments (one for each program) are in cells F75-F81 and change as the stochastic yield changes when F9 is pressed. This procedure for simulating alternative insurance options uses the same random yield across all seven insurance programs so the results can be directly compared.

The key output variables (KOV) for the insurance simulation model are the indemnity payments (F75-F81). The statistical summary results of a 500 iteration analysis (lines 98-103) show that the average indemnity is \$9.91/acre for the 50 percent yield coverage option. At the 90 percent yield coverage level, the average indemnity is \$25.44/acre. These average indemnities represent a minimum insurance premium rate before adding in the profit and the cost of administering the insurance program. The higher the rate of coverage (greater the percent of average yield covered) the higher the premium.

Additionally, counter variables are used as KOVs to indicate the frequency of insurance payments (E75-E81). A counter is created by using an IF statement to have a 1 if an indemnity is paid and a 0 if no payment is made or

=IF(D75 > 0, 1, 0)

The statistical summary for a counter indicates the probability of indemnities. For example, the results (lines 98-103) indicate there is a 23.8 percent chance of an indemnity at the 50 percent coverage level and a 45.4 percent indemnity at the 90 percent yield coverage level.

Machinery Replacement

One of the more difficult problems in simulating a business is the replacement of machinery. There are several ways not to replace machinery, such as: assume ten percent of the inventory is replaced each year or assume the inventory is all new at the outset of the planning horizon and all of it is replaced N years later. Machinery replacement is a lumpy expense for businesses because each machine has a different cost and expected life in the business. Also, businesses generally do not replace everything at once, because the economic life of machinery on a farm varies by machine.

The best solution I have found for simulating machinery replacement is to itemize the machinery complement and replace each item at the end of its “economic life in the business.” The information required to use this method is listed below. This information must be obtained for each machine at the start of the simulation period (at t=0):

- machine name
- year placed into use
- original purchase price (excluding trade-in)
- current market value of the machine
- current price of a replacement machine
- number of years machine will be used on the farm

In a simulation model, a machine is replaced only if it has passed its economic life on the farm. The value of a machine at the time it is replaced (traded-off) after N years is the current value at $t=0$ discounted to the year the machine is replaced. The cost of a new machine is the $t=0$ cost of a replacement inflated for N years. The net cost of the replacement is the updated cost less the trade-in value.

Financing machinery replacements can be handled on a machine-by-machine basis with a loan for each purchase. Alternatively the model can sum the new loan requirements and finance one new machinery loan for each year machinery is replaced. I prefer the later myself to reduce the number of loans to simulate. Additionally, with one loan per year your model can easily repay the principal early if surplus cash is generated.

To finance machinery replacements the model must have a value for the minimum down payment on machinery loans. First compare the trade-in value to the down payment requirement; if the trade-in is less than the minimum down payment, the deficit must come from cash reserves. Finance the remainder (cost less minimum down payment) at the prevailing terms for machinery loans (number of years and interest rate). If the trade-in value exceeds the minimum down payment, then finance the difference and do not pay a cash difference. Annual interest payments appear in the Income Statement and principal payments appear in the Cash Flow. The remaining debt and updated value of machinery appear in the Balance Sheet.

Refer to Machinery Demo.XLS for an example of how to simulate machinery replacement for a business. This demo program is set to finance the purchase of replacement machinery for 5 years using fixed interest rate mortgages. Ten different machinery items are included in the demo. Change the information for any or all of the machines to observe the impacts on cash outflow requirements, interest costs and the firm's balance sheet. Adding more machines to the spreadsheet can be done by inserting columns before column M, entering the new machines and copying the equations for the new machines. All of the equations should work, but it is your responsibility to check them.

Using this method for replacing machinery will produce an uneven cash flow requirement for machinery replacement. The uneven cash flow requirements need to be evaluated after solving for the replacement of each machine to see if “too much” machinery is to be replaced in any given year. Spread out these cash flow requirements by changing the economic life of individual machines until the cash flow requirements are reasonable for the business being analyzed.

Farm Level Simulation Model

A 10 year, whole farm simulation model is demonstrated in Farm Simulator Demo.XLS. The model is developed using the type of calculations in FLIPSIM (Richardson and Nixon). The model is capable of simulating one, two, or three crops. Input data required of the user are in bold in the Model and the Stochastic worksheets. Historical prices and yields for the three crops on the farm are entered in the Stochastic worksheet. The historical prices and yields are used to develop and simulate a MVE distribution for prices and yields. Charts showing the historical and stochastic prices and yields are provided at the bottom of the Stochastic worksheet.

Projected inflation rates for variable inputs are also entered in the Stochastic worksheet. Separate inflation rates for fuel, labor, and other inputs are entered so the model can simulate the effects of alternative inflation rates. Annual interest rates for financing operating costs are also included as input for the whole farm simulation model.

All other inputs necessary to simulate a crop farm are entered at the top (rows 8-71) of the Model worksheet. Assets and liabilities, family living, fixed costs, and other income are required for a farm risk analysis. Budgets for each crop are included along with base acres, payment yields, and price wedges. The price wedges are used to localize stochastic national prices for each crop. Annual values must be provided by crop for: planted acres, mean crop yields, mean crop prices, loan rates, direct payment rates, and target prices. Planted acres and mean yields can come from the producer or the farm plan being analyzed. Average annual crop prices can be obtained from FAPRI or developed from other sources. Policy values for 2002-2007 are specified in the 2002 farm bill. Policy values for 2008-2011 are assumed to remain at their 2007 levels.

The whole farm simulation model is made up of several sections where the calculations are done. A detailed income statement is provided in rows 75-104. The values reported in the statement are calculated in specific sections for each crop. A cash flow statement is provided in rows 106-123 for the express purpose of tracking ending cash reserves. The farm's balance sheet is in rows 125-136 to calculate net worth.

Values found in the cash flow and balance sheet are used to calculate net present value, one of the KOVs for the farm (rows 140-147). Three more KOVs useful for evaluating a farm with risky cash flows are: probability of a cash flow deficit, probability of negative cash flows, and probability of losing real net worth. Definitions for these variables are:

- Cash flow deficits occur when cash outlays exceed net cash farm income.
- Negative cash flows occur when beginning cash plus net cash income and other cash earnings are less than total cash outflows. This is the amount of the operating loan, which must be refinanced or carried over.
- Firms lose real net worth when the present value of ending net worth is less than beginning net worth.

All three of the probability KOVs are simulated using counter variables in rows 149-151. When the condition is true the counter is 1, else it is zero. The probability of a cash flow deficit can be simulated for each year using 10 counter variables. This type of KOV facilitates pinpointing years that the firm will likely need to refinance its operating loan.

The section of the model where crop receipts, expenses, and government payments to crop are calculated, is in rows 154-189. Stochastic production equals planted acres times stochastic yields. National prices are adjusted by the local wedge (or basis) to derive stochastic local prices. Market receipts equals the product of production and local prices. LDP, direct and CCP government payments are calculated using annual loan rates, target prices and direct payment rates. Base acres and DP and CCP payment yields are used for these calculations. An AWP fraction is used to convert the stochastic national price to a stochastic adjusted world price used to calculate the LDP rate.

Costs of production for each crop are calculated using annual inflation rates and the crop budgets. For crop 1 the costs of production are calculated in rows 272-285. The same steps are repeated for crops 2 and 3. Results from this section are used in the Income Statement. Other costs for the farm are calculated next. Of particular note are interest costs for the operating loan and the operating loan carryover, i.e., for financing cash flow deficits.

Interest on cash reserves are calculated so they can enter the cash flow statement as earned interest. Family living costs or withdrawals are calculated so these cash outflows can appear in the cash flow statement. Land value is calculated using the user's assumed annual inflation rates for land. Land value shows up in the asset side of the Balance Sheet.

The calculations necessary to amortize the original land loan and a beginning farmer loan are included in the model in rows 369-442. Annual interest costs in these schedules are used in the Income Statement to calculate total costs. Annual principal payments for the two loans appear in the Cash Flow statement as these are not production expenses.

The final section of the model includes the calculations to compute annual federal income taxes for the farm. A corporate income tax schedule is used to compute federal income taxes.

At the top of the model worksheet is a KOV table, rows 10-31. The table is cell referenced to each of the KOVs found in the model. When the farm model is simulated, the user can highlight the same KOVs every time and they are always in the same order. Given that the KOVs are always in the same order, you can develop summary tables for the outputs and permanent charts to interpret the results. An example of such a table is provided in columns AB – AK in SimData1.

Simulating an Econometric Model

Recursive supply demand models for crops and livestock can be estimated and simulated using tools provided in Simetar. A recursive supply demand model for the US soybean sector is estimated and simulated in this section. The Excel demo program is named Soybean Model Demo.XLS.

The objective of the model is to simulate annual prices, ending stocks, and the present value of total government payments, 2002-2004, for the 1996 and the 2002 farm programs. The objective requires that the model incorporate different types of farm programs and endogenously solve for price.

The hypothesized equations for the model are:

$$\text{Planted Acres}_t = f(\text{E Price}_t, \text{Planted Acres}_{t-1})$$

$$\text{Harvested Acres}_t = f(\text{E Price}_t, \text{Planted Acres}_t)$$

$$\text{Yield}_t = f(\text{E Price}_t, \text{Yield}_{t-1}, \text{Trend}_t)$$

$$\text{Domestic Use}_t = f(\text{Price}_t, \text{Domestic Use}_{t-1}, \text{Income/Pop}_t)$$

$$\text{Exports}_t = f(\text{Price}_t, \text{Exports}_{t-1}, \text{GDP EU}_t)$$

$$\text{Ending Stocks}_t = f(\text{Price}_t, \text{Stocks}_{t-1})$$

$$\text{Define Eprice}_t = \text{Max} [\text{Price}_{t-1} \text{ or Loan Rate}_t]$$

Based on these equations historical data for the endogenous and exogenous variables, 1980-2001, were obtained and added to the Data worksheet in the demo program.

A significant time saving step is to create a worksheet for each equation to be estimated. The variables hypothesized for each equation are cell referenced from the Data worksheet to be equation's worksheet. For example the PltAcres worksheet has the values for Planted Acres_t, Eprice_t, and Planted Acres_{t-1}. The multiple regression for each equation is estimated in the equation's own worksheet so it is easy to locate.

The residuals and the standard deviation of residuals from each of the regression models are cell referenced to the Stoch worksheet. The residuals are used to define and simulate a MVN distribution of shocks to the econometric equations. The stochastic shocks for each year in the planning horizon are tested to insure they are appropriately correlated as part of the validation.

The intercept and slope parameters for the six econometric equations are cell referenced from their worksheets to the Model worksheet where the completed model is simulated. The coefficients for each of the equations are assembled in rows 24-40 of the Model worksheet. Solving for price in a recursive model is accomplished by first estimating the total demand equation, re-writing the equation so that price is a function of quantity, and then solving for price using quantity supplied.

Total demand is the sum of domestic use, exports and ending stocks. A separate total demand equation must be estimated for each year 2002-2004 as the exogenous variables in these equations change each year. The total demand is estimated as:

$$DD_t = a + bP_t + c DP_{t-1} + d I_t + \tilde{e}_1$$

$$ED_t = k + fP_t + g ED_{t-1} + h G_t + \tilde{e}_2$$

$$ES_t = i + jP_t + \tilde{e}_3$$

$$TD_t = a + k + i + (b+f+j) P_t + cDP_{t-1} + d I_t + gED_{t-1} + h G_t + (\tilde{e}_1 + \tilde{e}_2 + \tilde{e}_3)$$

$$\text{let } A = a + k + i$$

$$B = b + f + j$$

$$C = cDP_{t-1} + d I_t + g ED_{t-1} + h G_t$$

$$\tilde{e} = \tilde{e}_1 + \tilde{e}_2 + \tilde{e}_3$$

Then total demand can be written as:

$$TD = A + C + B P_t + \tilde{e}$$

or

$$P_t = \frac{A}{B} + \frac{C}{B} + \frac{1}{B} TD + \frac{\tilde{e}}{B}$$

which simplifies to

$$\tilde{P}_t = \alpha + \beta Q + \beta \tilde{\epsilon}$$

$$\text{let } \alpha = \frac{A}{B} + \frac{C}{B}$$

$$\beta = \frac{1}{B}$$

The total demand is thus specified as price is a function of quantity and an error term which makes the intercept stochastic (Figure 13.1).

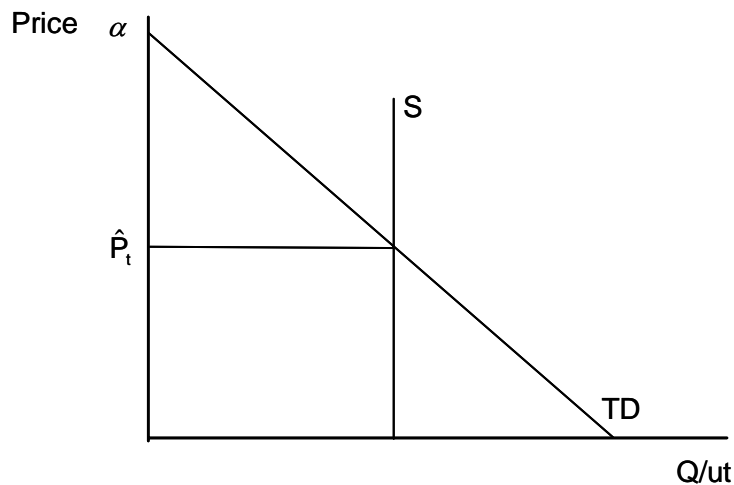


Figure 13.1. Price Dependent Demand Function and Supply Determine Price.

Given a quantity supplied (Q) in the equation, equilibrium price is found directly by solving the demand function. This is permitted because of the assumption that supply equals total demand (Figure 13.1).

In Soybean Model Demo.XLS the process of deriving total demand for each year is summarized in rows 42-52. The intercepts and slopes on price are summed in rows 44-47. The shift effects for domestic demand ($d I_t + c DP_{t-1}$) are calculated in row 44 and added to the domestic demand intercept in row 50. Similar steps are done for exports. The total demand intercept changes each year (cells E54:H54) because the exogenous variables (I , G , DP_{t-1} and ED_{t-1}) and the stochastic shocks change annually.

Because quantity supplied is a function of expected price it is determined in row 73 and then used in the price equation (row 56) to display price in the current year. This specification is consistent with the fact that the season average price is largely determined after harvest and supply is known with certainty (Figure 13.1). Supply is calculated as a sum of fixed imports (3.0 each year) and production. Stochastic shocks on yields, planted acres and harvested acre equations cause supply to be a stochastic variable in addition to responding to lagged prices and

current policy variables. It is assumed that for the simulation period the expected price is defined as:

- 1996 farm bill
E Price_t = Max [Price_{t-1}, Loan Rate_t)

- 2002 farm bill
E Price_t = Max (Price_{t-1}, Target Price_t)

Once price is calculated the value is used to simulate domestic demand and exports. Domestic use and exports (rows 76 and 77) are calculated using the stochastic price and the appropriate exogenous variables and stochastic shocks used in rows 44 and 45 to estimate total demand. Ending stocks are solved as an identity:

$$\text{Ending Stocks}_t = \text{Supply}_t - \text{Domestic Use}_t - \text{Exports}_t$$

One of the KOVs is the present value of total receipts plus government payments. Direct and counter cyclical payments are calculated under the 2002 farm bill using base acres, direct payment yields and counter cyclical payment yields. A CDF chart of the PV of total revenue under each policy is provided in the SimData worksheet. The simulated annual soybean prices are summarized using a fan graph in the SimData worksheet.

References

- Adams, Gary M. "Impact Multipliers of the U.S. Crops Sector: A Focus on the Effects of Commodity Interaction." Vol. I, II, University of Missouri-Columbia, Department of Agricultural Economics, Ph.D. Dissertation, 1994.
- Bell, D.E. and A. Schleifer, Jr. Decision Making Under Uncertainty. Cambridge, MA: Course Technology, Inc., 1995.
- Bernstein, P.L. Against the Gods: The Remarkable Story of Risk. New York: John Wiley & Sons, Inc. 1996.
- Brown, D. Scott. "A Structural Approach to Modeling the U.S. Livestock and Dairy Industries: With Emphasis on the Dynamics Within the System." University of Missouri-Columbia, Department of Agricultural Economics, Ph.D. Dissertation, 1994.
- Catlett, L.B. and J.D. Libbin. Inventory in Futures and Options Market. New York: Delmar Publishers, 1999.
- Fleisher, B. Agricultural Risk Management. Boulder, CO: Lynne Rienner Publishers, Inc., 1990.
- Gempesaw, C.M., II, and J.W. Richardson. "The Use of Simulation in Management Decision Making: The Case of Poultry and Aquaculture Production." In Simulation Systems, G.W. Zobrist and J.V. Leonard, editors. Amsterdam: Gordon and Breach Science Publishers, 2000.
- Gill, R.C., II. "A Stochastic Feasibility Study of Texas Ethanol Production: Analysis of Texas State Legislature Ethanol Subsidy Proposal." Texas A&M University, Department of Agricultural Economics, M.S. Thesis, December 2002.
- Gray, A.W. "Agribusiness Strategic Planning Under Risk." Texas A&M University, Department of Agricultural Economics, Ph.D. Dissertation, August 1998.
- Naylor, T.H. "Corporate Simulation Models." In Business Simulation for Decision Making. W.C. House, editor, New York: PBI, 1977.
- Ragsdale, C.T. Spreadsheet Modeling and Decision Analysis. Cincinnati, OH: South-Western College Publishing, 1998.
- Ray, D.E. and J.W. Richardson. "Detailed Description of POLYSIM." Agricultural Experiment Station, Oklahoma State University and USDA, Technical Bulletin, T-151, December 1978.
- Richardson, J.W. and H.P. Mapp, Jr. "Use of Probabilistic Cash Flows in Analyzing Investment Under Conditions of Risk and Uncertainty." Southern Journal Agric. Econ. Dec. 1976.

Richardson, J.W. and C.J. Nixon. "A Description of FLIPSIM V: A General Firm Level Policy Simulation Model." Bulletin 1528, Texas Agricultural Experiment Station, July 1986.

Robison, L.J. and P. Barry. Present Value Models and Investment Analysis. New Port, AL: The Academic Press, 1996.

Winston, W.L. Simulating Modeling Using @Risk©. New York: Duxbury Press, 1996.

Chapter 14

Simulation Applications for Business Management

Simulation is a useful tool for risk management in business. Several different examples of using simulation for business risk management are presented in this chapter. The first section describes how to develop a financial risk management model. The second section demonstrates how simulation can be used to conduct portfolio analyses. The third section deals with project management under risk while the fourth section demonstrates how to use simulation for developing a bid for a risky project. Simulation can be used to conduct feasibility studies for risky investments or projects as demonstrated in the fifth section. The last section of the chapter presents an example of using simulation for inventory management.

Financial Risk Management

Financial risk management can take on different meanings, depending on whether you are the borrower or the lender. Borrowers need to manage risk to insure their ability to repay loans, remain solvent and ultimately increase real net worth. Lenders are interested in making loans to borrowers who have high probabilities of repaying their loans on time. A financial risk management model for the borrower will generally work for the lender, at least to determine whether the borrower is a good risk for a particular loan.

The length of loan under consideration will dictate the type of financial risk management model developed. A one-year operating loan requires a much simpler model than a 5 to 10 year loan for machinery or buildings. To demonstrate a financial risk management model a one year operating loan is analyzed. Risks affecting the repayment capacity of the borrower are: production, production costs, and price of the output. Risk management tools the borrower could make use of are production insurance, options or futures markets, and government payments.

The financial risk management demo program, Financial Risk Management Demo.XLS, is for a one year analysis of a corn farm. The variables the user may change are bolded and confined to rows 7-63. The user can change crops by inserting the relevant values for risk, costs, loans, crop insurance, and hedging strategies. The demo program is capable of simulating and comparing 9 alternative risk management strategies at once (see row 69 for the nine scenario names). The risk management strategies are combinations of hedging and crop insurance that a producer could consider and a lender may require. The base scenario is a no insurance, no hedge strategy.

The information for 3 levels of MPCCI and 3 levels of CRC insurance is provided as input in rows 41-44. The user specifies which level of MPCCI coverage to simulate using a 1, 2, 3, switch in row 20. A similar switch for CRC insurance is provided on row 21. Table lookup functions in the model use the switches to get the appropriate insurance values for the simulation.

The marketing risk management strategies involve hedging a fraction of the crop using the futures market or buying put options. Information to specify the marketing strategies is provided on rows 25-29. Both hedging and options are included because lenders and borrowers should consider the certain costs of the option premium versus the uncertain risk of hedging losses in the futures market.

Provisions in the 2002 farm bill contain safety net payments. The farm's parameters for the farm program (base acres and payment yields) are entered on rows 32-37. National farm program provisions must be entered as well so government payment rates for CCP and LDP programs can be simulated (rows 32-35).

The loans to be evaluated are specified in rows 10-22. Information for the farm's operating (self-liquidating) loan is provided first, followed by the parameters for a proposed loan and then parameters for three existing loans. The information for an operating loan consists of the interest rate and a value for the fraction of the year that the loan is outstanding. Operating loans are one year loans that are repaid after the crops are sold. Note that the repayment of the principal is not shown in the financial tables, because the loan is to purchase inputs for production. Including the costs of inputs in the income statement automatically accounts for the "repayment" of the operating loan to purchase the inputs. As a result an operating loan can be simulated by: (1) assuming the operating loan is for the total amount of fixed costs, variable production costs, harvesting costs, and costs for options; and (2) assuming the full amount of the loan is paid interest for only a fraction of the year. For example a \$100,000 operating loan, issued as a line of credit, is not fully borrowed until the end of the season; therefore interest may only be paid on 66 percent of the total loan amount. Interest on operating loans is calculated in rows 140-150 for the 9 risk management strategies.

Calculations for MPCPI production insurance and cash receipts insurance (CRC) are in rows 111-137. The premium costs are calculated at the first of the year. Indemnities paid are calculated using stochastic yield and prices assuming the crop is harvested and year end prices are known. Market receipts and government payments are calculated assuming all payments are received once the crop is sold (rows 153-175). Government payments are actually paid at different times during the year. This simplifying assumption is not significant because the bank holds a lien on the payments no matter when they are paid.

The gains and losses from hedging are calculated in rows 178-193. If the high price for the futures contract exceeded the hedge price objective, a hedge is placed. To simulate whether a hedge is placed or not a random variable for the December corn futures high prices is used (rows 47-61). When the simulated high price exceeds the hedge objective then a hedge is placed and gains or losses are calculated. A gain is observed if the stochastic October futures price is less than the hedge price objective. A loss occurs if the October futures price exceeds the hedge price objective. More complicated hedging strategies could be programmed to test different risk management schemes that utilize the futures market.

The gains from buying puts at a specified strike price are calculated given the stochastic futures price at harvest (rows 182-192). The premium is paid at the start of the year based on the option premium specified in the input.

Each loan is amortized separately and the principal and interest costs are reported for the year being simulated. Federal income taxes for the farm are calculated assuming that the farm is taxed as a corporation. This assumption avoids calculating self-employment taxes and medicare taxes.

The results of the costs, receipts, and tax calculations are summarized in pro forma financial statements in rows 67-103. The key output variables for the business are: net cash income, ending cash balance, annual probability of cash flow deficits, and annual probability of negative ending cash reserves (refinancing). The results for the two types of probability KOVs can be summarized in a bar chart to show the risk exposure under alternative risk management strategies (Figure 14.1).

Financial risk management analyses for a business are easily programmed as a simulation model. All businesses face stochastic production and output prices. If input costs are stochastic they can be included in the model. The range of risk management options will differ by business but the examples provided in this section should suggest a variety of strategies which could be analyzed. A range of risk management strategies should be analyzed as the financial risk of a borrower will differ significantly under alternative strategies.

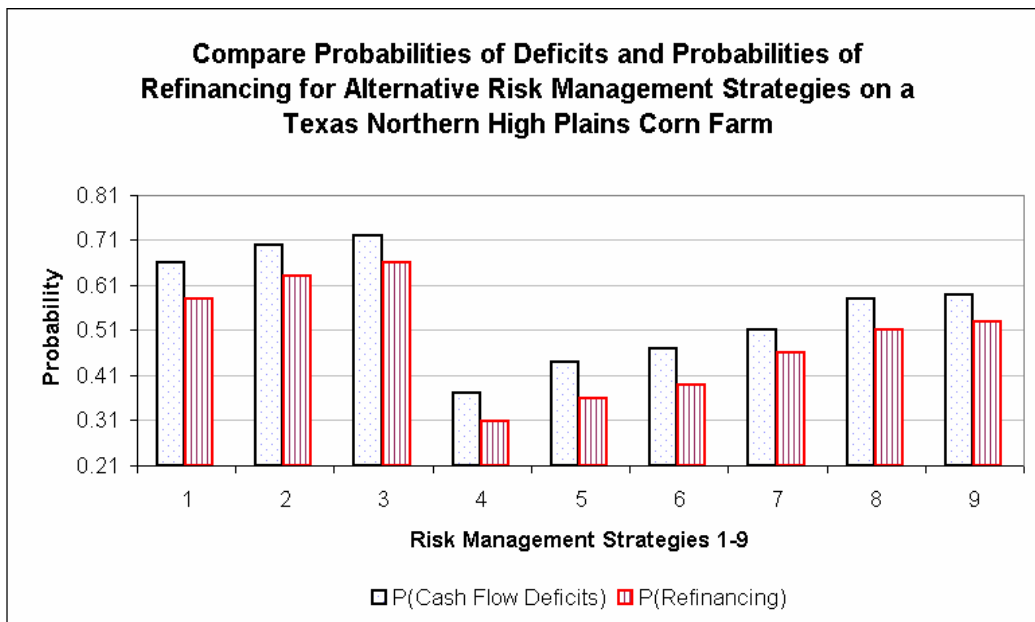


Figure 14.1. Example of Using a Bar Chart to Present Probabilities of Deficits and Refinancing for Alternative Risk Management Strategies.

Portfolio Analysis

A portfolio is a combination of different investment options that are combined based on fractions of separate investment instruments. Given two risky investment instruments, X and Y, one can define an infinite number of portfolios by selecting, α , the fraction of X in the portfolio or:

$$P_i = \alpha_i X + (1 - \alpha_i) Y$$

at the extremes, α equals zero or one and the portfolio has all X or all Y, respectively. Any value of α between zero and one defines a possible portfolio of X and Y.

Given two risky investment instruments the mean return equals

$$\text{Mean} = \alpha r_x + (1 - \alpha) r_y$$

where r_x and r_y are the average returns to X and Y. The variance on returns for a portfolio equals

$$\text{Var} = \alpha^2 S_x^2 + (1 - \alpha)^2 S_y^2 + 2 \alpha(1 - \alpha) \rho S_x S_y$$

where S_x^2 and S_y^2 are the variances of returns for X and Y, respectively, and ρ is the correlation between returns on X and Y.

The portfolio that minimizes the variance on returns is defined by Bell and Schleifer as

$$\alpha = \frac{S_y (S_y - \rho S_x)}{S_x^2 + S_y^2 - 2 \rho S_x S_y}$$

Diversification of an existing portfolio will reduce risk if and only if

$$\sigma_{xy} > \sigma_{xy,z} * \rho_{xy,z}$$

or

$$\rho_{xy,z} < \frac{\sigma_{xy}}{\sigma_{xy,z}}$$

where σ_{xy} is the standard deviation of returns for the existing portfolio between X and Y, defined by α , $\sigma_{xy,z}$ is the standard deviation of returns for the proposed portfolio that includes Z, and $\rho_{xy,z}$ is the correlation of returns between the existing XY portfolio and the proposed X, Y, Z portfolio. A new investment, Z, is always of value for diversification if the returns to Z are independent or negatively correlated to returns for the existing portfolio (Bell and Schleifer).

The process of calculating pair wise standard deviations and correlations for alternative portfolios is time consuming and can be confusing. Simulation and risk analysis offers a direct approach for evaluating alternative portfolios with many different investment instruments. In a simulation context, portfolio analysis is simply the evaluation of alternative investments that could have occurred.

To simulate how the different possible investments might have performed we simulate the returns for proposed portfolios as stochastic variables. Assume there are 10 separate investment instruments available with adequate history of returns, r_i , to define a MVE distribution. (If the investment is long term use annual rates of return and if the investment period is short term use monthly returns to define the MVE distribution.) Simulate stochastic rates of return for each investment (\tilde{r}_i) using the MVE and evaluate the performance of possible portfolios as:

$$\begin{aligned}\tilde{P}_1 &= \alpha_1 \tilde{r}_1 + \alpha_2 \tilde{r}_2 + \dots + \alpha_{10} \tilde{r}_{10} \\ \tilde{P}_2 &= \alpha_1 \tilde{r}_1 + \alpha_2 \tilde{r}_2 + \dots + \alpha_{10} \tilde{r}_{10} \\ &\cdot \\ &\cdot \\ &\cdot \\ \tilde{P}_N &= \alpha_1 \tilde{r}_1 + \alpha_2 \tilde{r}_2 + \dots + \alpha_{10} \tilde{r}_{10}\end{aligned}$$

The values for the α_i weights change from one portfolio to another. For all portfolios the following restriction must hold:

$$1.0 = \sum_{i=1}^{10} \alpha_i$$

The KOVs for the simulation model are the \tilde{P}_i returns for the N alternative portfolios being evaluated. Stochastic dominance and dynamic certainty equivalents can be used to rank the alternative portfolios.

To demonstrate how simulation can be used for portfolio analysis the historical returns for ten mutual funds are analyzed in Portfolio Analysis Demo.XLS. The steps followed in the analysis are:

- Check the data for a trend in the annual rate of return.
- Calculate the parameters to simulate the individual mutual funds as a multivariate empirical distribution.
- Simulate each mutual fund as a random variable (lines 30-37).
- Identify alternative portfolios that are made up of different portions (fractions) of each fund (lines 119-130).
- Simulate each of the nine mutual funds as if they are mutually exclusive investments (lines 101-110).
- Multiply the simulated rates of return by alternative portfolio proportions and sum across the mutual funds to determine the annual rates of return for each portfolio (lines 133-144). These six annual rate of return variables are the key output variables (KOVs) for the simulation analysis.
- The final step is to analyze the simulated outcomes for the six portfolios using StopLight, Stochastic Dominance, Certainty Equivalents, and CDF graphs.

The results of the portfolio analysis are presented in the different worksheets in Portfolio Analysis Demo.XLS. The simulation results are in SimData. The StopLight and CDF results are in their own worksheets. The summary statistics for two separate simulations of the workbook are summarized in lines 148-163 to demonstrate that all of the portfolios are associated with lower relative risk than any of the individual mutual funds. Portfolio 10 had the lowest CV (63.97 percent) while the CVs of the individual funds ranged from 90 to 215 percent. The Certainty Equivalent analysis indicates that decision makers who are risk neutral to moderately risk averse prefer portfolio 11, but those who are more risk averse prefer portfolio 10.

Project Management

Project management refers to the development and analysis of a plan for completing a project which has multiple parts or activities. For example, a manager may be asked to develop a detailed plan for expanding the business by adding a new division or analyzing the development of a new product. The detailed plan would list all of the activities to be done, an estimate of the time to complete the project, and a cost estimate. The project management plan may or may not include risk variables.

Project management is a natural application for stochastic simulation because any plan or feasibility analysis should account for the affects of risk on its success. Imagine how much more complete a report for a proposed business expansion would be, if it included probability distributions for days to completion and costs, as well as probabilities for achieving the decision makers key output values.

The fundamentals for developing a project management plan under uncertainty are presented in this section. The example project selected to demonstrate the techniques is the development of a large scale economic simulation model. The steps involved in carrying out a project management analysis are:

- Identify the activities and their order,
- Develop a network diagram of the project,
- Determine the time step (days, weeks, or months) for each activity,
- Calculate the start and end times for each activity,
- Determine costs for each activity on a dollar per day, week, or month basis,
- Assign risk to the length of time for each activity, and
- Simulate and validate model before writing the report.

– Identify Activities and Their Order

List all of the activities necessary to successfully complete the project. Start at the beginning and keep them in the order they will be started. An example of a list of project activities is provided in Column B of the Project Management Demo.XLS workbook (Figure 14.2). In the demo program you will find all of the steps necessary to develop an economic simulation model. Each activity is assigned a number in Column A for reference (Figure 14.2).

The order that the activities must be started is indicated by specifying the immediate predecessor activity in Column C of Figure 14.2. Activity 1 precedes activities 2 and 3 and activity 4 cannot start until both activities 2 and 3 are completed. Following down Column C in the demo program we find bottlenecks will likely occur at activities 4, 14, and 19. Bottlenecks occur when an activity cannot proceed until all of its predecessors are completed.

	A	B	C	D	E	F	G	H	I	J	K	L	M
7									PDF for Time Expended				
8			Immediate	Time to	Start	End	Cost per	Total	Pessimistic	Expected	Optimistic	Formula for	
9	No	Project Activities	Predecessor	Complete	Time	Time	Time Unit	Cost	Minimum	Expected	Maximum	Start Time	End Time
10	1	Develop Diagram of the System	0	24.7	-	24.7	100	2,473	10	20	30	-	=D10
11	2	Define the KOVs	1	1.5	24.7	26.3	100	152	1	2	4	=F10	=D11+E11
12	3	Determine the Intermediate Outputs	1	4.0	26.3	30.2	100	398	1	2	5	=F11	=D12+E12
13	4	Design Complete Output Format	2 & 3	5.0	30.2	35.2	100	498	3	5	10	=MAX(F11,F12)	=D13+E13
14	5	Validate Format on User	4	9.2	35.2	44.5	200	1,849	5	10	20	=F13	=D14+E14
15													
16	6	Define Input and Calculated Variables	4	2.4	35.2	37.6	100	239	2	3	6	=F13	=D16+E16
17	7	Gather Primary Data	6	286.9	37.6	324.5	50	14,346	100	200	300	=F16	=D17+E17
18	8	Gather Secondary Data	6	18.1	37.6	55.7	50	903	10	15	30	=F16	=D18+E18
19	9	Estimate Econometric Relations	8	8.9	55.7	64.6	150	1,342	10	30	60	=F18	=D19+E19
20													
21	10	Identify Stochastic Variables	5	4.2	44.5	48.7	150	831	1	2	4	=F14	=D21+E21
22	11	Gather Data for Stochastic Variables	10	44.7	48.7	93.4	50	2,235	20	30	60	=F21	=D22+E22
23	12	Parameter Estimation for Stochastic Variables	11	6.0	93.4	99.3	150	894	5	10	20	=F22	=D23+E23
24	13	Distribution Testing	12	2.3	99.3	101.6	200	453	2	3	6	=F23	=D24+E24
25													
26	14	Assemble Model & Begin Testing	7, 9 & 13	105.7	324.5	430.3	250	26,432	50	100	200	=MAX(F17,F19,F24)	=D26+E26
27	15	Validation and Verification by	14	0.0	430.3	430.3			0	0	0	=F26	=D27+E27
28	16	Personal	15	43.6	430.3	473.9	200	8,721	30	45	90	=F27	=D28+E28
29	17	Peers	15	142.0	430.3	572.2	100	14,198	60	100	200	=F27	=D29+E29
30	18	Producers	15	49.7	430.3	480.0	150	7,456	40	60	120	=F27	=D30+E30
31	19	Model Refinement Based on Validation	16, 17 & 18	36.8	572.2	609.0	250	9,196	10	25	50	=MAX(F28,F29,F30)	=D31+E31
32	20	Model Application & Report Preparation	19	10.5	609.0	619.5	200	2,100	5	10	20	=F31	=D32+E32
33													
34		Key Output Variables for Project Management			Total	619.5	Total	94,617					

Figure 14.2. Project Management with Simulation.

– **Network Diagram**

A simple network diagram of the activities can be developed using the Drawing toolbar in Excel. (The network diagram in the demo program was drawn using the Auto Shapes > Basic Shapes and Auto Shapes > Connectors on the Drawing toolbar. For a tutorial on developing network diagrams with Excel refer to Appendix A.) Simple abbreviations to describe the activities can be typed into the flow chart boxes. Arrows from the Connectors menu simply attach the boxes in the order specified in Column C. Double clicking on a drawn object allows one to move it any place on the worksheet, so the relational order of activities can be easily changed. The network diagram should visually show how the activities are related, the predecessors and the bottlenecks.

– **Time for Each Activity**

The unit of time (days, weeks, or months) will depend on the project. The demo project is specified in days. In the initial phase, the project management table should be developed ignoring risk on time required to complete each project, as shown in Column D of Figure 14.2. After the equations for the next step (Columns E and F) are developed and validated, the time to complete each project will be made stochastic. (Set the Simetar Simulation Engine option to Expected Value and then click the Save button to see the demo program as it looked in the initial phase with deterministic days to completion.)

– **Calculate Start and End Times**

The formulas used to calculate the start and end times for each activity are provided in Columns L and M of Figure 14.2. The end time for activity 1 is the number of days to complete the activity (or D1). The start time for activities 2 and 3 is the end time for activity 1 because activity 1 is the immediate predecessor activity for 2 and 3. End time for activity 2 is its start

time plus the time required to complete activity 2. Repeat these formulas for calculating the start and end times until you get to a bottleneck activity. The basic formulas are:

$$\begin{aligned}\text{Start time activity } i &= \text{End time for predecessor to } i \\ \text{End time activity } i &= \text{Start time } i + \text{Time to Complete } i\end{aligned}$$

The start time for a bottleneck activity that is dependent on two or more predecessors is calculated using the Excel maximum command. For example, the start time for activity 14 is dependent on completing activities 7, 9, and 13 so the equation is written as the maximum of (End time 7, End time 9, or End time 13) or in Excel this is:

$$=\text{MAX}(\text{F17}, \text{F19}, \text{F24})$$

in cell E26 (Figure 14.2). The end time for the last activity is the key output variable (KOV) for the length of the project.

– **Determine Costs**

The cost per unit of time should be specified separately for each activity. This allows experimentation with alternative cost scenarios. For example, you may want to contract out certain activities to free up internal people so the project can be done faster.

The per day costs in the demo program for each activity are in Column G (Figure 14.2). Total cost of each activity is the product of time to complete in Column D and cost per time unit in Column G. The sum of total costs over all activities is the second key output variable (cell H34 in the demo program).

– **Assign Risk**

Simple GRK probability distributions are used for the demo program to make the time to complete each activity stochastic. The time to complete cells in Column D are stochastic and use the minimum, expected, and maximum days to complete in Columns I, J, and K, respectively. Project management literature refers to these GRK parameters as pessimistic, expected, and optimistic time to completion.

The GRK distribution is used in the demo because it readily conforms to the optimistic-pessimistic paradigm. The GRK distribution could be replaced with any distribution you believe best fits the times to complete variables.

– **Simulation and Validation**

Verification of the formulas used to calculate the start and end times must be done while Simetar is in Expected Value mode. (See Chapter 16 for using this setting in the Simetar Simulation Engine.) These formulas must be checked carefully because millions of dollars may rest on their accuracy. Once the formulas are verified and found to be correct, it is time to simulate the model stochastic.

Validation of the stochastic model should start with checking the simulated time to complete variables. This is necessary to insure that the program is accurately simulating the specified distributions. Check for negative times to complete and values that are un-realistically large. Remember if you use the GRK distribution, the simulated minimums and maximums will extend beyond the specified minimum and maximum in Columns I and K about 3 percent of the time.

Once the model is validated, simulate it and collect the statistics for time to complete and total cost. CDFs and StopLight summaries of these key output variables will be useful for summarizing the project management analysis. See the Project Management Demo.XLS workbook for an example of the outputs generated with a simulation model for project management.

Bid Analysis

Budgets are a standard tool for decision makers and are usually the foundation for preparing bids on proposed projects. In the area of project analysis, pro forma budgets are prepared for evaluating new projects and prior to bidding on contracts. Generally, such budgets are prepared with less than perfect knowledge as to the actual expenses that will be incurred for each cost category. In the best of cases, experts within each area (or activity) of a project are consulted as to the expected cost for an activity. Risks associated with project costs are generally not included in the budget, thus contributing to cost over runs and project managers looking for new jobs.

Successful businesses are not awarded every bid they submit, but they make a profit on every award received. The key then is to submit bids that provide an acceptable probability of making a profit, given the risk and hope that it is the bid that is accepted. Adding risk to a project budget analysis and bid development is a simple task that can be done with simulation. An example of a project budget analysis for developing a bid under risky conditions is presented in this section. The steps to conducting this type of analysis are:

- Identify the cost categories for the project,
- Develop probability distributions for the cost categories, and
- Validate and simulate the model.

– Identify Cost Categories

Identify all of the cost categories (activities) for the proposed project. The order of the cost categories is not crucial but putting them in chronological order may insure that none of the activities are left out. In the second step, develop probability distributions for each cost category so consider this when developing the list of activities.

The more focused the cost categories list the easier it is to develop the pdfs for cost. For example, there is considerably more uncertainty associated with defining a pdf for the cost of a new building than the uncertainty surrounding the pdf for the cost of carpet for the building.

The Bid Analysis Demo.XLS workbook demonstrates how to construct and use a bid analysis simulation model. The example is based on the cost categories a general contractor might go through to develop the bid for a house. Although a general contractor may have built many homes, each bid must be developed based on the specifications for the proposed building

and current materials prices. The list of cost categories in the demo worksheet reflect the activities the contractor wants to consider for the bid.

– **Develop PDFs for the Cost Categories**

Each of the cost categories can be a random variable. Initially assume that the random variables are independent. The parameters to define the PDFs can come from several sources, such as:

- Historical cost data for similar projects,
- Experts in the field, and
- Personal experience of the project manager, i.e., subjective distribution.

Historical information for similar projects will likely be the source of data for most project analysis simulation models. The project manager may not have supervised exactly the same type of project, but he/she could have a history of the costs for particular activities. For example, site preparation may call for leveling the ground, tree removal and soil testing. Based on past building projects in the area, this cost has ranged from, say, \$500-\$1,000 and appears to be uniformly distributed.

Another example of using historical data would be the cost of permits and fees. These types of costs are published or available from the county courthouse and are deterministic. The more cost categories identified in Step 1 the more PDFs that must be parameterized, but it is easier to find historical data for narrowly defined cost categories.

Experts in the area are an excellent source of information for parameterizing the cost category PDFs. The more narrowly defined the cost categories, the easier it is to find an expert and to solicit information from the expert. For a construction project we would call on an excavation company to quote a bid to supply sand and gravel to the site. Early on in the process they may give a range of, say, \$3,000 to \$5,000 with an average of \$4,000. This information fully defines a GRK distribution. Similar, information can be obtained from other experts or subcontractors on the project. You do not have the contract yet, so you do not have to pin down the subcontractors to a single value; just get a range for the cost of each part of the project.

The final source of information for parameterizing PDFs about cost activities is the project manager's personal experience. This may largely be based on the manager's subjective expectations for the cost categories, given the specifications for the proposed project. Some of the subjective distributions may have less risk (see framers costs in demo) while others may have considerably more risk (see trim carpenters costs in demo). Subjective distributions are based on past experiences of the project manager and expectations for the particular project under analysis.

– **Validate and Simulate the Model**

Once the parameters are developed for each of the cost categories they can be assembled into a bid analysis simulation model. See Bid Analysis Demo.XLS for an example of a simulation model for conducting a project analysis. The cost categories are listed in column A. Parameters for the distributions on the respective cost categories are in columns G-J. For the demo worksheet, the simulated values for each distribution appear in column E under the heading Cost.

Interest costs are a function of all cost categories in rows 11-37, the number of days to complete the project, and a stochastic interest rate. Interest rate is assumed to be normally distributed and the number of days to completion are distributed GRK based on the project manager's subjective distribution. (Days to completion could be generated from a Project Management simulation model.)

The bid analysis should include the contractor's profit as an explicit cost category to calculate total costs. In an effort to be the low bidder, however, the firm may be willing to accept a reduced profit. This strategy is risky, particularly if unforeseen cost overruns erode the contractor's reserve for profit and risk returns. To quantify this risk a cost overrun calculation is recommended. Cost overrun is calculated using the formula

$$= \text{IF} ((\text{Total Cost} > \text{Bid Price}), (\text{Total Cost} - \text{Bid Price}), 0)$$

In the demo program the Simetar Scenario option was used to analyze the probability distribution for cost overruns under alternative bid assumptions. Four bids of \$190K, \$195K, \$200K, and \$205K are simulated based on expected total cost of \$198,525. The simulation results for the four bid scenarios are saved in the Cost Overrun worksheet. (The cell E47 in Sheet 1 is the KOV.) The results for the four bid scenario simulation quantify the PDF for potential cost overruns given the bids. CDFs and a Target Probability analysis of the Cost Overrun PDFs are included in Sheet 1 of Bid Analysis Demo.XLS for ease of reference.

Results of the four bids are significantly different. The probability of a cost overrun ranges from 21 to 87 percent. The probability of a cost overrun exceeding \$10,000 ranges from 0 to 36 percent depending on the bid. If the contractor thinks that the only way to win the contract is with a bid of \$200,000 and the 6.5 percent chance of a \$10,000 or greater cost overrun is acceptable then he/she will submit a bid. A \$190,000 bid has a 36 percent chance of a \$10,000 or greater cost overrun which may discourage the contractor from submitting a low-ball bid.

Project Feasibility

Economists are frequently called on to develop project feasibility studies. Requests for feasibility studies of new value added ventures such as gasohol plants, livestock processing plants (e.g., cattle, goats, sheep, poultry), cotton gins, grain elevators, flour mills, etc. appear to be growing at an ever increasing rate. As farmers and agribusiness try to gain a larger part of the consumer's dollar, their interest in value added ventures will expand. The purpose of this section is to develop and demonstrate a comprehensive project feasibility simulation model. A secondary purpose is to suggest ways in which project feasibility models can be used to manage the project if it is undertaken.

Feasibility studies have historically ignored the presence of risk, which is interesting in that, many proposed projects seek to reduce producer's income risk. Although Richardson and Mapp's feasibility study for a business demonstrated the usefulness of stochastic simulation for project feasibility more than 20 years ago, the methodology is not widely used. Gray developed a comprehensive stochastic simulation model of an agribusiness to test management strategies but not as a feasibility analysis. Outlaw, et. al. and Gill have recently (2002) used stochastic simulation to analyze the feasibility of gasohol plants in Texas. In the process of doing these two studies more than 50 feasibility studies from across the country were reviewed and none

used a stochastic simulation model. Most all feasibility studies use deterministic simulation models that simulate the annual income, expenses and cash flows of a business.

Using a stochastic simulation model for a project feasibility study offers many advantages over simple deterministic simulation models. The deterministic modeling approach ignores risk so to compensate most economists use sensitivity analyses on critical values or resort to Best Case-Worst Case scenarios. Another drawback to deterministic modeling is that the studies result in a single answer for the project's feasibility and one has no idea if the answer is the mean of the distribution or not.

The advantage to using a stochastic simulation approach is that the feasibility of a project is reported as a distribution of possible outcomes, not one value. Therefore whether the project fails or succeeds, the analyst can say, "I was right, I told you there was an X% chance of failure or there was a Y% chance of success." The additional cost of doing a stochastic feasibility study over a deterministic study is small given the availability of computers, the wide spread use of Excel for calculating (simulating) income/expenses and cash flows for new projects, on the ease of using simulation Add-Ins for Excel. Besides most deterministic models for feasibility studies use a spreadsheet approach to calculate the pro forma financial statements.

– **Project Feasibility Model**

Project feasibility models can range from very simple to very complex simulation models of a proposed investment. The KOV is always, "will it be profitable?" This leads to defining several KOVs in the model such as:

- Probability of solvency,
- Probability of a positive net present value, or economic success,
- Probability of increasing real net worth,
- Probability that the return on equity exceeds x,
- Probability of annual cash flow deficits or refinancing,
- Net present value (NPV),
- Annual net worth,
- Annual cash reserves,
- Annual net cash income,
- Annual expenses (by category and total), and
- Annual receipts (by category and total).

The list of KOV's is similar to the pyramid in Chapter 2 that shows the steps to building a simulation model, start with the most important KOV and work backwards to define secondary KOV's needed to calculate the preceding KOV.

The next step in developing a stochastic simulation model for project feasibility is to gather data to define the input/output relationships. The amount of raw input per unit of saleable output depends on the proposed plant size. In some cases, the availability of raw inputs may dictate the size of the plant, for example a closed cooperative may limit inputs to the production of raw materials from its member/owners. In any event, the input/output relationship and plant size will dictate the amount of raw input required and output to generate. Next one must determine if the input/output relationship is fixed or stochastic and if it is stochastic obtain parameters to simulate

the relationship. Generally there is a fixed input/output relationship due to some biological or technology constraints that dictate the value and hold it constant.

Once plant size and the input/output relationship is determined, the analyst must develop costs for the given plant size. The overall fixed cost of land, buildings, machinery, trucks, etc. for the project can be obtained through various sources, including information for similar plants in other regions and costs for local sites. Operating costs will be unique for the plant size and input/output relationship so a matrix of costs must be developed if multiple plant sizes are being considered. It is recommended that operating costs be separated into functional categories rather than estimated as large aggregates. For example, an estimate of electricity usage is more reliable than an estimate for all energy (electricity, natural gas, diesel, and gasoline) for a facility. It is also recommended that costs be broken down into the quantity purchased and its per unit price. This approach facilitates stochastic simulation, i.e., electricity cost = electrical use kwh * stochastic price per kwh.

Parameters for all stochastic variables can be estimated from historical values or the experience base of experts. Historical prices for inputs such as electricity, labor, fuel, and raw input (corn, steers, etc.) and price of the output (ethanol, beef, etc.) can be used to estimate parameters for these probability distributions. Where historical data are not available, such as quantities used for various inputs, then use the experts in the field to define GRK distributions. The more narrowly the analyst defines the cost categories the easier it is to find an expert to specify the distribution in question. For example, plant managers can more easily define ranges on electrical rates than for all fuel costs; or more easily specify wage rates by function area than total salary for the plant.

The cost of the raw input and the price of the output are the two most important stochastic variables to include in the model. Other variables that are subject to change from one period to another, and constitute a large portion of costs should be simulated stochastically. Once the parameters for the stochastic variables are determined the analyst can verify and validate the model and proceed with the analysis. A demonstration feasibility study is presented next.

– Preliminary Project Evaluation

The feasibility study presented in Project Feasibility Demo.XLS is abstracted from a feasibility study developed at Texas A&M (Gill). The original model was used to analyze the economic viability of developing a 30 million gallon ethanol plant in Texas (Gill). The model has all of the components outlined above as well as several components unique to the area. For example, Texas is a corn deficit region so the stochastic corn price is simulated using risk about a national average FAPRI forecast plus a local wedge (a) or

$$PCorn_{TX} = a + \widetilde{PCorn}_{Nat.}$$

Another aspect of the feasibility study is that the stochastic price for DDGS was simulated using the relationship between DDGS and soybean meal. This is done much like the corn price because there is no national or state forecasts of annual DDGS prices.

The conversion rate between corn and ethanol is held constant as this is a fermentation process not subject to external forces such as weather. A learning curve is used to reduce output less and less, relative to capacity, each of the first three years as the plant comes on line.

Variable costs are separated into 10 categories that corresponded to ethanol plant costs widely available in the literature. Of the input costs only prices for corn, electricity and natural gas are stochastic. Trend projections for costs of electricity and natural gas are used to develop means for these random variables. Annual average price projections for corn and soybean meal came from FAPRI. Mean ethanol prices are specified as constant values that could be changed via a scenario analysis table. A MVE distribution of the stochastic prices is developed using residuals about their respective trend lines (see Stochastic worksheet).

The proposed firm is assumed to be taxed as a corporation. The federal income taxes show up as a cash outflow and no state income taxes are calculated. (The federal excess tax exemption for ethanol is ignored in the model.) Dividends are calculated as a fixed fraction of positive net cash income. Ending cash balance is an asset if positive and a liability if negative. Repayment of cash flow deficits appears in the next year's cash flow and interest to finance a negative cash flow is an expense in the next year.

The balance sheet includes cash reserves and other assets as well as the remaining balance of the original loan for the plant. Net worth is calculated and in turn is used to calculate net present value:

$$NPV = \sum \frac{\text{Dividends}_t}{(1+i)^t} - \frac{\text{Beginning Net Worth}}{\text{Net Worth}} - \frac{\text{Ending Net Worth}}{(1+i)^{10}}$$

A switch variable is used to count the number of times NPV is positive so the probability of economic success can be calculated directly by Simetar. The probability of negative annual cash flows is calculated using a counter for each year that is 1.0 if cash flow is negative and 0.0 otherwise. Other KOVs can be added to the model to suit the decision maker's interests.

– Uses of a Project Feasibility Model

Once a feasibility model is completed it should be verified and validated thoroughly. Remember, lots of money, your reputation, and job are riding on the model so check it closely. After the model is verified then proceed with the feasibility study.

Scenario or sensitivity analyses should be done on variables that are crucial to success, such as:

- Project cost,
- Loan terms: years, interest rate and down payment,
- Dividend fraction,
- Inflation rate for variable costs,
- Input/Output relationship,
- Variable costs,
- Policy variables as state and federal subsidy rate, and
- Mean levels of stochastic prices.

Charts of the NPV, present value of ending net worth, and ending cash reserves can easily be developed to communicate the results to potential investors. Be careful not to overwhelm them with numbers. Hold the scenario analyses to the end.

After a project is completed the feasibility model can be used to test alternative management strategies for the business. Additionally, as the plant is being built, refinements to the actual plant cost estimates can be made and the feasibility analysis updated. This updating activity is highly recommended if early estimates of plant costs are exceeded during plant construction. Also, updated information can give investors early warning if the distribution on returns is shifting to far to the left, given cost over-runs during construction.

Inventory Management

All business managers face inventory management decisions. When to re-order and how much to order are the variables of interest for inventory management. Factors that should be considered when formulating a inventory management rule are:

- Cost of storage or the holding cost
- Cost of placing an order
- Cost of lost sales due to losing a customer
- Delivery time from the time the order is placed
- Can demand be back lagged if inventory runs out?

An inventory management rule considers the factors listed above to establish the two parameters in the rule:

- Amount to order
- Reorder point

The amount ordered when an order is placed is an obvious part of the rule. The reorder point is the level of inventory when an order is placed. To develop an inventory management rule when demand per period is random a simulation model can be developed.

An inventory management simulation model is a periodic model (weekly, monthly, etc.) based on the frequency of inventory checking. The problem should be simulated for numerous periods, say 50 or more, to test the economic benefits of alternative inventory management rules. Alternative management rules are specified, simulated, and analyzed to determine which one is preferred in terms of reducing total costs. The alternative management rules can be specified as if they are alternative management scenarios and simulated by Simetar.

The example inventory management model in Figure 14.3 assumes a period of one week, weekly demand is stochastic and is distributed $N(40,6)$, shortage cost is \$10/unit, and holding cost is \$3/unit/week. Ordering costs are: \$200 fixed cost to place an order, per unit purchase cost is \$4, and delivery takes one week. Beginning inventory for week 1 is 100 units and the order up to this amount is 150 units, which defines the largest order permitted. The five different reorder points to test for an inventory management rule are 50, 60, 70, 80, and 90.

The simulation model starts with stochastic weekly quantity demanded values in row 10. Beginning inventory in row 13 equals ending inventory for the previous week (row 19). The quantity ordered each week (row 14) is solved using an if statement:

$$=IF(\text{Beg. Invent.} < \text{Reorder point}, (\text{Order up to Amount} - \text{Beg. Inventory}), 0)$$

Amount received in each week (row 15) is the amount ordered in the previous week. Available supply equals beginning inventory plus the units received for each week. Weekly sales equal the lesser of stochastic demand or available supply or:

$$=IF(\text{Demand} < \text{Supply}, \text{Demand}, \text{Supply})$$

Inventory at the end of the week equals available supply minus sales. Lost sales are calculated as stochastic demand minus sales.

	A	B	C	D	E	F	G	H	
1	Price of Good Sold	\$ 23							
2	Shortage Cost	\$ 10	Mean Weekly Demand	40					
3	Holding Cost	\$ 3	Std Dev Weekly Deman	6					
4	Fixed Cost	\$ 200							
5	Unit Purchase Cost	\$ 4							
6	Beginning Inventory	100							
7	Order up to this amount	150							
8	Reorder Point	50	Reorder Point Scenario Table						
			50	60	70	80	90		
9	Generate a Common								
10	Stochastic Demand	40.48	35.15	41.31	42.50	40.10	43.11	33.30	
11	One Week Delivery								
12	Week	1	2	3	4	5	6	7	
13	Beginning Inventory	100	59.5	24.4	0.0	83.1	193.0	149.9	
14	Quantity Ordered	0.0	0.0	125.6	150.0	0.0	0.0	0.0	
15	Amount Received	0.0	0.0	0.0	125.6	150.0	0.0	0.0	
16	Available Supply	100.0	59.5	24.4	125.6	233.1	193.0	149.9	
17	Stoch. Demand	40.5	35.2	41.3	42.5	40.1	43.1	33.3	
18	Sales	40.5	35.2	24.4	42.5	40.1	43.1	33.3	
19	Ending Inventory	59.5	24.4	0.0	83.1	193.0	149.9	116.6	
20	Lost Sales	0.0	0.0	16.9	0.0	0.0	0.0	0.0	
21	Cost								
22	Storage Cost	300.0	178.6	73.1	0.0	249.4	579.1	449.8	
23	Order Cost	0.0	0.0	200.0	200.0	0.0	0.0	0.0	
24	Purchase Cost	0.0	0.0	502.5	600.0	0.0	0.0	0.0	
25	Penalty for Shortage	0.0	0.0	169.4	0.0	0.0	0.0	0.0	
26	Total	300.0	178.6	945.1	800.0	249.4	579.1	449.8	
27	Revenue	910.8	791.0	548.3	956.3	902.3	969.9	749.3	
28	Profits	610.8	612.4	-396.8	156.3	652.9	390.8	299.6	
29	KOVs								
30	Average Cost	485.1							
31	Average Inventory	75.1							
32	Average Lost Sales	1.7							
33	Average Revenue	843.5							
34	Average Profit	358.4							

Figure 14.3. Example of an Inventory Management Simulation Model.

Costs for the business are broken into several categories (Figure 14.3). Storage cost is the per unit weekly holding cost times beginning inventory. Order cost equals the fixed cost of placing an order, if an order is placed that period. Purchase cost is the per unit cost of the item times the number ordered. If a shortage occurs the penalty cost equals the units of lost sales times the shortage cost. The shortage cost is used to penalize an inventory management scheme that tries to maximize profits by holding down storage costs. (Shortage cost may be zero if lost sales can be back lagged).

Revenue to the business is simply quantity sold times price per unit of good sold (Figure 14.3). Profit equals total revenue minus the sum of costs. Five possible KOVs are suggested for evaluating the five inventory management rules specified in the Scenario Table. The manager may want to maximize average weekly profit over the period or minimize average weekly total cost. The model is simulated and results of a dynamic certainty equivalents analysis are available in Inventory Management Demo.XLS.

The inventory simulation model can be modified to handle more complex inventory management scenarios. For example, delivery time could be multiple periods (2 or 3 weeks) and delivery time could be made stochastic as well. Anticipated inflationary costs for the goods purchased can be changed each period. Shortage cost can be a function of lost sales so it is zero for small shortfalls but very large if shortfalls exceed a value, say 10 units per week.

The final point to consider is how many periods should be included in the model. The example in Inventory Management Demo.XLS uses 10 periods. The correct number of periods depends on the lag time from order to delivery and the order up to this amount and maximum order amount. Generally the model should include sufficient periods for inventory to reach equilibrium with any management rule being simulated. This may be 50 or more periods for most problems.

References

- Bailey, D., B.W. Brorsen, and J.W. Richardson. "Dynamic Stochastic Simulation of Daily Cash and Futures Cotton Prices." Southern Journal of Agricultural Economics, 16-2(1984): 109-116.
- Bell, David E. and A. Schleifer, Jr. Risk Management. Cambridge, MA: Course Technology, Inc., 1995.
- Bell, D.E. and A. Schleifer, Jr. Decision Making Under Uncertainty. Cambridge, MA: Course Technology, 1995.
- Eichberger, J. and I.R. Harper. Financial Economics. Oxford: Oxford University Press, 1997.
- Ferrara, W.L. and J.C. Hayga. "Toward Probabilistic Profit Budgets." In Business Simulation for Decision Making. W.C. House, editor, New York: PBI, 1977.
- Gerloff, D.C. "Defining Efficient Commercial Loan Portfolios for Regional and Multi-Regional Lenders." Texas A&M University, Department of Agricultural Economics, Ph.D. Dissertation, 1992.
- Gill, R.C., II. "A Stochastic Feasibility Study of Texas Ethanol Production: Analysis of Texas State Legislature Ethanol Subsidy Proposal." Texas A&M University, Department of Agricultural Economics, M.S. Thesis, December 2002.
- Gray, A.W. "Agribusiness Strategic Planning Under Risk." Texas A&M University, Department of Agricultural Economics, Ph.D. Dissertation, August 1998.
- Lerner, E.M. "Simulating a Cash Budget." In Business Simulation for Decision Making. W.C. House, editor, New York: PBI, 1977.
- Mullick, S.K. and D.P. Haussener. "Production Decisions for New Products." In Business Simulation for Decision Making. W.C. House, editor, New York: PBI, 1977.
- O'Brien, D., M. Hayenga, and B. Babcock. "Deriving Forecast Probability Distributions of Harvest-Time Corn Futures Prices." Review of Agricultural Economics, 18(1996): 167-180.
- Outlaw, J.L., D.P. Anderson, S.L. Klose, J.W. Richardson, et. al. "Comprehensive Ethanol Feasibility Study." Texas A&M University, Department of Agricultural Economics, December 2002.
- Ragsdale, C.T. Spreadsheet Modeling and Decision Analysis. Cincinnati, OH: South-Western College Publishing, 1998.
- Richardson, J.W. and H.P. Mapp, Jr. "Use of Probabilistic Cash Flows in Analyzing Investment Under Conditions of Risk and Uncertainty." Southern Journal Agric. Econ. Dec. 1976.

Richardson, W.A. "Stochastic Analysis of Selected Hedging Strategies for Cotton in the Texas Southern High Plains." Texas A&M University, Department of Agricultural Economics, Senior Honors Thesis, April 2002.

Scott, D.F., Jr. and L.J. Moore. "Financial Planning in a Simulation Framework." In Business Simulation for Decision Making. W.C. House, editor, New York: PBI, 1977.

Chapter 15 Probabilistic Forecasting

Much of the emphasis in this book is on estimating parameters for the probability distribution of a random variable. Students of simulation are advised to develop the best model possible for explaining or predicting the deterministic component of each random variable, so as to minimize the stochastic component. Little has been said thus far about different models for forecasting and parameter estimation. The purpose of this chapter is to present several forecasting techniques for estimating the parameters of the deterministic component for random variables and using these parameters to develop probabilistic forecasts.

Probabilistic Forecasting

Probabilistic forecasts are generally inter-temporal forecasts of a random variable with stochastic components incorporated into the forecast. For example, if the deterministic model forecasts annual quantities sold, a probabilistic forecast would provide forecasts of the probability distributions for future annual sales. Assume that the annual sales data is explained by a linear trend, then the deterministic forecast model is

$$\hat{Y}_t = \hat{a} + \hat{b}T_t \text{ for } t = 1, 2, 3, \dots, T$$

and the probabilistic forecast for year $T + i$ is

$$\tilde{Y}_{T+i} = \hat{a} + \hat{b}T_{T+i} + \tilde{\sigma} * \text{SND}$$

where $\tilde{\sigma}$ represents the stochastic component of Y and SND is a stochastic standard normal deviate. Probabilistic forecasts thus incorporate risk into forecasting.

Probabilistic forecasting techniques involve several steps after the data are collected and checked for errors.

- Select the appropriate forecasting technique (or just try all of them).
- Estimate parameters for the forecasting technique, \hat{a} and \hat{b} .
- Estimate the residuals of the observed values minus the forecast values to quantify the forecast error. Residuals are the stochastic part of the forecast, $\hat{e}_t = Y_t - \hat{Y}_t$ for all t .
- Select the forecasting technique with the smallest forecast error (MAPE) and use it to forecast the deterministic component of the random variable, \hat{Y}_{T+i} .
- Simulate the stochastic component of the forecast and add it to the deterministic forecast for a probabilistic forecast, or $\tilde{Y}_{T+i} = \hat{Y}_{T+i} + \tilde{e}_{T+i}$.
- Present the probabilistic forecast showing the risk in each period.

The format used throughout this chapter is to describe a forecasting technique, present an Excel worksheet example, develop a deterministic forecast, and then develop a stochastic forecast using the deterministic forecast as the mean. The forecast techniques described in this chapter are:

- Trend regression
- Multiple regression
- Seasonal forecasts using dummy variables
- Seasonal forecasts using harmonic regression
- Seasonal indices
- Cycles
- Moving Average
- Time series decomposition -- additive and multiplicative
- Exponential smoothing -- simple, Holt, Holt-Winters, multiplicative and additive, dampened
- Time series analysis -- autoregressive and vector autoregressive

Quantifying Forecast Error

The forecast error is the stochastic portion of the variable and is calculated as the residuals (\tilde{e}_t) for each historical value. Three measures of forecast error are described in this section. Start with the residuals between the observed and historical values over the historical forecast period or

$$e_t = Y_t - \hat{Y}_t$$

for $t = 1, 2, 3, \dots, T$.

Mean absolute error is the average absolute residual over the historical period, (T).

$$\text{MAE} = \left(\sum_{t=1}^T |e_t| \right) / T$$

Root mean square error is like a standard deviation in that it is the square root of the average squared residual over all the historical period.

$$\text{RMSE} = \sqrt{\left(\sum_{t=1}^T e_t^2 \right) / T}$$

Mean absolute percentage error expresses the forecast error as a percentage of the historical observations and is easier to understand. It is recommended that you not use this test if your data have a mean of zero.

$$\text{MAPE} = 100\% * \left(\sum_{t=1}^T \left| \frac{e_t}{Y_t} \right| \right) / T$$

These measures of forecast error are available as functions in Simetar. The format for each function is specified as:

- Mean absolute error (MAE)
= MAE (array of residuals)
- Root mean square error (RMSE)
= RMSE (array of residuals)
- Mean absolute percentage error (MAPE)
= MAPE (array of residuals, array of observed values)

where array of residuals refers to the 1 to T array of \hat{e}_t values, and array of observed values refers to the 1 to T observed values used to calculate the residuals.

All three measures of forecast error are tied to the same variable, so a forecasting procedure that minimizes one usually minimizes the other two, as well. Select one measure and use it to determine which forecasting procedure is best for a particular variable. The MAPE procedure appears to be preferred by practicing forecasters because it expresses the forecast error as a percent of the variable to forecast.

Trend Regression Forecasts

A simple regression model can be used to project a random variable if a trend exists in the series. The parameters for the trend regression are estimated using Simetar's multiple regression function. The model estimated for a linear trend projection is

$$\hat{Y}_t = \hat{a} + \hat{b}T_t + e_t$$

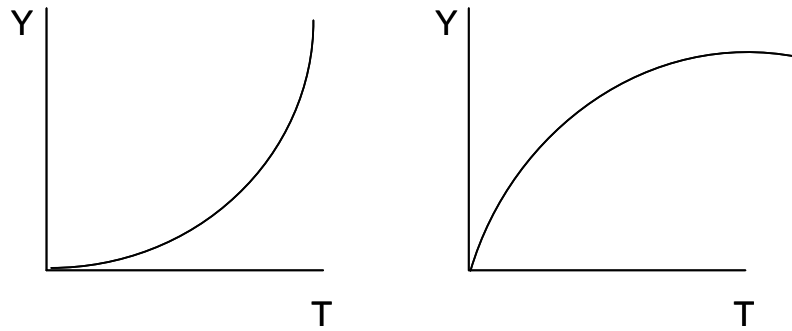
where T_t is the trend variable that increments by 1 unit for each period.

The trend variable could be a series of values such as 1, 2, 3, ..., T or it could be the years for the data such as 1950, 1951, ..., 2005. The use of years for T is useful because you always know the value to use to forecast a particular year. For example, to forecast Y in 2010 the formula is

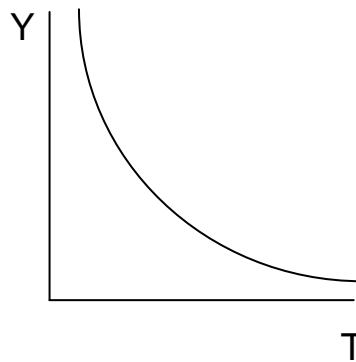
$$\hat{Y}_{2010} = \hat{a} + \hat{b}(2010).$$

When the data have a non-linear trend, add a second or third trend variable to the OLS equation to capture the effect. The model to estimate becomes:

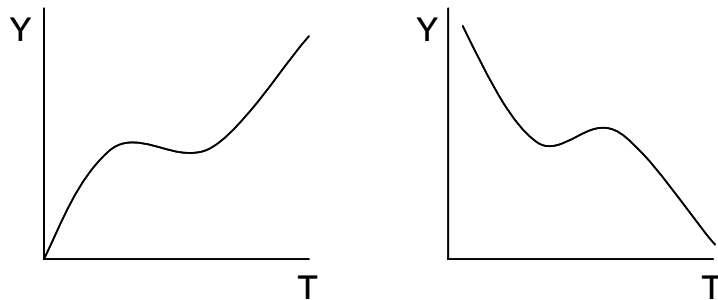
- Non-linear increasing trend as: $\hat{Y}_t = \hat{a} + \hat{b}_1 T_t + \hat{b}_2 T_t^2 + e_t$



- Non-linear decreasing trend as: $\hat{Y}_t = \hat{a} + \hat{b} \frac{1}{T_t} + e_t$



- Two or more changes in trend as: $\hat{Y}_t = \hat{a} + \hat{b}_1 T_t + \hat{b}_2 T_t^2 + \hat{b}_3 T_t^3 + e_t$



The residuals from a trend regression are:

$$\hat{e}_t = Y_t - \hat{Y}_t$$

The $\hat{\epsilon}_t$ values are referred to as “de-trended” data, because the residuals represent the portion of the random variable not explained by trend. If the trend regression has a high R^2 and low MAPE the residuals can be used with confidence to quantify the stochastic portion of the variable.

An example of using a linear trend regression to develop deterministic and probabilistic forecasts is provided in the Trend Forecasts Demo.XLS workbook. The monthly Amarillo steer prices have a statistically significant trend as evidenced by a t statistic of 4.096 on the slope coefficient (Figure 15.1). The cyclical and seasonal variability about the trend is quite large as evidenced by a MAPE of 12.84 percent and a standard deviation on the residuals of \$11.92/cwt. The deterministic projection increases price \$0.058/cwt every month regardless of the past seasonal and cyclical variability so this is not the best method for projecting this series.

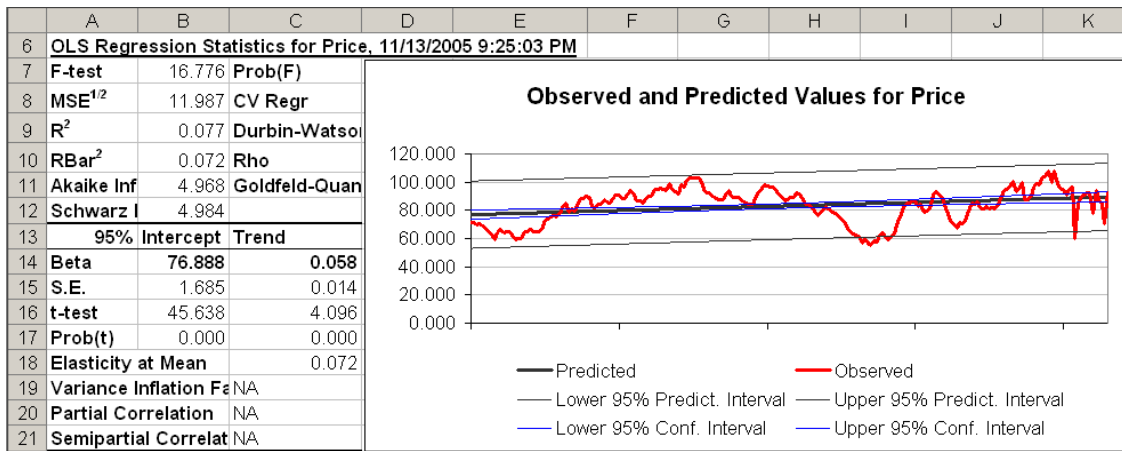


Figure 15.1. Example of a Linear Trend Regression.

A non-linear trend projection of the series is also included in Trend Forecasts Demo.XLS. The results of a non-linear trend regression (Figure 15.2) show the benefit of adding T^2 and T^3 to the regression when the variable has a non-linear trend. The R^2 is 56.8 percent, all of the betas have highly significant t statistics and the MAPE is reduced to 8.34 percent.

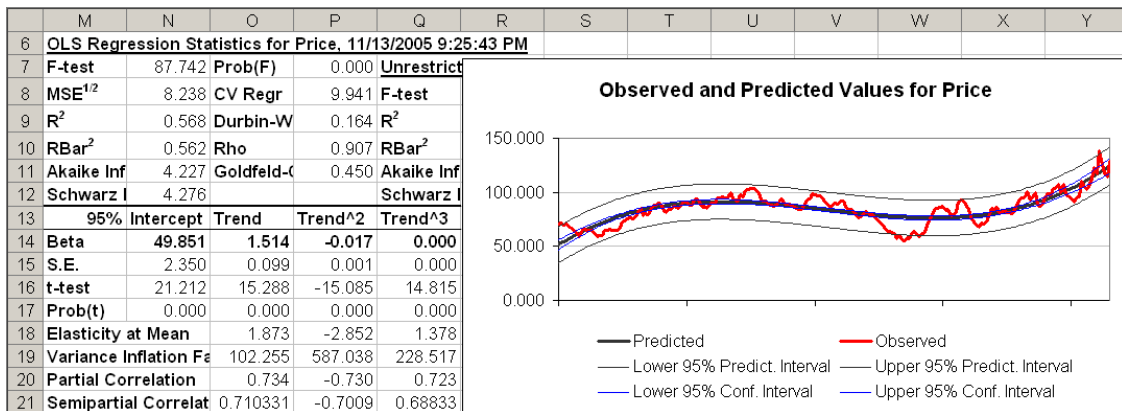


Figure 15.2. Example of a Non-Linear Trend Regression.

The probabilistic forecast from a trend regression is made assuming the residuals are normally distributed with mean zero and standard deviation equal to the standard deviation of the residuals (or S.D. Residuals) from the regression. In other words, the deterministic forecasts presented in Figures 15.1 and 15.2 are simulated using the following formulas for period 205, as there are 204 observations in the data set. The deterministic forecast formula is:

- Linear Trend

$$\hat{Y}_{205} = \hat{a} + \hat{b} * (205) = 76.888 + 0.058 (205) = 88.854$$

- Non-linear Trend

$$\begin{aligned} \hat{Y}_{205} &= \hat{a} + \hat{b}_1 (205) + \hat{b}_2 (205)^2 + \hat{b}_3 (205)^3 \\ 108.382 &= 49.851 + 1.514 (205) + (-0.017) (205) + 0.00005327 (205) \end{aligned}$$

The stochastic forecast formula for the linear and non-linear regressions are:

- Linear Trend

$$\begin{aligned} \tilde{Y}_{205} &= \hat{Y}_{205} + \text{SEP} * \text{SND} \\ \tilde{Y}_{205} &= 88.854 + 11.92831 * \text{SND} \end{aligned}$$

- Non-Linear Trend

$$\begin{aligned} \tilde{Y}_{205} &= \hat{Y}_{205} + \text{SEP} * \text{SND} \\ \tilde{Y}_{205} &= 108.382 + 8.15683 * \text{SND} \end{aligned}$$

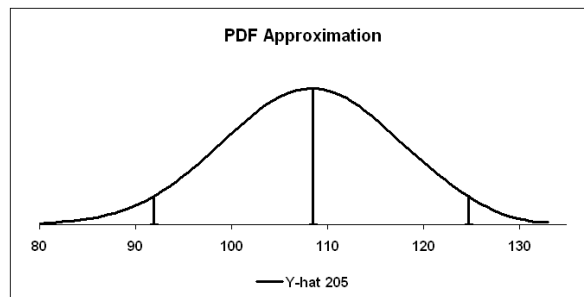


Figure 15.3. The Probabilistic Forecast for \hat{Y}_{205} .

where SEP is the standard error of prediction for the respective regression models. The multiple regression option in Simetar will simulate both the deterministic and stochastic forecasts if the X matrix contains more values than the Y vector. See the regression results in Trend Forecasts Demo.XLS in cells A229-B240. The stochastic or probabilistic forecast uses the SEP rather than the standard deviation for the residuals because the forecasts are out of sample forecasts. If the trend regression model provided an acceptable model for forecasting the series, then probabilistic forecasts using the above formulas will provide acceptable forecasts. The PDF in Figure 15.3 depicts the probabilistic forecast for \tilde{Y}_{205} in the non-linear trend regression.

Multiple Regression

Multiple regression models are useful for forecasting variables that are part of a system. In these cases, structural models add to the explanatory ability of a simple trend regression by including other variables to help explain the variability of the dependent variable. Econometric models of crops and livestock rely on structural regression models to forecast the endogenous variables in the system. For example, in a U.S. wheat model one might forecast planted acres as a function of exogenous variables such as: expected price, idled acres, trend, and a lagged

dependent variable, or:

$$\hat{Y}_t = \hat{a} + \hat{b}_1 P_{t-1} + b_2 I_t + b_3 T + b_4 Y_{t-1} + e_t$$

Forecasting with this type of model presents a problem because one must project values for the exogenous variables to forecast the endogenous variable. This problem is usually overcome in a complete econometric model because separate equations for the exogenous variables would be included.

An example of estimating and forecasting with a multiple regression model is provided in Multiple Regression Forecasts Demo.XLS. A summary of the results in Figure 15.4 suggest that the lagged price and planted acres are the most explanatory variables. The idled acres (CRP_t) is included for policy analysis and trend (years) was restricted out because lagged acres did a better job (had a higher t statistic). Overall an R² of 81.5% and an F of 39.55 are reasonable. The 3.98% MAPE is encouraging that on average the model has an absolute error of about 4%.

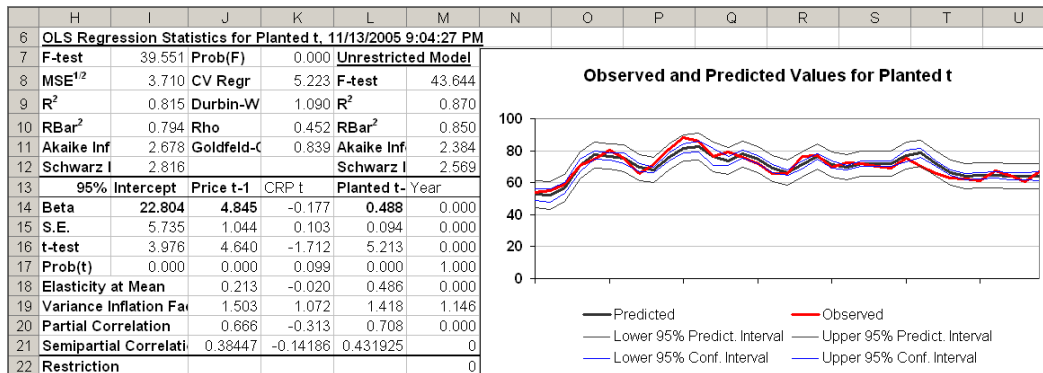


Figure 15.4. Using Multiple Regression for Forecasting.

For the wheat acreage equation a one period ahead deterministic forecast is:

$$\hat{Y}_{2006} = 22.804 + 4.845 * (P_{2005}) - 0.177 (CRP_{2005}) + 0.488 (Planted_{t-1})$$

The deterministic forecast yields a value of 65.67 million acres when P₂₀₀₅ = 2.88, CRP = 9.6 and Y = 61.0. A probabilistic forecast for one period ahead, using the standard error of the prediction (SEP) as a measure of the unexplained variability is presented in Figure 15.5 and is simulated as:

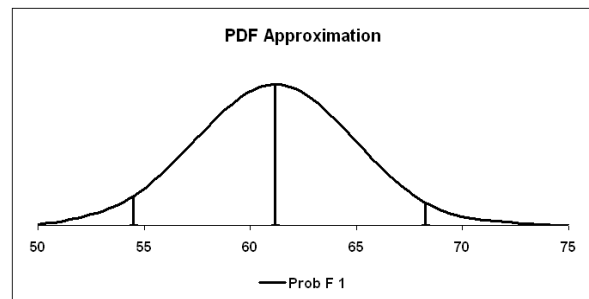


Figure 15.5. Probabilistic Forecast for Assumed Values of Exogenous Variables.

$$\tilde{Y}_{2006} = \hat{Y}_{2006} + SEP * SND$$

or

$$\tilde{Y}_{2006} = 65.67 + 3.58 * SND$$

Seasonal Forecasts Using Dummy Variables

Seasonal variability is within year variability that repeats itself each year. Monthly and quarterly data generally exhibit a seasonal pattern. Seasonal patterns are caused by production and demand patterns tied to weather, holidays, or tradition. Many agricultural price series are seasonal in nature due to their production pattern. For example, wheat prices tend to be low at harvest and rise throughout the marketing year until near harvest the next year.

Business sales also show a seasonal pattern. Agribusiness firms that sell inputs see larger sales pre-planting than post-planting or during harvest. Their sales may also show an end of year jump as farmers pre-pay for inputs to reduce income taxes. Retail sales firms observe seasonal patterns often due to seasonal travel and holiday purchases.

Two regression based methods for forecasting seasonal patterns are presented here: dummy variables without a trend and dummy variables with a trend. Dummy variable regression models use 0's and 1's to identify seasons. In the case of monthly data, the dummy variable model would include 11 dummy variables with the affect of the missing month being captured in the intercept. The dummy variable for January would have a one when the observation is for January and zero otherwise. The same pattern would be used for each of the other 10 months. The regression model to be estimated is:

$$Y_t = \hat{a} + b_1 \text{Jan}_t + \hat{b}_2 \text{Feb}_t + \dots + \hat{b}_{11} \text{Nov}_t + b_{12} \text{DVH}_t + e_t$$

where DVH is a dummy variable for months with holidays, if that is relevant for the data series.

An example of using a seasonal dummy variable regression model without trend to analyze monthly Amarillo steer data is provided in Figure 15.6. (See Regression For Seasonal Forecasts Demo.XLS for the complete analysis summarized in Figure 15.6). Each of the monthly dummy variables is statistically significant as evidenced by large t statistics. The F statistic is quite large (735) and the R^2 is 97.9. The Durbin-Watson statistic shows considerable autocorrelation.

Forecasting with the seasonal model simply calls for substituting in values (0 and 1) for the months or:

$$\hat{Y}_{\text{Jan}} = 82.389 * (1.0)$$

$$\hat{Y}_{\text{June}} = 83.822 * (1.0)$$

$$\hat{Y}_{\text{Dec}} = 81.414 * (1.0)$$

Probabilistic forecasts are accomplished using the standard error of the prediction (SDP) for the regression model and the deterministic forecasts.

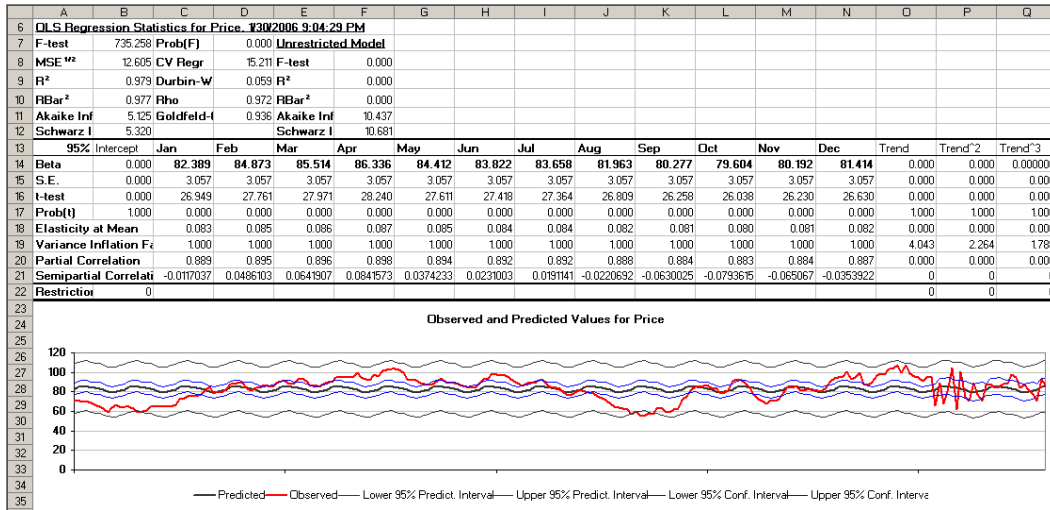


Figure 15.6. Seasonal Forecast Using Dummy Variables Without Trend.

$$\hat{Y}_{Jan} = 82.389 + 12.971 * SND$$

$$\hat{Y}_{June} = 83.822 + 13.326 * SND$$

The residuals (\hat{e}_t) from the seasonal dummy variable regression are referred to as de-seasonalized data. The deseasonalized data (d_t) could be analyzed further to try to improve the MAPE. The further analysis could be done using a separate model to analyze the data for a trend as:

$$\hat{d}_t = \hat{a} + \hat{b}_1 T_t + \hat{b}_2 T_t^2 + \hat{b}_3 T_t^3 + e_t$$

However, an easier way to do this is to re-run the seasonal dummy variable regression and include the trend variables as:

$$Y_t = \hat{a} + \hat{b}_1 Jan_t + \hat{b}_2 Feb_t + \dots + \hat{b}_{11} Nov + \hat{b}_{12} T + \hat{b}_{13} T^2 + \hat{b}_{14} T^3 + e_t$$

The results of the seasonal/trend regression with the three trend components is summarized in Figure 15.7 and is provided in Regression For Seasonal Forecasts Demo.XLS. The Durbin-Watson statistic of 0.068 suggests the residuals are still autocorrelated even though the other goodness of fit statistics are quite good.

The deterministic forecast with the model is now a wavy trend line, as indicated in Figure 15.7, and is developed using the following formulas for periods 205 and 206:

$$\hat{Y}_{Jan, 205} = 50.077 + 1.553 (205) - 0.017 (205)^2 + 0.000055 (205)^3$$

$$\hat{Y}_{Feb, 206} = 52.264 + 1.553 (206) - 0.017 (206)^2 + 0.000055 (206)^3$$

·
·
·

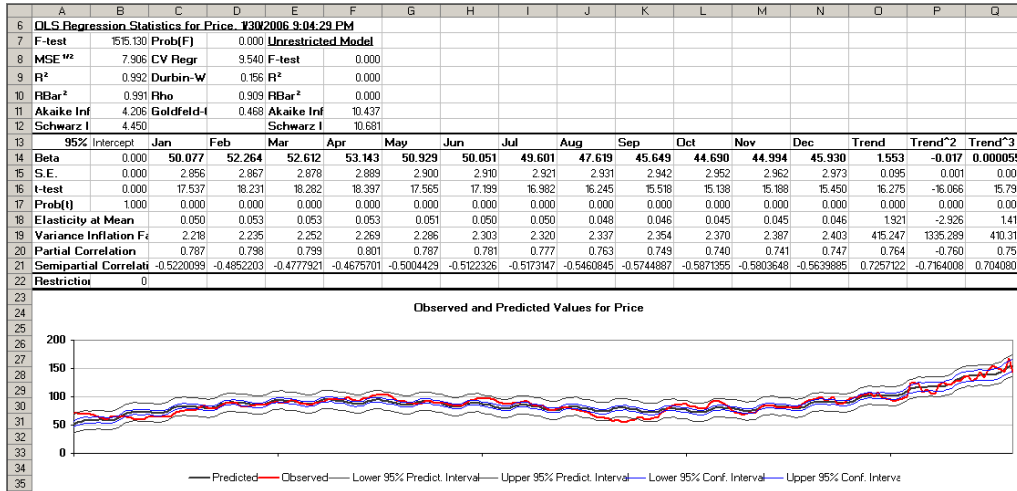


Figure 15.7. Seasonal Forecast Using Dummy Variables With a Trend.

Note that for the forecast, the slope coefficient for the month being forecast is the only seasonal beta that is relevant because the other slope parameters would be multiplied by zero. (To see this examine the X matrix beyond the historical data for the regression in Regression For Seasonal Forecasts Demo.XLS). The probabilistic forecast for the seasonal/cyclical model uses the standard error of the predictions from the regression model in lines 229-255 of the Dummy Variables Worksheet.

$$\hat{Y}_{Jan, 205} = 110.392 + 8.447 * SND$$

$$\hat{Y}_{July, 211} = 118.521 + 8.819 * SND$$

Seasonal Forecasts Using Harmonic Regression

For data series that display a stable seasonal pattern with a linear trend, a regression that uses Sin and Cos functions of time can be used to develop a forecast model. This is referred to as a harmonic regression model. The seasonal length, SL, is set equal to 12 for monthly data and 4 for quarterly data. The harmonic regression model that is estimated is:

$$Y_t = \hat{a} + \hat{b}_1 (2 \pi T_t) + \hat{b}_2 \text{Sin} (2 \pi T_t/SL) + \hat{b}_3 \text{Cos} (2 \pi T_t/SL) + e_t$$

The $2 \pi T_t$ term captures the linear trend in the data and the Sin and Cos terms capture the seasonal pattern in the data.

The X matrix for a seasonal harmonic regression with SL equal to 12 is presented in Table 15.1. Note that the Sin and Cos X's repeat themselves in a regular pattern, thus mimicking a seasonal pattern. The T_i values used to calculate the values in Table 15.1 are 1, 2, 3, ..., 12.

Table 15.1. Seasonal Harmonics for the First 12 Months in the X Matrix.

Month	T	2PiT	Sin (2PiT/SL)	Cos (2PiT/SL)
Jan	1	0.523599	0.50	0.87
Feb	2	1.047198	0.87	0.50
Mar	3	1.570796	1.00	0.00
Apr	4	2.094395	0.87	-0.50
May	5	2.617994	0.50	-0.87
Jun	6	3.141593	0.00	-1.00
Jul	7	3.665191	-0.50	-0.87
Aug	8	4.188790	-0.87	-0.50
Sep	9	4.712389	-1.00	0.00
Oct	10	5.235988	-0.87	0.50
Nov	11	5.759587	-0.50	0.87
Dec	12	6.283185	0.00	1.00

The results of a harmonic regression assuming a seasonal length of 12 months for the Amarillo steer prices is presented in Figure 15.8. The model has a MAPE of 12.80 and the error terms are autocorrelated because the trend in the data is not linear and this model does not capture the cycle in the data.

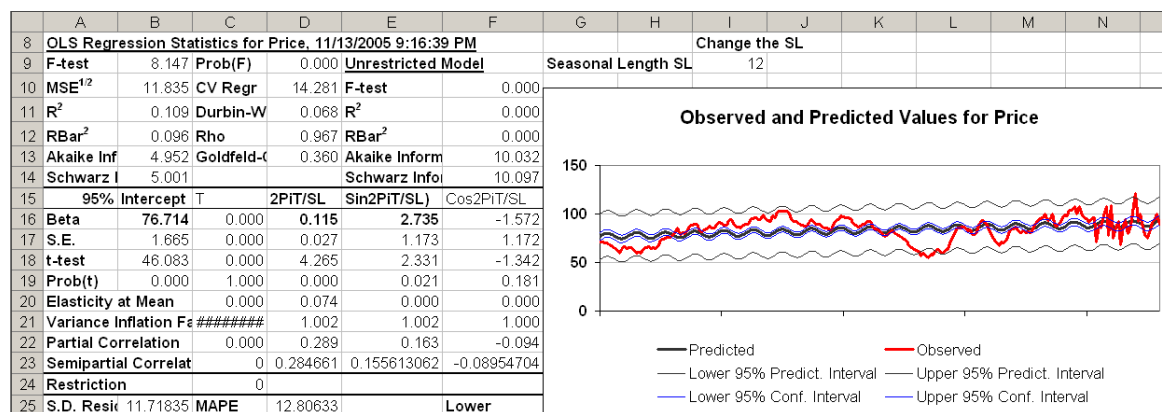


Figure 15.8. Harmonic Regression to Forecast Seasonal Patterns.

Deterministic forecast values from the harmonic regression are easy to generate because the X values are simply functions of time. For example, the Amarillo steer data has 204 observations so to forecast the next observation, 205, the equation is:

$$\hat{Y}_{205} = 76.714 + 0.115 * (2 \pi(205)) + 2.735 * (\sin (2 \pi(205)/12)) - 1.572 * (\cos (2 \pi(205)/12))$$

Note that this seasonal forecast procedure does not depend on matching each period (month) to its particular dummy value. In other words, the harmonic model does not care if the 205th observation is January or March. To this extent the procedure is easier to use than the dummy variable approach, particularly if the data have a linear trend. The example summarized in Figure 15.8 is in the Harmonic Regression worksheet of the Regression For Seasonal Forecasts Demo.XLS workbook.

The probabilistic forecast of the series for month 205 is:

$$\tilde{Y}_{205} = \hat{Y}_{205} + \text{SDP} * \text{SND}$$

$$\hat{Y}_{205} = 89.034 + 12.010 * \text{SND}$$

where SDP is the standard error of the prediction. The PDF for forecasting period 205 using the harmonic regression model is presented in Figure 15.9.

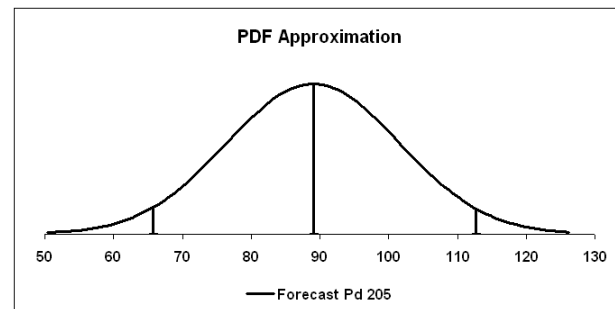


Figure 15.9. Probabilistic Forecast of Period 205 Using Harmonic Regression Model.

Seasonal Indices

Seasonal variability is within year variability that repeats itself each year and is observed for monthly, weekly or quarterly data. Indices that quantify this variability are called seasonal indices. A seasonal index can be used in conjunction with trend or multiple regression models to forecast a data series that exhibits a seasonal pattern. A seasonal index is calculated two ways:

– Simple Average Seasonal Index

-- Calculate the average value for each month (or period such as quarter)

$$\bar{X}_{\text{Jan}} = \sum_{i=1}^T (X_{i\text{Jan}}) / T$$

where $X_{i\text{Jan}}$ is the observed value for each January in the population for years 1 through T.

-- Calculate the overall average for all N observations or

$$\bar{X} = \sum_{i=1}^N (X_i) / N$$

where $N = 12 * T$ years

-- Calculate the index for each month or period as

$$I_{\text{Jan}} = \bar{X}_{\text{Jan}} / \bar{X}$$

$$I_{\text{Feb}} = \bar{X}_{\text{Feb}} / \bar{X}$$

·
·
·

The resulting seasonal index has a mean of 1.00. The seasonal index is useful for converting an annual average to a monthly value. Dividing each period's index value by the number of periods, say, 12 yields the Fractional Contribution Index which is useful for forecasting the monthly totals for annual sums (such as total sales). The seasonal index in Figure 15.10 demonstrates how

Simetar calculates a Simple Average Seasonal Index (see Seasonal Index Forecast Demo.XLS).

	D	E	F	G	H	I	J	K	L	M	N	O	P
8	Cyclical Indices for Monthly Price, 12/15/2005 4:47:48 AM												
9		1	2	3	4	5	6	7	8	9	10	11	12
10	1	71.06	71.47	70.06	70.31	68.75	67.08	64.8	63.12	59.44	63.94	66.62	64.12
11	2	65.12	65.25	62.72	59.15	60.19	60	64.5	65.25	66.13	64.7	64.94	64.68
12	3	66.88	72.6	73.66	75.43	76.38	75.63	77.53	81.5	85.4	78.81	80.71	79.5
13	4	83.66	88	88.3	89.75	89.5	82.8	81.5	84.1	84.25	87.5	85.7	85.33
14	5	89.13	90.88	90.5	88.25	88.4	92.83	93.83	90.7	86.5	87	85.1	88.08
15	6	89.85	89.63	95.13	95.25	95.8	94.63	94.38	99.2	94.75	93.3	91.88	98.17
16	7	96.25	101.75	102.75	103.3	103.19	102.69	99.63	92.94	92.19	91.85	89.38	88.25
17	8	87.44	87.69	91.15	93.88	90	89.4	89.25	88.01	85.75	85.44	84.25	84.13
18	9	88.63	92.88	94.35	98.32	97.44	96.45	96.34	95.07	90.5	88.82	86.5	87.67
19	10	89.57	89.5	92.4	91.88	87.55	84	84.34	83.1	79.32	76.57	77.85	80.08
20	11	82.45	80.51	78.88	78.19	75.9	73.87	71.83	67.4	65	63.5	62.5	61.5
21	12	56.9	60.07	56.49	54.94	58.3	57.28	62.67	63.94	60.47	59.14	62.31	63.01
22	13	71.99	75.8	81.49	85.48	85.15	86.6	86.63	82.98	84	78.79	79.14	81.32
23	14	90.83	93.17	91.86	89.43	83.85	77.815	72.92	70.915	67.275	71.63	70.445	72.835
24	15	79.465	84.82	84.405	86.25	81.755	81.16	83.04	81.215	81.52	80.805	84.18	90.25
25	16	93.675	94.99	96.125	100.36	93.265	95.245	100	87.925	87.22	90.31	96.63	98.975
26	17	97.72	103.825	103.47	107.545	99.585	107.5	99	96	95	91.16	95.135	96.14
27	SUM	1400.62	1442.835	1453.74	1467.715	1435.005	1424.98	1422.19	1393.365	1364.715	1353.265	1363.27	1384.04
28	AVERAGE	82.38941	84.87265	85.51412	86.33618	84.41206	83.82235	83.65824	81.96265	80.27735	79.60382	80.19235	81.41412
29	ST DEV	11.57589	11.9041	12.96476	14.22	12.60769	13.73284	12.45774	11.46838	11.57323	10.96217	10.82638	11.95711
30	INDEX	0.994	1.024	1.032	1.042	1.019	1.011	1.009	0.989	0.969	0.961	0.968	0.982
31	FRAC. CONT.	0.083	0.085	0.086	0.087	0.085	0.084	0.084	0.082	0.081	0.080	0.081	0.082
32	INDEX LCI	0.719	0.749	0.735	0.719	0.726	0.690	0.718	0.715	0.686	0.691	0.703	0.695
33	INDEX UCI	1.270	1.299	1.329	1.365	1.311	1.333	1.301	1.263	1.251	1.230	1.232	1.270

Figure 15.10. Example of a Simple Average Seasonal Index.

– Centered Moving Average Seasonal Index

-- Calculate a moving average centered on the expected seasonal period, i.e., 12 period moving average for monthly data starts in period 6.

$$M_6 = (Y_1 + Y_2 + \dots + Y_{12}) / 12$$

$$M_7 = (Y_2 + Y_3 + \dots + Y_{13}) / 12$$

·
·
·

-- Divide the original observed data by the moving average for the same period.

$$F_6 = M_6 / Y_6$$

$$F_7 = M_7 / Y_7$$

·
·
·

-- Calculate the average fraction for each period across all N years.

$$I_{Jan} = (F_{Jan, 1} + F_{Jan, 2} + \dots + F_{Jan, N}) / N$$

where I_{Jan} is the index for the first period, or January for a monthly data series, and $F_{Jan, i}$ is the fraction for January in year i.

An example of calculating seasonal indices is provided in the Moving Average Index worksheet in the Seasonal Index Forecasts Demo.XLS workbook and is summarized in Figure 15.11. The seasonal index indicates that the February price is 97.2% of the average annual price and October prices are 104.7% of the annual average price.

	A	B	C	D	E	F	G	H	I	J	K	L	M
7	Cyclical Indices, 1/30/2002 4:02:03 PM												
8		1	2	3	4	5	6	7	8	9	10	11	12
9	1							1.025977	1.045258	1.100465	1.010961	0.957958	0.985145
10	2	0.965295	0.96454	1.009307	1.075472	1.056252	1.058819	0.986447	0.980926	0.979397	1.018573	1.035642	1.060303
11	3	1.04328	0.977887	0.983907	0.979258	0.983378	1.009983	1.002214	0.969847	0.940213	1.034143	1.023964	1.050183
12	4	1.003511	0.957131	0.954563	0.942642	0.951643	1.034093	1.056365	1.027844	1.028528	0.990657	1.0102	1.018941
13	5	0.985952	0.975646	0.983817	1.009726	1.007494	0.96038	0.951686	0.984284	1.033704	1.033333	1.063455	1.031827
14	6	1.01259	1.019283	0.967689	0.972835	0.972938	0.992396	1.002305	0.961383	1.015215	1.037991	1.061035	0.99961
15	7	1.025312	0.969476	0.956464	0.949754	0.949171	0.948754	0.97006	1.029634	1.026413	1.020677	1.038343	1.039136
16	8	1.037483	1.02725	0.983059	0.948769	0.984329	0.986624	0.986914	1.00384	1.034373	1.041852	1.062443	1.071135
17	9	1.023398	0.982917	0.972818	0.936983	0.947853	0.960083	0.963117	0.974913	1.02169	1.037079	1.057028	1.032304
18	10	0.999032	0.988655	0.947186	0.941922	0.97856	1.01186	1.000514	1.007366	1.043547	1.066219	1.035126	0.994969
19	11	0.954927	0.963338	0.967367	0.961328	0.974731	0.982379	0.98468	1.020969	1.031212	1.025623	1.0148	1.008137
20	12	1.070782	1.005521	1.063352	1.08661	1.020733	1.03987	0.961471	0.962458	1.045753	1.108401	1.090388	1.115418
21	13	1.007119	0.980134	0.933463	0.910939	0.93232	0.933617	0.951162	1.011182	1.012664	1.087199	1.083786	1.049565
22	14	0.929355	0.894487	0.894185	0.907353	0.959856	1.0251	1.082571	1.101595	1.151409	1.075219	1.090186	1.055127
23	15	0.974155	0.922684	0.939338	0.930558	0.993397	1.016672	1.009524	1.044717	1.051996	1.074624	1.044221	0.985806
24	16	0.963571	0.960614	0.954651	0.920679	1.000527	0.988991	0.947265	1.083483	1.09997	1.069038	1.004942	0.988951
25	17	1.006451	0.95011	0.959755	0.926732	1.000538	0.925192						
26	SUM	16.00221	15.53967	15.47094	15.40156	15.71372	15.87481	15.88229	16.2097	16.61655	16.73159	16.67352	16.48656
27	AVERAGE	1.000138	0.97123	0.966934	0.962597	0.982108	0.992176	0.992643	1.013106	1.038534	1.045724	1.042095	1.03041
28	ST DEV	0.035496	0.031938	0.035346	0.05145	0.030371	0.037783	0.036856	0.040255	0.047178	0.030377	0.033998	0.035195
29	INDEX	1.000	0.971	0.967	0.963	0.982	0.992	0.993	1.013	1.039	1.046	1.042	1.030
30	FRAC. CONT. INDEX	0.083	0.081	0.081	0.080	0.082	0.083	0.083	0.084	0.087	0.087	0.087	0.086

Figure 15.11. Moving Average Seasonal Index.

A seasonal index can be used for forecasting when combined with a forecast of the annual value. The formula for forecasting monthly values with a seasonal index is demonstrated as follows for year i:

$$\hat{Y}_{Jan, i} = \hat{Y}_i * I_{Jan}$$

$$\hat{Y}_{Feb, i} = \hat{Y}_i * I_{Feb}$$

The Centered Moving Average Index has a smaller standard deviation than the simple average index. The reason is that the centered moving average removes some of the variability in the process. The Moving Average Index is the preferred index if the data are sufficiently long to afford the loss of observations at the beginning and end of the series.

Cycles

Cycles are year-to-year variability that exhibit a pattern which repeats itself with some regularity. With real data the cycle will not be the same length each time but may vary by a year or two each time. Variability about the cycle is caused by stochastic forces in the economy. For example, the cattle cycle is never the same length because forces such as droughts, trade policies, farm programs, imports, feed prices, domestic demand and general business cycles affect the sell-off of cattle and thus cattle prices.

Cycles can be analyzed several ways, such as using moving averages or using harmonic regression equations. In the case of forecasting cycles a harmonic regression seems to work the easiest. The harmonic regression for a cycle of length CL years is:

$$\hat{Y}_t = \hat{a} + \hat{b}_1 (2\pi T_t) + \hat{b}_2 \sin((2\pi T_t) / CL) + \hat{b}_3 \cos((2\pi T_t) / CL) + e_t$$

If this model is used the analyst must guess at the cycle length and estimate different regressions for each possible CL. The regression model associated with the best goodness of fit statistics is the best estimate of the cycle length. The second term in the regression is a trend variable $(2\pi T_t)$ so this model analyzes the data for a cycle after implicitly de-trending the data.

By taking advantage of Excel's calculation abilities and the interactive/dynamic features of Simetar, the process of testing a data series for different cycle lengths is very simple. A cell can be used to hold the CL value. Next make the harmonic variables for the regression $(\sin(2\pi T / CL)$ and $\cos(2\pi T / CL)$) a function of the cell containing the CL value.

Manually changing the CL value (say, from 7, 8, 9, 10, 11, etc.) will update the X matrix used for the regression and thus the Simetar regression results. An example of this type of estimation framework is provided in Figure 15.12 and Probabilistic Cycle Forecasts Demo.XLS.

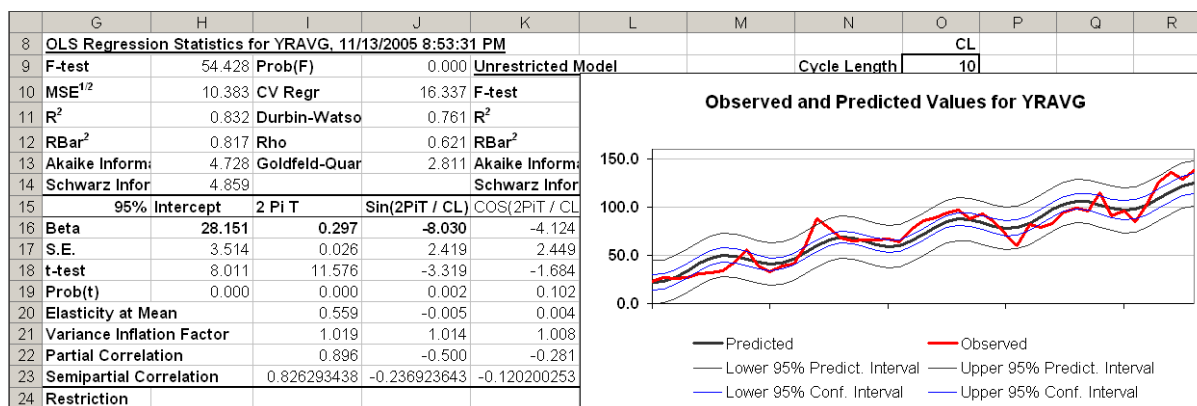


Figure 15.12 Harmonic Regression Equation to Analyze and Forecast Cycles.

The procedure described above was used to estimate the length of the cycle for the Amarillo steer prices. Cycle length, CL, values of 3 to 12 were tested. The results of testing the annual Amarillo steer price over the 1965-2001 period for a cycle are provided below:

Table 15.2 MAPE's for Testing Alternative Cycle Lengths

<u>CL</u>	<u>MAPE</u>	<u>R²</u>	<u>F-test</u>
3	16.00	0.770	36.83
4	16.04	0.759	34.69
5	15.38	0.772	37.15
6	15.76	0.768	36.32
7	15.59	0.780	38.90
8	15.37	0.767	36.21
9	14.26	0.800	44.09
10	14.12	0.832	54.42
11	14.55	0.834	55.32
12	15.99	0.819	49.82

The MAPE is minimized for a 10 year cycle and the R² is maximized at an 11 year cycle, as is the F-test statistic. These results would support the hypothesis of a 10 to 11 year cycle for Amarillo 500-600 lb. steer prices over the 37-year study period.

Forecasting the cycle is easy using a harmonic regression. The annual average price for the 38th year in the series, assuming a 10-year cycle and a trend, is forecast as:

$$\hat{Y}_{38} = 105.512 = 28.151 + 0.297 (2 \pi (38)) - 8.030 \sin(2 \pi (38/10)) - 4.124 \cos(2 \pi (38/10))$$

The probabilistic forecast with this model is

$$\tilde{Y}_{38} = \hat{Y}_{38} + \text{SDP} * \text{SND}$$

$$\tilde{Y}_{38} = 105.512 + 11.210 * \text{SND}$$

The \tilde{Y}_{38} is the stochastic forecast of the average annual price for Amarillo 500-600 lb. steers, based on cyclical and trend information in the data. The \tilde{Y}_{38} value can be used with the seasonal index for the same series to develop a probabilistic forecast of prices in December of year 38 as:

$$\tilde{Y}_{38\text{Dec}} = \tilde{Y}_{38} * (1.03176 + \text{SNP} * 0.039335)$$

where 1.03176 is the seasonal index for December and 0.039335 is the standard deviation for the seasonal index (Figure 15.11).

Moving Average Forecasts

A moving average can be used to develop a forecast of a time series. Business forecasters often use a moving average forecast as a naive forecast from which they can measure the performance of alternative forecasting methods. Moving average forecasts are based on the

simple idea, that the average of past values is an unbiased forecast of future values. A moving average systematically estimates and re-estimates the average over time by dropping the oldest value and adding the most recent value. This formula puts equal weight on each historical value and systematically adjusts for trend, cycles and seasonal variability.

For a 3-year simple moving average the forecast equations would be:

<u>Historical</u>	<u>Forecast</u>
Y_4	$\hat{Y}_4 = (Y_1 + Y_2 + Y_3) / 3$
Y_5	$\hat{Y}_5 = (Y_2 + Y_3 + Y_4) / 3$
Y_6	$\hat{Y}_6 = (Y_3 + Y_4 + Y_5) / 3$
.	.
.	.
.	.

The resulting \hat{Y} series can be compared to observed Y_t values to determine the moving averages predictive ability. A MAPE statistic should be calculated based on the residuals from the moving average or

$$\hat{e}_t = Y_t - \hat{Y}_t$$

A question remains, how long should the moving average be? Should the moving average be 3, 4, 6, 10, or 12 periods? The simple answer is best moving average length is the one with the lowest MAPE.

As the length of the moving average grows, the forecast depends less and less on the more recent values. More specifically the weight of last period's observation is 33% if the moving average is 3 periods and 20% if it is a 5 period moving average. Data series that have long memories can be forecast with longer period moving averages. The MAPE values in Table 15.3 summarize the MAPE's across alternative moving average lengths for the Amarillo monthly 500-600 lb. steer prices for 1965-2001.

Table 15.3. Comparison of Moving Average Lengths to Explain a Time Series.

<u>MA Length</u>	<u>MAPE</u>
1	3.11
2	3.81
3	4.45
4	4.98
5	5.46
6	5.83
7	6.18
8	6.48
9	6.69
10	6.91

A trade off occurs, however, as the length of the moving average decreases the future forecast values approach a constant value, namely, the most recent value. An example of a moving average forecast for the Amarillo steer price series is provided in Figure 15.13. (See Moving Average Forecasts Demo.XLS for the workbook behind this example.) The Moving Average Forecast function in Simetar allows one to change the number of periods and dynamically observe the effect on the MAPE and the forecast values. Note that with a period length of 5 the forecast values all look very stable at approximately 94. A period length of 12 results in forecast values that range from 99.3 to 97.0 seven periods out.

The residuals from the moving average calculation can be thought of as de-trended data. The moving average is a flexible trend line. It is also a function that looks like a way of measuring cycles. Overall, the moving average provides a very quick forecast a few periods out.

A probabilistic forecast of the moving average procedure is made assuming the residuals about the deterministic forecast are distributed normal with mean zero and standard deviation equal to the standard deviation of the residuals (SDR). For period 205 the stochastic forecast of the Amarillo steer prices is:

$$\tilde{Y}_{205} = \hat{Y}_{205} + \text{SDR} * \text{SND}$$

assuming a 5-month moving average as indicated in Figure 15.13.

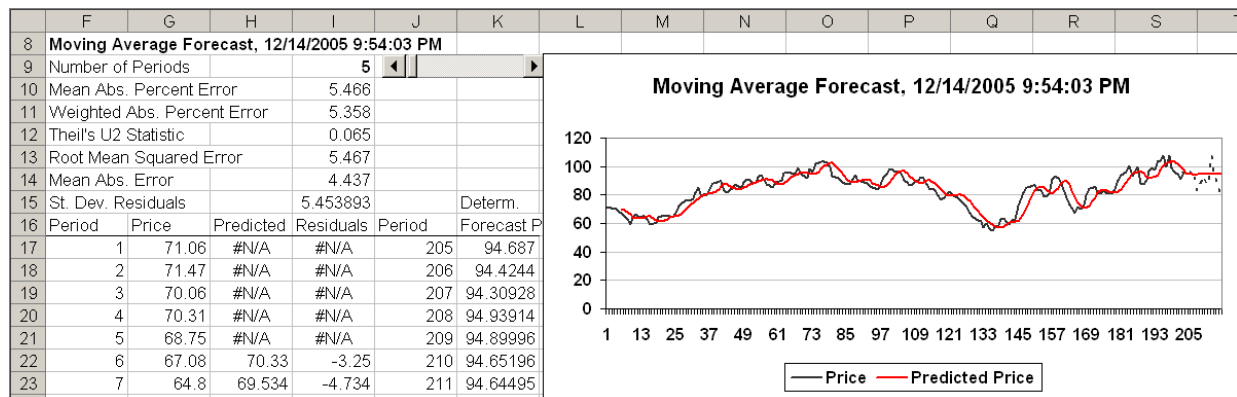


Figure 15.13. Example of a Moving Average Forecast.

Decomposition Forecasting

Decomposition forecasting is a more complex form of forecasting as it incorporates, trend seasonal, cyclical and irregular (error) into forecasting all at once. The procedure is an intuitive method for forecasting variables that show trend, seasonal, and cyclical effects. Bowerman, O'Connell, and Koehler indicate the procedure has been useful for forecasting variables that have constant, increasing or decreasing seasonal variation. The parameter estimation steps for this procedure are described completely by Bowerman, O'Connell and Koehler. The presentation of decomposition forecasting in this chapter will focus on model selection and application for this procedure.

Decomposition forecasting offers considerable flexibility as the model can be specified several different ways. The model can incorporate a multiplicative seasonal effect

$$\hat{Y}_t = TR_t * CL_t * SN_t * IR_t$$

where TR_t is the trend effect,
 CL_t is the cyclical effect,
 SN_t is the seasonal effect, and
 IR_t is the irregular error effect;

or incorporate an additive seasonal effect

$$\hat{Y}_t = TR_t + CL_t + SN_t + IR_t$$

The multiplicative form of the procedure is used when the data display increasing or decreasing seasonal patterns with respect to trend or cycle. The additive form of the model is used when the data display a constant seasonal variation with respect to trend or cycle. The cyclical component can be included or excluded in the two forms of the decomposition forecasting procedure. The cyclical term should be included if sufficient data are available to observe two or more complete cycles.

The four forms of the decomposition forecasting procedure are summarized in Table 15.4. The Simetar options for forecasting with this procedure are provided in the right most column of the table. A multiplicative decomposition model that has insufficient data (many years of monthly values) to observe several complete cycles is estimated by setting the ADDITIVE option to False and the CYCLE option to False. This model is used if the seasonal variation changes in proportion to the trend or cycle. In contrast, if the seasonal variation is constant with respect to trend/cycle and sufficient data is available to express the cycle, the Simetar options are set ADDITIVE to True and CYCLE to True.

Table 15.4. Alternative Multiplicative Decomposition Model Specifications.

<u>Model</u>	<u>Type of Data</u>	<u>Trend</u>	<u>Cycle</u>	<u>Seasonal</u>	<u>Simetar Setting</u>
Multiplicative w/o Cycle	Size of seasonal swing related to trend	$TR_t *$	1.0	$* SN_t$	ADDITIVE = False CYCLE = False
Multiplicative w/ Cycle	Size of seasonal swing related to trend	$TR_t *$	CL_t	$* SN_t$	ADDITIVE = False CYCLE = True
Additive w/o Cycle	Size of seasonal swing not related to trend	$TR_t *$	1.0_t	$+ SN_t$	ADDITIVE = True CYCLE = False
Additive w/ Cycle	Size of seasonal swing related to trend	$TR_t *$	CL_t	$+ SN_t$	ADDITIVE = True CYCLE = True

TR, CL, and SN refer to the presence of the trend, cyclical, and seasonal, respectively, components in the model.

Examples for all four of these models are presented in the four worksheets of Seasonal Decomposition Forecasts Demo.XLS. The results of the four model specifications are summarized in terms of this MAPE in Table 15.5. For the data set tested in the example, the best specification appears to be a multiplicative decomposition model with a cycle, given its MAPE of 1.204.

Table 15.5 Comparison of Four Decomposition Forecasting Models for Amarillo Price Series.

<u>Model</u>	<u>Simetar Settings</u>		<u>MAPE</u>
	<u>Additive</u>	<u>Cycle</u>	
Additive Decomp w/ Cycle	True	True	2.784
Additive Decomp w/o Cycle	True	False	2.898
Multiplicative Decomp w/ Cycle	False	True	1.204
Multiplicative Decomp w/o Cycle	False	False	1.222

See Seasonal Decomposition Forecasts Demo.XLS for the models used to estimate these results.

– **Multiplicative Decomposition**

The multiplicative decomposition model specification is written as:

$$\hat{Y}_t = TR_t * CL_t * SN_t$$

If the model assumes no cycle the model becomes:

$$\hat{Y}_t = TR_t * 1 * SN_t$$

Assuming a quarterly model the deterministic forecast is

$$\hat{Y}_1 = TR_1 * SN_1$$

$$\hat{Y}_2 = TR_2 * SN_2$$

$$\hat{Y}_3 = TR_3 * SN_3$$

$$\hat{Y}_4 = TR_4 * SN_4$$

$$\hat{Y}_5 = TR_5 * SN_1$$

⋮
⋮

The SN_j (where $j = 1, 2, 3, 4$) terms are the seasonal (quarterly) adjustment factors for the data series and they are used over and over with successively newer TR_i values. When working with monthly data there are 12 SN_j values. The TR_i values are generally assumed to be a linear trend

forecast or:

$$TR_t = \hat{a} + \hat{b} T_t$$

To make the multiplicative decomposition a probabilistic forecast assume a normal distribution for the residuals from the $\hat{Y}_t = TR_t * SN_t$ model and simulate it as:

$$\tilde{Y}_t = TR_t * SN_t * NORM(\bar{R}, SDR)$$

where SDR is the standard deviation of the residuals and \bar{R} is the mean of the residuals (IR_t).

Adding the cycle to this multiplicative decomposition model yields the following model:

$$\hat{Y}_t = TR_t * CL_t * SN_t$$

with a probabilistic forecast of:

$$\tilde{Y}_t = TR_t * CL_t * SN_j * NORM(\bar{R}, SDR)$$

In both of these multiplicative specifications it is noted that the IR_t term is ignored for forecasting. The reason being is that IR_t is assumed to be one unless it exhibits a regular pattern which can be forecasted. The IR_t should be a random error term that is simulated by the normal distribution for IR.

– Additive Decomposition

The additive decomposition model specification is written as:

$$\hat{Y}_t = TR_t + CL_t + SN_t$$

If the model assumes no cycle the model becomes:

$$\hat{Y}_t = TR_t + 0 + SN_t$$

For a quarterly model the deterministic forecast is:

$$\hat{Y}_1 = TR_1 + SN_1$$

$$\hat{Y}_2 = TR_2 + SN_2$$

$$\hat{Y}_3 = TR_3 + SN_3$$

$$\hat{Y}_4 = TR_4 + SN_4$$

$$\hat{Y}_5 = TR_5 + SN_1$$

.

.

.

The SN_j (where $j = 1, 2, 3, 4$) terms are the seasonal (quarterly) adjustments to \hat{Y} for the series. The SN_j values remain constant for all future periods. The TR_t values are forecasted by a linear trend regression model of:

$$TR_t = \hat{a} + \hat{b} T_t$$

If the data exhibits a cycle the additive decomposition model becomes:

$$\hat{Y}_t = TR_t + CL_t + SN_t$$

To make a probabilistic forecast with the additive decomposition model the IR_t term for irregular residuals (errors) are assumed to be distributed normal. With this assumption the probabilistic forecast for the additive decomposition model is:

$$\tilde{Y}_t = TR_t + CL_t + SN_t + \text{NORM}(\bar{R}, \text{SDR})$$

where \bar{R} is the mean of the residuals and SDR is the standard deviation of the residuals.

– Examples of Decomposition Models

An example of an additive seasonal decomposition model with a cycle is demonstrated in Figure 15.14. The two Simetar parameters are Additive = True and CYCLE = True produce a forecast with a MAPE of 2.784. The Data and Trend chart shows the raw data and the underlying linear trend. The Cycle and Seasonal chart shows the cycle and seasonal components of the decomposed data series. The forecasts for the TR_t , CL_t , SN_t and \hat{Y}_t are the dotted points at the end of each respective line in the charts.

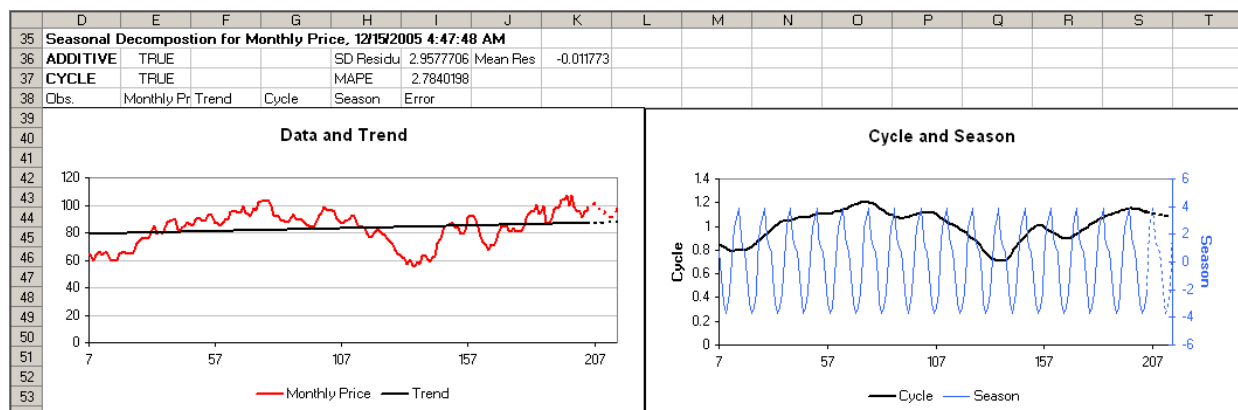


Figure 15.14. Example of an Additive Seasonal Decomposition Model With a Cycle.

The results for estimating an Additive Seasonal Decomposition model without a cycle are summarized in Figure 15.15. The Simetar parameters are ADDITIVE = True and CYCLE = False. The Cycle and Season chart shows a straight line at 1.0 for the cycle which indicates the model was estimated without the cycle (CL_t) component. The MAPE for the model is 2.898. The standard deviation of residuals (irregular component) is 2.957 and the mean of the residuals is -0.011773. This model is simulated as:

$$\tilde{Y}_t = \hat{Y}_t + \text{NORM}(\bar{R}, \text{SDR})$$

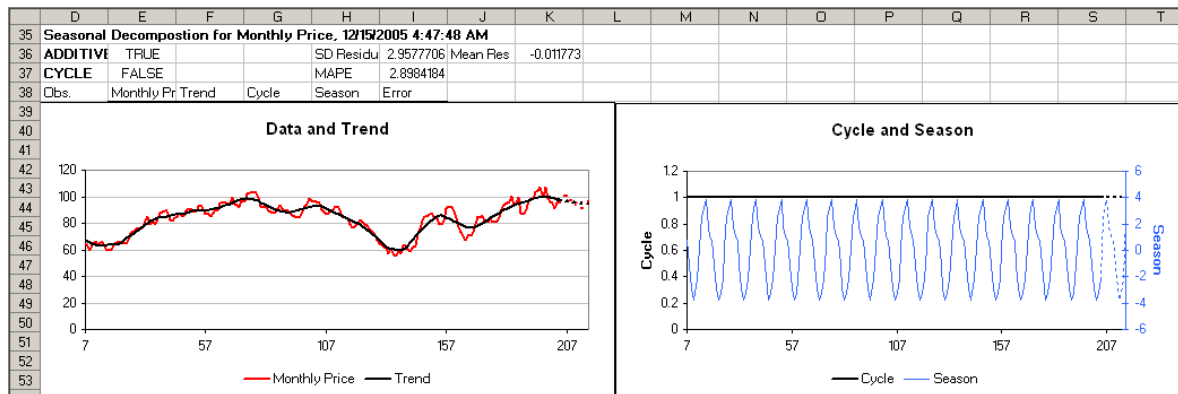


Figure 15.15. Example of an Additive Seasonal Decomposition Model Without a Cycle.

The results of a multiplicative seasonal decomposition with a cycle are summarized in Figure 15.16. The MAPE for this model is 1.204. The cycle and seasonal components are summarized in the Cycle and Season chart with their forecasted values as dots. The underlying trend and the complete forecast, \hat{Y}_t , are presented in the Data and Trend chart. The standard deviation for the residuals is 0.0367 and the mean of the residuals is 0.9988. To simulate the probabilistic forecast using this model see the example in line 241 and beyond. Because the model is multiplicative, the probabilistic forecast is simulated as:

$$\tilde{Y}_t = \hat{Y}_t * \text{NORM}(0.9988, 0.0367)$$

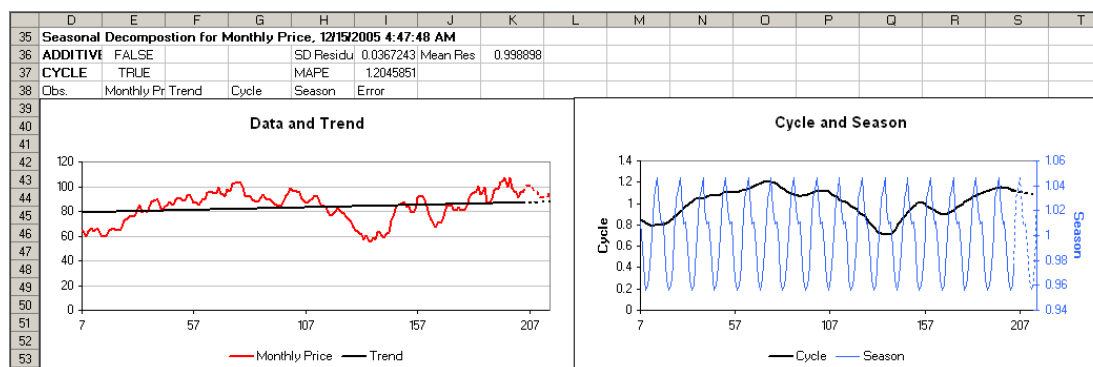


Figure 15.16. Example of a Multiplicative Seasonal Decomposition Model With a Cycle.

The fourth example model is for a multiplicative seasonal decomposition model without a cycle (Figure 15.17). The Simetar parameters are ADDITIVE = False and CYCLE = False. The MAPE is 1.222 and the standard deviation for the residuals is 0.0367 and the mean is 0.9988. The charts demonstrate that there is no cycle in the estimated model because the cycle in the Cycle and Season chart is 1.0 for all periods. The forecasted \hat{Y}_t values in the Data and Trend chart are the dotted lines for the original data series. Interestingly the trend line in the Data and Trend chart is a non-linear trend because the trend component (TR_t) is multiplied by the seasonal component (SN_t). To simulate this model for a probabilistic forecast use the formula:

$$\tilde{Y}_t = \hat{Y}_t * \text{NORM}(\bar{R}, \text{SDR})$$

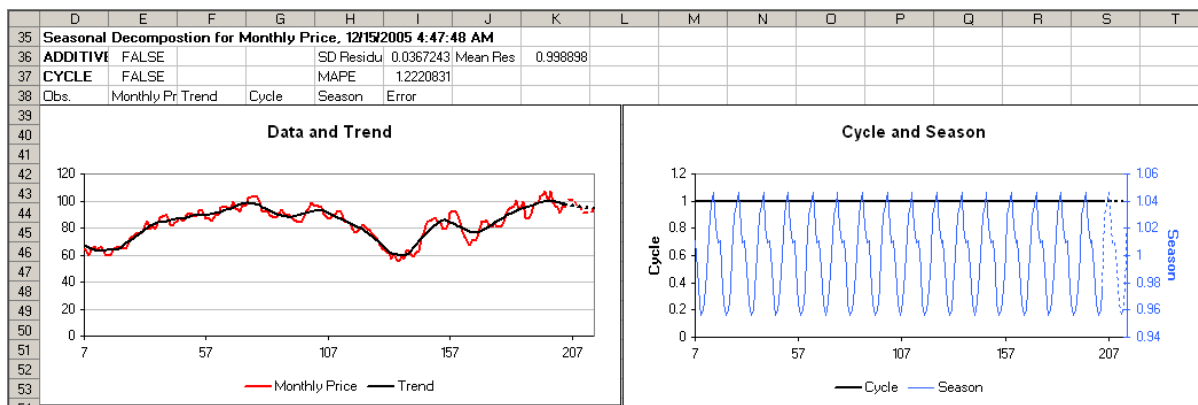


Figure 15.17. Example of a Multiplicative Seasonal Decomposition Model Without a Cycle.

Exponential Smoothing

Exponential smoothing is a very popular forecasting method with business forecasters. It provides a quick and accurate method for forecasting time series that may have trend and seasonal parameters changing over time. Several different methods of exponential smoothing are presented in this chapter. Exponential smoothing derives its name from forecasting a time series using weighted averages of past observations. The more recent observations receive higher weights. The different exponential smoothing models are:

- Simple exponential smoothing (no trend),
- Holt's exponential smoothing (for series with a changing trend present)
- Holt-Winter's additive exponential smoothing for data with constant seasonal variability, and
- Holt-Winter's multiplicative exponential smoothing for series with increasing seasonal variability.
- Dampened trend exponential smoothing alone and combined with additive or multiplicative seasonal variability.

For a very detailed description of these exponential smoothing procedures see Bowerman, O'Connell and Koehler.

– Simple Exponential Smoothing

Data series that have no trend are best forecasted with their mean as the series is by definition varying about its mean. In this case we could forecast:

$$\hat{Y}_t = b_0 + e_t$$

which is equivalent to:

$$\hat{Y}_t = \bar{Y} + e_t$$

If the mean changes over time then we can forecast it with a simple exponential smoothing model that picks up the gradual change in the mean (or level) by weighting the most recent observations most. Bower, O'Connell and Koehler suggest using a moving average of the first N observations as the starting level or l_0 .

$$l_0 = \sum_{t=1}^N Y_t / N$$

where N is 4 for quarterly data and 12 for monthly data. Assume in period T a new observation Y_t is observed and we have the previous periods estimate of the level, l_{T-1} . Using Y_t and l_{T-1} we want to estimate the level for period T or l_T . This is done using the smoothing equation:

$$l_T = \alpha Y_T + (1-\alpha) l_{T-1}$$

The α term is the level smoothing constant and must lie between 0 and 1. Thus the forecast for the next level, l_T , is a fraction of Y_T and the previous level. If α equals 0.25 then 25% of the effect of Y_T is used while 75% of the effect of l_{T-1} is used to forecast the next level. If the series is slow to change α will be small.

For applied forecasting with exponential smoothing it is assumed the best forecast of the next period, T+1, is the level forecast for T. The deterministic simple exponential smoothing forecast is:

$$\hat{Y}_{T+1} = l_T = \alpha Y_t + (1-\alpha) l_{T-1}$$

The probabilistic forecast for this methodology is simulated using the standard deviation of the residuals (SDR) for the forecast and treating the forecasted level as the mean of a normal distribution. The SDR is calculated using the residuals, which in Excel is done using the =STDEVP() function on the residuals of:

$$\hat{e}_t = Y_t - \hat{Y}_T$$

The probabilistic forecast is simulated as:

$$\tilde{Y}_{T+1} = \hat{Y}_{T+1} + \text{SDR} * \text{SND}$$

where SDR is the standard deviation of the residuals, and SND is a random standard normal deviate.

An example of simple exponential smoothing is provided in the Simple ES worksheet for the Exponential Smoothing Forecasts Demo.XLS. Simetar optimizes the α smoothing constant using Solver to minimize the MAPE. Simetar then forecasts the \hat{Y} 's for M periods, deterministically and probabilistically. Simetar's probabilistic forecast can be activated by changing the "Stochastic Forecast" option from the word False to True. See Figure 15.18 for an example.

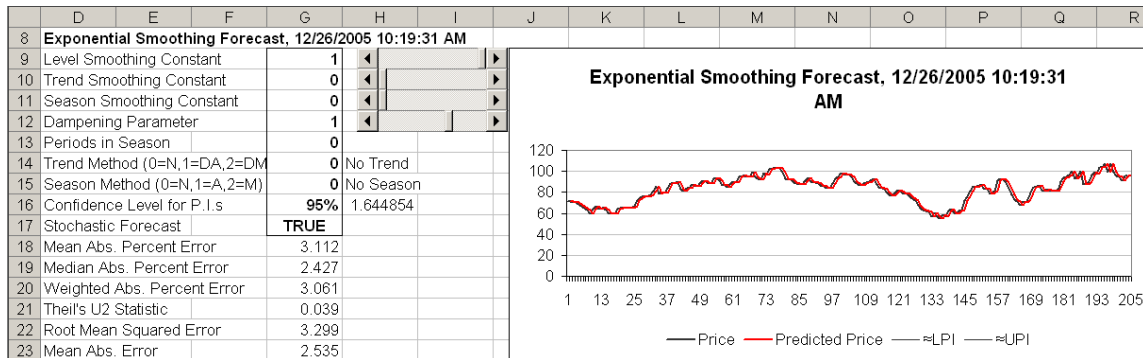


Figure 15.18. Example of Simple Exponential Smoothing.

– Holt's Exponential Smoothing

Time series with a linear trend that is increasing or decreasing at a constant rate are usually forecasted using a trend regression. This method is covered earlier in the chapter and was expressed as:

$$\hat{Y}_T = a + b T_t + e_t$$

In the trend model b is the growth rate for a one period change in T . The problem is that in the trend model, b does not change. So if a data series has a trend which changes we cannot forecast it with a constant trend.

Holt's exponential smoothing procedure is designed to forecast time series with a changing level and a changing trend or growth rate. The level at time T is estimated using the smoothing constant as:

$$l_T = \alpha Y_T + (1-\alpha)(l_{T-1} + b_{T-1})$$

where b_{T-1} is the estimated growth rate for the series in T-1. The growth rate in period T is estimated as:

$$b_T = \gamma (I_T - I_{T-1}) + (1-\gamma) b_{T-1}$$

where γ is the trend smoothing constant. The α and γ , seasonal and trend, respectively, smoothing constants are estimated using Solver to minimize MAPE when Simetar estimates the Holt exponential smoothing function.

When using Simetar to estimate and forecast using Holt's method, the user provides initial guesses for the seasonal and trend smoothing factors. Simetar forecasts, both deterministically and stochastically, for M periods if the user requests a forecast. See the example of Holt's exponential smoothing forecast in Figure 15.19.

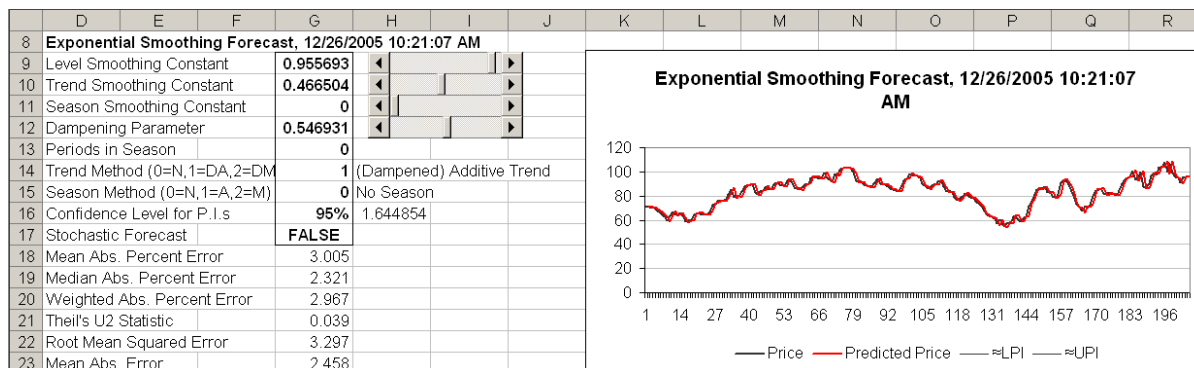


Figure 15.19. Example of Holt's Exponential Smoothing Procedure.

– Holt-Winter's Exponential Smoothing

Holt-Winter's method forecasts time series that exhibit both trend and seasonal effects. The seasonal effect can be multiplicative or additive. (See Decomposition Forecasting section of this chapter for other examples of multiplicative and additive seasonal effects.) The additive Holt-Winter's procedure is used when the seasonal effect is constant from year to year. The multiplicative Holt-Winter's procedure is used when the seasonal effect is increasing from year to year.

-- Additive Seasonal Holt-Winter's Exponential Smoothing

The presence of a linear trend and a changing level, growth rate and seasonal pattern describes a time series forecasted best by the additive Holt-Winter's method. Three equations come into play for this method:

- Level is forecasted as:

$$I_T = \alpha (Y_T - SN_{T-L}) + (1-\alpha) (I_{T-1} + b_{T-1})$$

- Growth rate is forecasted as:

$$b_T = \delta (Y_T - I_T) + (1-\delta) SN_{T-L}$$

- Seasonal change is forecasted as:

$$SN_T = \delta (Y_T - I_T) + (1 - \delta) SN_{T-L}$$

where SN_{T-L} is the seasonal factor or effect for the same season (month or quarter) the year before, and δ is the season smoothing constant. The three parameters, α , γ , and δ are estimated in Simetar by minimizing the MAPE for the additive Holt-Winter's procedure.

-- Multiplicative Seasonal Holt-Winter's Exponential Smoothing

This procedure is used to forecast time series that have a linear trend and a multiplicative seasonal pattern and the level, growth rate, and seasonal pattern can be changing. In this case the model is forecasted using three equations for level, growth rate, and seasonal effects.

$$I_T = \alpha (Y_T/SN_{T-L}) + (1-\alpha) (I_{T-1} + b_{T-1})$$

$$b_T = \gamma (I_T - I_{T-1}) + (1-\gamma) b_{T-1}$$

$$SN_T = \delta (Y_T/I_T) + (1-\delta) SN_{T-L}$$

Simetar estimates the level, trend and season smoothing constants for the multiplicative Holt-Winter's exponential smoothing procedure. Simetar also provides a deterministic and a probabilistic forecast of the time series. An example of Holt-Winter's model is presented in Figure 15.20. Note the option to specify the "Season Method" with an indication of possible settings:

- 0 is a model with no seasonal effects,
- 1 is a model with an additive seasonal effect, and
- 2 is a model with a multiplicative seasonal effect.

Changing the values for the Season Method automatically changes the nature of the model and the forecast. If you change the Season Method setting you must re-run Solver to optimize the α , γ , and δ parameters referred to, respectively, as level, trend, and season smoothing constants in Simetar.

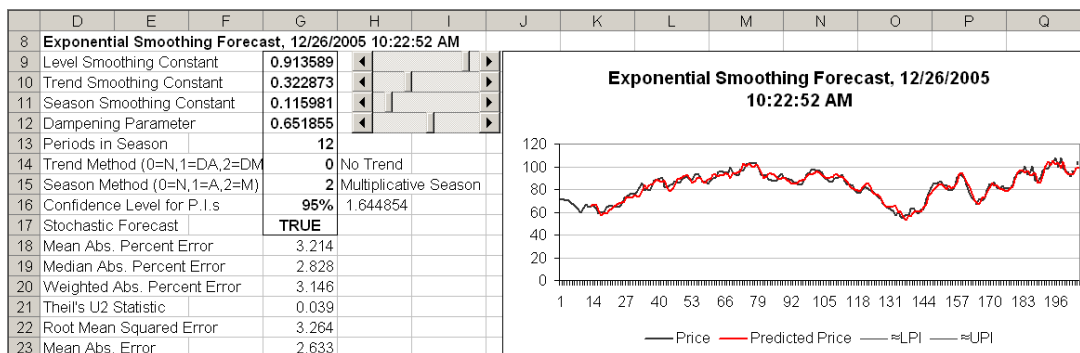


Figure 15.20. Example of Holt-Winter's Exponential Smoothing Procedure with Multiplicative Seasonality.

– Trend Dampened Exponential Smoothing

Gardner and McKenzie proposed a trend dampening method to forecast time series where the growth rate will not be sustained over a long period in the future. As the word “dampen” implies their procedure reduces the effect of the trend growth rate so the rate of decrease or increase is reduced. Both the Holt (trend) and the Holt-Winter's (trend and season) exponential smoothing procedures can be dampened using a dampening factor between 0 and 1. In the case of the dampened trend model we get:

$$I_T = \alpha Y_T + (1-\alpha) L_{T-1} + \theta b_{T-1}$$

$$b_T = \gamma(L_T - L_{T-1}) + (1-\alpha) \theta b_{T-1}$$

where θ is the dampening parameter.

Simetar automatically calculates the dampening parameter when estimating the parameters for an exponential smoothing model. The user can change the form that the dampening parameter enters the forecast, to experiment with alternative model specifications. The Holt-Winter's model example in Figure 15.20 has a dampening parameter of 0.677. The effect of the dampening parameter on the model is determined by the setting for the “Trend Method” option. The Trend method option can take on three values:

- 0 for no trend dampening,
- 1 for dampened additive trend, and
- 2 for dampened multiplicative trend.

The dampening parameter is calculated by Solver so as to minimize the MAPE for the forecast model. Therefore when you change the Trend method option you must re-run Solver to update the parameters for the model. The model in Figure 15.21 was estimated by setting the Trend Method to 0 running Solver, observing MAPE and repeating the process for Trend Method equal to 1 and 2. The Trend Method was set at 2 because this option resulted in the smallest MAPE.

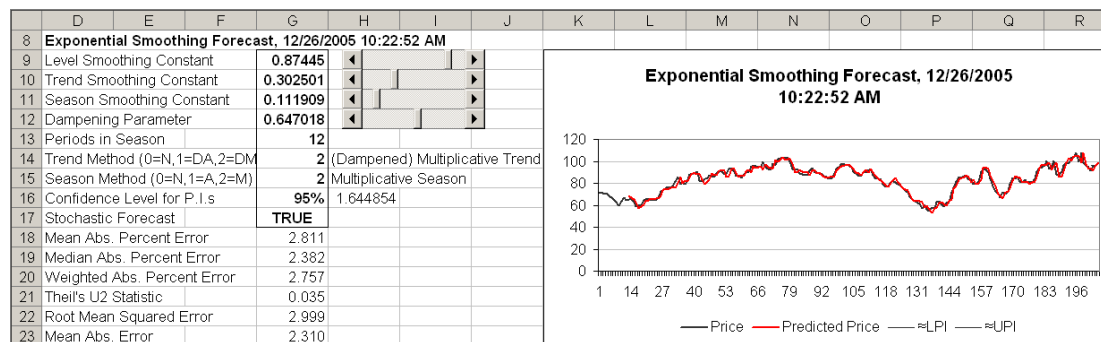


Figure 15.21. Example of Holt-Winter's Exponential Smoothing with Multiplicative Seasonality and Multiplicative Trend.

– Exponential Smoothing Summary

Exponential smoothing is a very flexible forecasting method as it can accommodate time series with trend, seasonal and level changes. With the added flexibility of multiplicative and additive trend dampening and seasonal corrections, exponential smoothing is a very powerful forecasting tool. Caution should be used in forecasting to many periods with this procedure. Exponential smoothing is best used for forecasting one period ahead. Some well behaved series can be safely forecasted for more periods ahead but be cautious in its application.

Simetar is designed to estimate and forecast all of the combinations of the exponential smoothing models. Caution should be used in estimating these models. Choose the model with the lowest MAPE, but be sure to re-estimate or update the parameters after each change of the “Trend Method” or “Season Method” options. Otherwise the parameters and the goodness of fit statistics (MAPE, etc.) will not reflect the assumptions indicated by these options.

Time Series Analysis

The most comprehensive forecasting technique is time series analysis. It has the advantage of not requiring separate forecasts for cycles, seasons or trends and it does not require forecasts for exogenous variables. Time series is based on the premise that future values are a function of past observations for the same series.

The steps for estimating a time series model are:

- Test for stationarity using the Dickie-Fuller test. If the series is not stationary then make it stationary by differencing the data.
- Once the series is stationary determine the number of lags to include in the time series model. The number of lags to include in the model can be based on the Schwarz Criterion and the autocorrelation lag function.
- Estimate the time series model based on the number of lags indicated by the tests.

These tests depend on the analysts understanding what is meant by differences and lags so an example of differences is provided in Table 15.6. For a time series, Y , of any length we can have k differences:

Table 15.6. Equations for Calculating K First Differences.

Actual Data	Differences			
Y	$D_{1,i}$	$D_{2,i}$...	$D_{k,i}$
Y_1	--	--		--
Y_2	$D_{1,2} = Y_2 - Y_1$	--		--
Y_3	$D_{1,3} = Y_3 - Y_2$	$D_{2,2} = D_{1,2} - D_{1,1}$		--
Y_4	$D_{1,4} = Y_4 - Y_3$	$D_{2,4} = D_{1,3} - D_{1,2}$		--
.	.	.		.
.	.	.		.
.	.	.		.
Y_t	$D_{1,t+1} = Y_t - Y_{t-1}$	$D_{2,t+2} = D_{1,t+2} - D_{1,t+3}$		$D_{k,t-k}$

If the Dickie-Fuller test shows that the actual data (Y_t) are not stationary, we then test the differences D_1 , D_2 , etc. until we find a stationary transformation of the actual data. Assume the second difference is stationary, then the series which is used for all further analysis is the $D_{2,i}$ series defined in Table 15.6.

The number of lags to use in estimating the time series model is based on the number of lags for the “stationary” series. (Tests are available to indicate the optimal number of lags for the stationary series.) If the raw data are stationary without differencing, then the number of lags in the model are lags of the actual data. In this case a 3 lag model is:

$$Y_t = a + b_1 Y_{t-1} + b_2 Y_{t-2} + b_3 Y_{t-3}$$

If the third difference is stationary, then the number of lags in the model are the number of lags for the $D_{3,i}$ series. In this case a 3 lag model is:

$$D_{3,t} = a + b_1 D_{3,t-1} + b_2 D_{3,t-2} + b_3 D_{3,t-3}$$

Time series models usually employ several lags of differenced data so it is important to keep track of whether the model uses actual data or differenced data and the number of lags of the series.

– Test for Stationarity

Testing for stationarity using the Dickie-Fuller test is done by solving the Dickie-Fuller test statistic for alternative numbers of differences, say 0, 1, 2, 3, The importance of a trend in the differenced series can also be tested, as well as the need for lags of the differenced series using the Augmented Dickie-Fuller test.

An example of a Dickie-Fuller Test table is provided in Figure 15.22. The top portion of the table assumes no trend (note “FALSE” in the =DF function) and the DF test statistic for the

actual data (zero lags) is -1.87 indicating the actual data are not stationary. The first difference of the data are stationary with a DF value of -12.95 . (See Time Series Forecasting Demo.XLS for the examples presented in this section.)

	F	G	H	I	J	K	L
9	Test if the data are stationary using the Dickie-Fuller Test.						
10	=DF(Range Raw Data, Trend, No. of Lags, No. Differences of Raw Data)						
11	No. of Differences		DF No Constant				
12	0		-1.87		=DF(\$B\$9:\$B\$212,FALSE,,F12)		
13	1		-12.95		=DF(\$B\$9:\$B\$212,FALSE,,F13)		
14	2		-24.97		=DF(\$B\$9:\$B\$212,FALSE,,F14)		
15	3		-33.15		=DF(\$B\$9:\$B\$212,FALSE,,F15)		
16	4		-38.75		=DF(\$B\$9:\$B\$212,FALSE,,F16)		
17							
18	No. Differences		DF With Constant				
19	0		-1.95		=DF(\$B\$9:\$B\$212,TRUE,,F19)		
20	1		-12.92		=DF(\$B\$9:\$B\$212,TRUE,,F20)		
21	2		-24.90		=DF(\$B\$9:\$B\$212,TRUE,,F21)		
22	3		-33.07		=DF(\$B\$9:\$B\$212,TRUE,,F22)		
23	4		-38.65		=DF(\$B\$9:\$B\$212,TRUE,,F23)		

Figure 15.22. Example of a Dickie-Fuller Table to Test for Stationarity.

A DF statistic less than -2.9 is considered to be significant evidence of stationarity. Using additional differences would make the DF test statistic more negative, however it further reduces the degrees of freedom of the model. The middle portion of the DF Table in Figure 15.22 repeats the DF tests but tests for the need to include a trend (note “TRUE” in the =DF function). These results are similar to the first set, so no trend is needed. The last section of Figure 15.22 shows the effects on the DF Test statistic if higher order lags are imposed on the $D_{1,i}$ differenced series. The $D_{1,i}$ series appears to become less stationary as the number of higher order lags increases, as indicated by the decreasing DF values.

– Test for Number of Lags

The number of lags for the univariate time series model can be estimated using the Schwarz Criterion. The =ARLAG() function in Simetar runs the Schwarz test on the data and indicates the number of lags for a given difference for the series. The =ARLAG() function is demonstrated in Figure 15.23. For the data series used in the example, the best number of lags for a first difference series is 1 lag based on the Schwarz test. Only a first differenced series ($D_{1,i}$) was considered in this case because the DF test pointed to one difference being stationary.

	F	G	H	I	J	K	L
34	Test for the Number of Lags based on the Schwarz Criteria.						
35	=ARLAG(Range Raw Data, Constant, No. of Differences)						
36	No. Differences		Yes Constant				
37	1		1		=ARLAG(\$B\$9:\$B\$212,TRUE,F37)		
38							
39		No Constant					
40			1		=ARLAG(\$B\$9:\$B\$212,FALSE,F37)		

Figure 15.23. Table for Testing the Number of Lags for a First Difference Series.

	F	G	H	I	J	K	L	M	N
43	Test for the Number of Lags based on Autocorrelation Coefficients.								
44	=autocorr(Range of Raw Data, No. of Lags, No. Differences)								
45	No. of Diff	No. Lags	Auto Corr	t Stat	Std. Error				
46	1	0	1				=AUTOCORR(\$B\$9:\$B\$212,G46,F46)		
47	1	1	0.0878116	1.251122	0.0701862		=AUTOCORR(\$B\$9:\$B\$212,G47,F47)		
48	1	2	0.1150222	1.626322	0.0707254		=AUTOCORR(\$B\$9:\$B\$212,G48,F48)		
49	1	3	0.0087518	0.122162	0.0716409		=AUTOCORR(\$B\$9:\$B\$212,G49,F49)		
50	1	4	-0.10255	-1.431334	0.0716462		=AUTOCORR(\$B\$9:\$B\$212,G50,F50)		
51	1	5	-0.088926	-1.228838	0.0723657		=AUTOCORR(\$B\$9:\$B\$212,G51,F51)		
52	1	6	-0.148806	-2.041179	0.072902		=AUTOCORR(\$B\$9:\$B\$212,G52,F52)		
53	1	7	-0.095777	-1.287613	0.0743832		=AUTOCORR(\$B\$9:\$B\$212,G53,F53)		
54	1	8	-0.099463	-1.326378	0.0749882		=AUTOCORR(\$B\$9:\$B\$212,G54,F54)		
55	1	9	0.0994585	1.314974	0.0756353		=AUTOCORR(\$B\$9:\$B\$212,G55,F55)		
56									
57	Test for the Number of Lags based on Partial Autocorrelation Coefficients.								
58	=pautocorr(Range of Raw Data, No. of Lags, No. Differences)								
59	No. of Diff	No. Lags	P Auto Corr	t Stat	Std. Error				
60	1	0	1				=PAUTOCORR(\$B\$9:\$B\$212,G60,F60)		
61	1	1	0.0878116	1.251122	0.0701862		=PAUTOCORR(\$B\$9:\$B\$212,G61,F61)		
62	1	2	0.1081452	1.540832	0.0701862		=PAUTOCORR(\$B\$9:\$B\$212,G62,F62)		
63	1	3	-0.009945	-0.14169	0.0701862		=PAUTOCORR(\$B\$9:\$B\$212,G63,F63)		
64	1	4	-0.117175	-1.669481	0.0701862		=PAUTOCORR(\$B\$9:\$B\$212,G64,F64)		
65	1	5	-0.07404	-1.054903	0.0701862		=PAUTOCORR(\$B\$9:\$B\$212,G65,F65)		
66	1	6	-0.11531	-1.642909	0.0701862		=PAUTOCORR(\$B\$9:\$B\$212,G66,F66)		
67	1	7	-0.060582	-0.863166	0.0701862		=PAUTOCORR(\$B\$9:\$B\$212,G67,F67)		
68	1	8	-0.073647	-1.049302	0.0701862		=PAUTOCORR(\$B\$9:\$B\$212,G68,F68)		
69	1	9	0.1191755	1.69799	0.0701862		=PAUTOCORR(\$B\$9:\$B\$212,G69,F69)		

Figure 15.24. Test for the Number of Lags Using Autocorrelation and Partial Autocorrelation Coefficients.

Two other tests can be used to determine the number of lags for a stationary series: the autocorrelation and the partial autocorrelation coefficients for the differenced series. The results of the autocorrelation and partial autocorrelation tests for the data in this example are summarized in Figure 15.24.

The results suggest that the autocorrelation coefficients decline rapidly so no more than two or three lags should be used for the model. A bar chart of the autocorrelation coefficients in Figure 15.24 makes it easier to see that the autocorrelation coefficients drop off rapidly as the number of higher order lags increases (Figure 15.25).

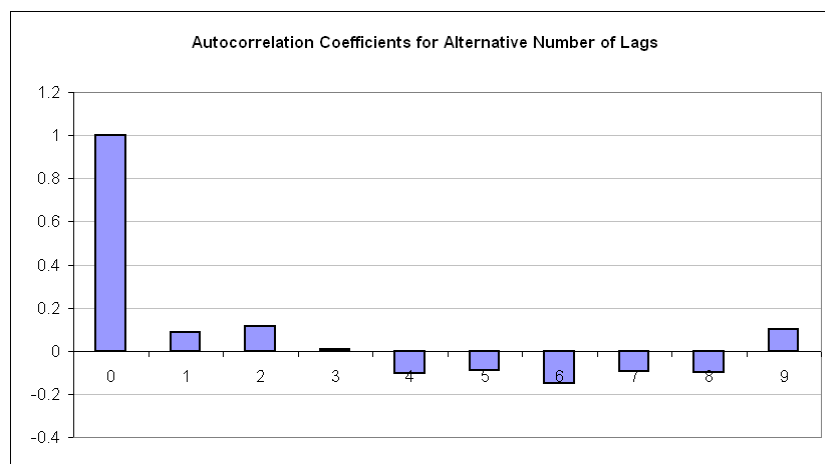


Figure 15.25. Chart of the Autocorrelation Coefficients for Alternative Lags.

– Estimated Time Series Model

Based on the tests a time series model with one difference and 6 lags was estimated for the data in this example. Six lags were included because Simetar allows the user to restrict out any lag you want. Given this option it is easy to fit the model with more lags than needed and then restrict out the undesirable lags. The Restriction Matrix also allows the analyst to experiment with alternative numbers of differences, i.e., to re-confirm what the Dickie-Fuller test indicated. Experimenting with this restriction value (try 0, 1, 2, 3) causes the model to be re-fit with a different number of differences to the original input series and will cause significant changes to the impulse response chart and the goodness of fit statistics. The time series results are summarized in Figure 15.26

The results of the model are quite good with a MAPE of 3.06% and a standard deviation on the residuals of 3.246.

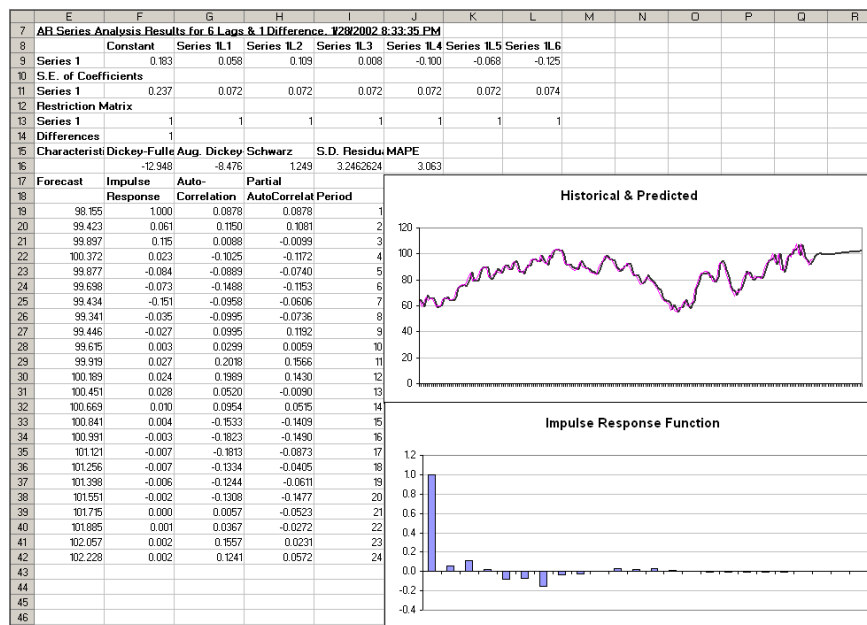


Figure 15.26 Sample Time Series Model Output.

The time series function in Simetar provides forecasts of the model if requested. In Figure 15.26 the forecast values for 24 periods are provided. The deterministic forecasts for the model could have been calculated by hand.

Probabilistic forecasts for a time series model can be developed two ways. First the deterministic forecast values can be treated as means for individual probability distributions distributed normal (\hat{Y}_{t+i} , $\hat{\sigma}$) where $\hat{\sigma}$ is the estimated standard deviation for the residuals. In other words a probabilistic forecast for period i of an auto regressive model (AR(p, q)) is:

$$\hat{Y}_{t+i} = \text{AR}(p, q)$$

$$\tilde{Y}_{t+i} = \hat{Y}_{t+i} + \hat{\sigma} * \text{SND}$$

A second method of developing a probabilistic forecast is to use the stochastic values for previous periods to update the mean in period i. Such a dynamic simulation forecast assuming the raw data (Y_t) are stationary is:

$$\hat{Y}_{t+i} = \hat{a} + \hat{b}_1 \tilde{Y}_{t+i-1} + \hat{b}_2 \tilde{Y}_{t+i-2} + \dots$$

$$\tilde{Y}_{t+i} = \hat{Y}_{t+i} + \hat{\sigma} * \text{SND}$$

If the raw data were not stationary then the forecast equations are:

$$\hat{D}_{t+i} = \hat{a} + \hat{b}_1 D_{t+i-1} + \hat{b}_2 D_{t+i-2} + \dots$$

$$\hat{Y}_{t+i} = \hat{Y}_{t+i-1} + \hat{D}_{t+i}$$

$$\tilde{Y}_{t+i} = \hat{Y}_{t+i} + \hat{\sigma} * \text{SND}$$

An example of how to simulate an AR model in a dynamic simulation mode is provided in Figure 15.27. Figure 15.27 is from the time series model in the Time Series Forecasting Demo.XLS workbook. The fan graph in the figure demonstrates that price risk increases as time progresses. The simulated results are in the Time Series Model worksheet.

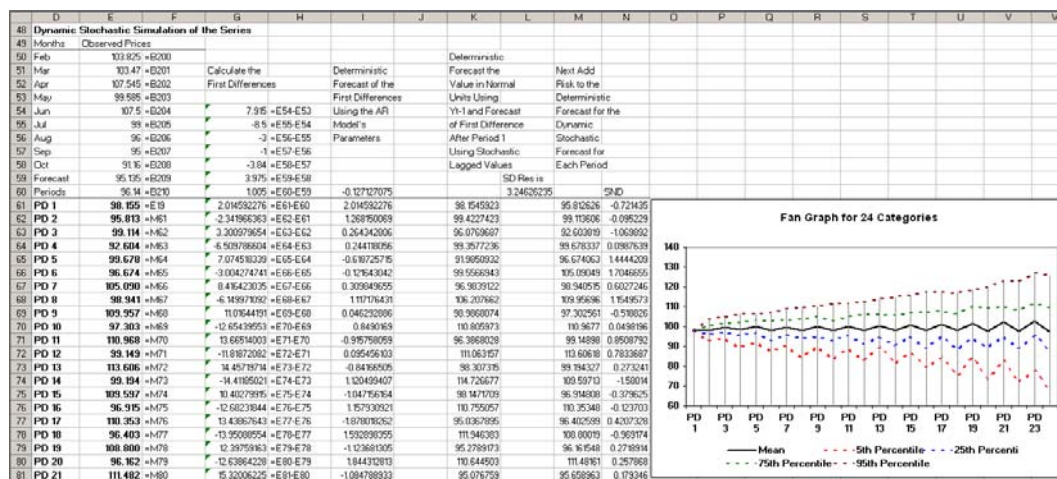


Figure 15.27. Example of a Dynamic Simulation of an AR Model.

– Presenting Probabilistic Forecasts

Stochastic simulation of a forecast creates more values than most decision makers want to see. The challenge is to present the information so it can be used for decision making. Each period's forecast is a probability distribution that has 100 or more values. Presenting the average value for each period ignores the information created by the stochastic simulation. Presenting a PDF for each period may be appropriate if the forecast is for only a few periods, otherwise the number of PDFs becomes confusing.

An alternative may be to present the stochastic results in a fan graph with the lines set to conventional confidence intervals (alpha equal 2 or 5 percent). Displaying the forecast mean and the 98 percent confidence interval about the mean indicates the downside risk and the upside risk in each year (Figure 15.28).

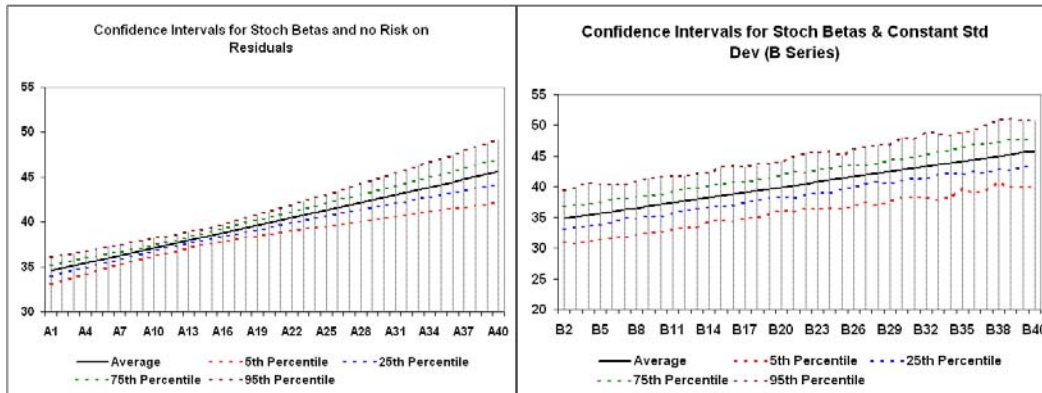


Figure 15.28. Example of a Fan Graph to Present Probabilistic Forecasts with Betas and Residuals Stochastic.

The confidence intervals in Figure 15.28 are simulated two different ways. The forecast in the left panel are simulated assuming the betas for a regression model are stochastic and there are no residuals. The stochastic betas gives the classical fanning out effect with a narrower confidence interval in the center. The procedure for simulating stochastic betas is presented in Probabilistic OLS Forecasts Demo.XLS. The formula for the simulation is:

$$\tilde{Y}_t = \tilde{a} + \tilde{b}_i X_{it}$$

where \tilde{a} and \tilde{b}_i are simulated MVN.

Note: a large sample size of 1000 or more is needed to develop smooth confidence interval lines from period to period.

The right panel in Figure 15.28 is generated from simulating the multiple regression forecast model assuming the betas are stochastic and the residuals are stochastic or

$$\tilde{Y}_t = \tilde{a} + \tilde{b}_i X_{it} + \hat{\sigma} * \text{SND}$$

The narrow middle for the confidence interval is largely lost in the right hand panel because the stochastic residual is much greater than the effect of the stochastic betas. The left panel shows the conventional confidence interval diagram shown in many statistics books. However, for probabilistic forecasting the confidence intervals in the right panel are more useful as they include both sources of risk.

The confidence intervals most often used in probabilistic forecasting come from simulating the model assuming the betas are fixed and the residuals are stochastic or:

$$\tilde{Y}_t = a + b_i X_{it} + \hat{\sigma} * \text{SND}$$

The resulting confidence intervals (Figure 15.29) are slightly narrower than for the right panel in Figure 15.28. Also, the confidence intervals will not necessarily display a widening effect over time that is presented in many statistics books.

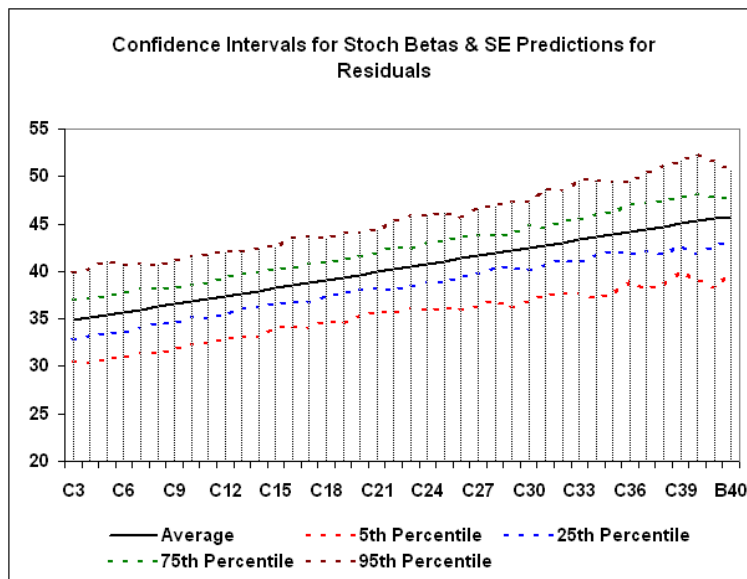


Figure 15.29. Example of a Fan Graph to Present Probabilistic Forecasts with Betas Constant and Residuals Stochastic.

In addition to presenting the forecast as a fan graph with confidence intervals for each period, it is recommended that a PDF and a CDF be presented for at least the first period (Figure 15.30). The confidence interval in the left panel PDF corresponds to the first year of the fan graph in Figure 15.29. The CDF in the right panel shows the probability (left axis) that the forecast will be less than a particular value on the bottom axis.

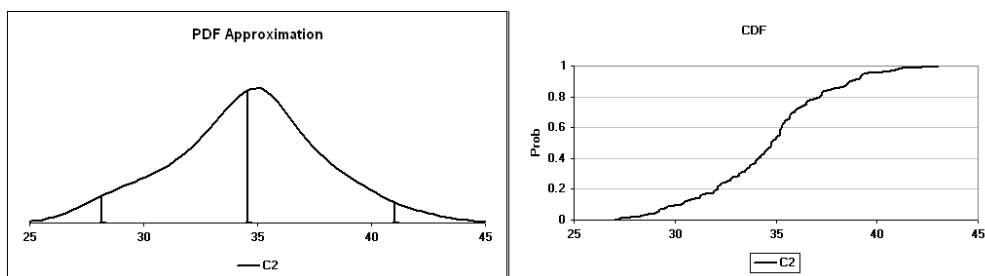


Figure 15.30. Example of Using a PDF and a CDF to Present the Probabilistic Forecast for the First Period to Demonstrate a Risk.

References

- Albright, S.C., W.L. Winston, C.J. Zappe. Managerial Statistics. Pacific Grove, CA: Duxbury 2000.
- Belsley, D.A., E. Kuh, and R.E. Welsch. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley & Sons, 1980.
- Bowerman, B.L. and R.T. O'Connell. Forecasting and Time Series: An Applied Approach. Pacific Grove, CA: Duxbury, 1993.
- Bowerman, O'Connell and Koehler. Forecasting, Time Series, and Regression: An Applied Approach. 4th Ed., United States: Duxbury: Thomson, Brooks/Cole, April 2004.
- Diebold, F.X. Elements of Forecasting. Cincinnati, OH: South-Western College Publishing, 2001.
- Schleifer, A., Jr. and D.E. Bell. Data Analysis, Regression, and Forecasting. Cambridge, MA: Course Technology, Inc., 1995.
- Ragsdale, C.T. Spreadsheet Modeling and Decision Analysis. Cincinnati, OH: South-Western College Publishing, 1998.

Table of Contents

1.0 What is Simetar?	1
1.1 Installing Simetar	1
2.0 Simulating Random Variables	3
2.1 Probability Distributions in Simetar.....	3
2.2 Simulation Engine in Simetar	5
2.3 Specifying Options in the Simulation Engine.....	6
2.3.1 Variable Names	6
2.3.2 Random Number Seed.....	6
2.3.3 Number of Iterations	6
2.3.4 Output Location.....	6
2.3.5 Scenarios	6
2.3.6 Conduct Sensitivity Elasticity Analysis	7
2.3.7 Random Number Seed.....	7
2.3.8 Conduct Sensitivity Analysis	7
2.3.9 Expected Value.....	7
2.4 User Defined Settings	7
3.0 Probability Distributions Simulated in Simetar	8
3.1 Uniform Probability Distribution.....	8
3.2 Normal Related Probability Distributions.....	8
3.2.1 Normal.....	8
3.2.2 Truncated Normal.....	9
3.2.3 Two-Piece Normal	10
3.2.4 Modified Two-Piece Normal.....	10
3.2.5 Student's t (Excel's).....	10
3.2.6 F (Excel's).....	10
3.2.7 Chi-Squared (Excel's).....	11
3.2.8 Log Normal (Excel's).....	11
3.2.9 Power Normal.....	11
3.2.10 Inverse Gaussian.....	11
3.3 Continuous Probability Distributions.....	11
3.3.1 Gamma (Excel's).....	11
3.3.2 Truncated Gamma	12
3.3.3 Exponential.....	12
3.3.4 Double Exponential.....	12
3.3.5 Weibull	12
3.3.6 Truncated Weibull.....	12
3.3.7 Cauchy.....	13
3.3.8 Logistic.....	13
3.3.9 Log-Log.....	13
3.3.10 Log-Logistic	13
3.3.11 Extreme Value.....	13
3.3.12 Pareto.....	14
3.4 Finite-Range Continuous Probability Distributions.....	14
3.4.1 Triangle	14
3.4.2 Beta (Excel's).....	14
3.4.3 PERT	14
3.4.4 Cosine.....	15
3.4.5 Semicircle.....	15

3.5	Analogs to Finite Range Probability Distributions	15
3.5.1	GRK	15
3.5.2	GRKS	15
3.6	Discrete Probability Distributions	16
3.6.1	Bernoulli	16
3.6.2	Binomial	16
3.6.3	Negative Binomial	17
3.6.4	Multinomial	17
3.6.5	Poisson	17
3.6.6	Geometric	17
3.6.7	Hypergeometric	17
3.7	Sample Based Probability Distributions	18
3.7.1	Empirical	18
3.7.2	Truncated Empirical	18
3.7.3	Discrete Empirical	19
3.7.4	Kernel Density Estimated Random Variable	19
3.7.5	Discrete Uniform	20
3.7.6	Random Sorting	20
3.7.7	Bootstrapping (Random Sample with Replacement)	20
3.8	Time Series Probability Distributions	21
3.1	Random Walk	21
3.9	Multivariate Distributions	21
3.9.1	Correlated Standard Normal Deviates	21
3.9.2	Correlated Uniform Standard Deviates	22
3.9.3	Multivariate Normal (MVN) Distribution in One Step	23
3.9.4	Multivariate Normal (MVN) Distribution in Two Steps	23
3.9.5	Multivariate Empirical (MVE) Distribution in One Step	24
3.9.6	Multivariate Empirical (MVE) Distribution in Two Steps	25
3.9.7	Multivariate Mixed Distribution	25
3.9.8	Multivariate Log Normal	26
3.9.9	Multivariate Student's t	26
3.9.10	Hotelling T-Squared	27
3.9.11	Wishart	27
3.9.12	Wilk's Lambda	27
3.9.13	Dirichlet	27
3.9.14	Uncorrelating Random Deviates (USD and SND)	28
3.10	Iteration Counter	28
4.0	Parameter Estimation for Parametric Probability Distributions	29
4.1	Parametric Probability Distributions	29
4.2	Empirical Probability Distributions	30
4.3	Multivariate Probability Distributions	30
4.4	GRKS Probability Distributions	32
5.0	Statistical Tests for Model Validation	32
5.1	Univariate Distribution Tests	32
5.2	Multivariate Distribution Tests	33
5.3	Test Correlation	34
5.4	Test Mean and Standard Deviation	35
5.5	Univariate Tests for Normality	35
5.6	Multivariate Tests for Normality	36
5.7	Compare Means (ANOVA)	36

5.8 Compare Two Cumulative Distribution Functions (CDFs)	37
6.0 Graphical Tools for Analyzing Simulation Results	38
6.1 Line Graph	38
6.2 CDF Graph.....	39
6.3 PDF Graph	40
6.4 Histograms	41
6.5 Fan Graph.....	41
6.6 StopLight Graph.....	42
6.7 Probability Plots.....	43
6.8 Box Plots.....	43
6.9 Scatter Matrix Graph.....	44
7.0 Scenario Analysis	45
8.0 Sensitivity Analysis	46
9.0 Sensitivity Elasticity Analysis	47
10.0 Simulation and Optimization	48
11.0 Numerical Methods for Ranking Risky Alternatives	49
11.1 Stochastic Dominance (SD)	49
11.1.1 First Degree Stochastic Dominance	49
11.1.2 Second Degree Stochastic Dominance	49
11.1.3 Generalized Stochastic Dominance with Respect to a Function (SDRF).....	49
11.2 Stochastic Efficiency with Respect to a Function (SERF).....	50
11.3 Risk Premiums	51
11.4 Target Probabilities for Ranking Risky Alternatives	51
11.5 Target Quantiles for Ranking Risky Alternatives	52
12.0 Tools for Data Analysis and Manipulation	52
12.1 Matrix Operations	52
12.1.1 Column Vector to a Matrix	53
12.1.2 Reverse a Column or Row of Values	53
12.1.3 Convert a Matrix to a Vector.....	53
12.1.4 Sort a Matrix.....	53
12.1.5 Factor a Square Matrix	53
12.1.6 Transpose a Matrix (Excel)	54
12.1.7 Generalized Inverse of a Rectangular Matrix.....	54
12.1.8 Invert a Nonsingular Matrix (Excel)	54
12.1.9 Multiply Two Matrices.....	54
12.1.10 Concatenate Two Matrices	54
12.1.11 Convert a Vector to a Diagonal Matrix	54
12.1.12 Find the Determinate of a Square Matrix	54
12.2 Data Manipulation.....	55
12.2.1 Create an Identity Matrix	55
12.2.2 Create a Sequence of Numbers.....	55
12.2.3 Create a Matrix of Ones	55
12.2.4 Create a Centering Matrix	55
12.2.5 Create an Equicorrelation Matrix	55
12.2.6 Create a Topelitz	55
12.3 Box-Cox Transformation	56

12.4 Workbook Documentation	56
12.4.1 Delete Numbers in a Cell	56
12.4.2 Delete Text in a Cell.....	57
12.4.3 View Cell Formulas.....	57
12.4.4 View All Formulas	57
12.4.5 Workbook and Worksheet Name	57
13.0 Regression Analysis	57
13.1 Simple Regression.....	58
13.2 Multiple Regression	58
13.3 Bivariate Response Regression	61
13.3.1 Probit Analysis	61
13.3.2 Logit Analysis	61
14.0 Cyclical Analysis and Exponential Forecasting	62
14.1 Seasonal Index	62
14.2 Seasonal Decomposition Forecasting.....	63
14.3 Moving Average Forecast	64
14.4 Exponential Smoothing Forecast	64
14.5 Measuring Forecast Errors	66
15.0 Time Series Analysis and Forecasting.....	67
15.1 Tests for Stationarity	67
15.2 Number of Lags.....	68
15.3 Sample Autocorrelation Coefficients	69
15.4 Maximum Likelihood Ratio Test	70
15.5 Estimating and Forecasting Autoregressive (AR) Models.....	70
15.6 Estimating and Forecasting Vector Autoregressive (VAR) Models	72
16.0 Other Statistical and Data Analysis Functions.....	74
16.1 Summary Statistics.....	74
16.2 Jackknife Estimator	74
17.0 Function Evaluation.....	75
17.1 Optimize a Function	75
17.2 Value of a Function.....	76
17.3 Integral of a Function	76
18.0 Getting Help with Simetar	77
19.0 Solutions to Problems in Simetar Application	78
20.0 List of All Simetar Functions.....	82
21.0 Cross Reference of Functions and Demonstration Programs.....	86

Simetar© 2008

Simulation & Econometrics To Analyze Risk

1.0 What is Simetar?

Simetar© 2008 is a simulation language written for risk analysts to provide a transparent method for analyzing data, simulating the effects of risk, and presenting results in the user friendly environment of Microsoft® Excel¹. Any Excel spreadsheet model can be made stochastic and simulated using Simetar functions. Simetar, an acronym for Simulation for Excel to Analyze Risk is an Excel add-in. Simetar requires little additional memory and operates efficiently on most PCs running Excel XP, Excel 2000, Excel 2003, and Excel 2007. Instructions for installing Simetar are provided in Section 1.1.

Simetar consists of Menu Driven and User Defined Functions for Excel. A common principle in Simetar, is that all functions are dynamic; so if changes are made to the original data most all parameters, hypothesis tests, regression models, and risk ranking strategies are automatically updated. This feature of having Excel dynamically recalculate parameters offers significant efficiencies during the development, validation, verification, and application of stochastic simulation models.

The more than 230 functions in Simetar can be categorized into six groups: (a) simulating random variables, (b) parameter estimation and statistical analyses, (c) graphical analysis, (d) ranking risky alternatives, (e) data manipulation and analysis, (f) multiple regression, and (g) probabilistic forecasting. Simetar can be used to perform all of the steps for developing, simulating, and applying a stochastic model in Excel, namely: estimate parameters for random variables, simulate stochastic variables, test the validity of the random variables, present the results graphically, and rank risky alternatives.

The next section describes the procedure for installing Simetar. After installing Simetar open the demonstration program to see learn how to apply the major functions in Simetar. More than 100 demonstration programs will be installed on your computer at **Start > Programs > Simetar > Demos**. Refer to these demonstration programs as you read the User's Manual to learn how the functions are applied in working simulation and forecasting models.

1.1 Installing Simetar

The first step in installation is to set the macro security level for Excel to low. (If you currently have Simetar installed be sure to uninstall Simetar and delete the C:\Program Files\Simetar folder.) After setting macro security to Low, close Excel and insert the Simetar CD in your computer's CD drive. (If you are installing from a file downloaded from the Simetar website, copy the file to your computer's hard drive and proceed with the installation.) From the Windows Explorer, double click on the Simetar.exe file name and the Setup Wizard will open to

1 . Simetar© is copyrighted by the authors. Microsoft, Excel, and Windows are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries.

Figure 1. Click the Next box to proceed with the installation. The License Agreement is provided in the second screen of the Wizard (Figure 2). Read the License Agreement and click on the I Agree box to proceed with installation.



Figure 1. Install Simetar.

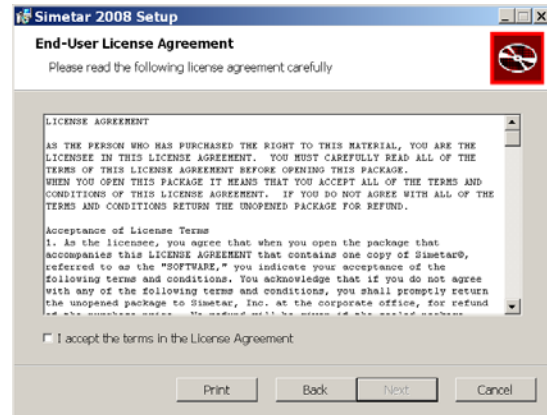


Figure 2. License Agreement.

Enter the License Code provided on your CD or with the Download Instructions (Figure 3). If you did not uninstall Simetar, a screen will appear that allows you to uninstall using our uninstaller. In the next screen select the “Typical” installation. Figure 4 is provided so you can change your mind as to the type of installation.

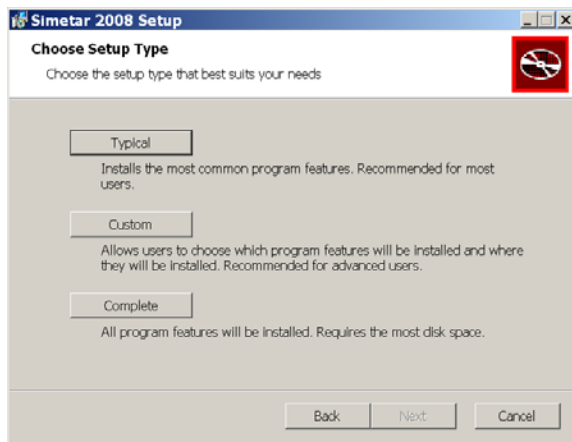


Figure 3. Type of Installation.

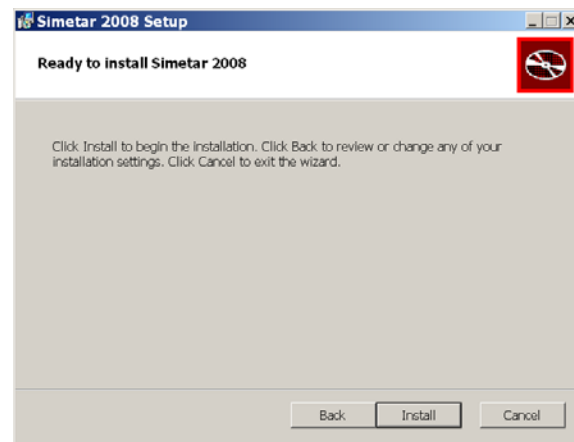


Figure 4. Final Chance to Change Your Installation Type.

Enter your license code in Figure 5, make sure all letters are caps and the dashes are included. The installation will take 2-3 minutes as the files are transferred and the appropriate files are updated so Simetar can operate in the Microsoft environment. The program will be stored in C:\Program Files\Simetar 2008. The last screen (Figure 6) indicates that Simetar has finished installing properly. Open Excel and you will see the Simetar toolbar in Excel 2003. For Excel 2007 you must click Add-Ins and then click on the word Simetar to see the Toolbar presented below. To test Simetar type the following command in cell A1 =NORM() press Enter and then press F9. You will see random draws of a standard normal random variable.

The installation procedure will place the word “Simetar” on the toolbar and add the Simetar icon

toolbar below:



Figure 5. Enter Your License Code.

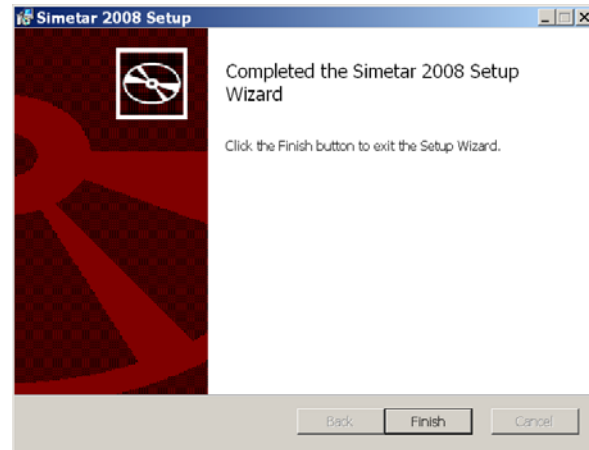


Figure 6. Final Installation Screen.



2.0 Simulating Random Variables

Simulating a stochastic model in Excel is accomplished by generating random values for each of the random variables, letting Excel update the model's equations, and saving the results of key output variables (KOVs) for statistical analysis and presentation. Repeating this process a large number of times (iterations or trials) causes the model to be simulated for a wide range of possible combinations of the random variables. The resulting array of 100 or more simulated values for a KOV defines an empirical probability distribution for each of the output variables. Probability distributions for the output variables are analyzed to gain a better understanding of the risk for the system being modeled. An example of simulation with Simetar is provided in example program Simulation Demo.xls.

2.1 Probability Distributions in Simetar

Simetar includes functions for generating pseudo-random numbers from more than 50 probability distributions plus six distributions included in Excel. An alphabetical list of probability distributions simulated by Simetar is provided in page 4. A detailed description of each Simetar function for simulating random numbers is provided in Section 3. See the Probability Distributions Demo.xls workbook for examples of how the functions are used in Excel. Access the Simetar demonstration programs from the Start menu:


Start > Programs > Simetar > Demos

Distribution Function Name and Parameters for each Probability Distribution in Simetar

Bernoulli	=BERNOULLIDIST(ProbabilityofTrueOutcome)
Binomial	=BINOMINV(n,Prob,[USD])
Bootstrap	=BOOTSTRAPPER(ListofPossibleOutcomes,RecalculationOff)
Cauchy	=CAUCHY(Median,Sigma,[USD])
Cosine	=COSINV(Center,Radius,[USD],MaxIterations,Precision)
Correlated SND	=CSND(RangeCorrelationMatrix, [ISNDs])
Correlated USD	=CUSD(RangeCorrelationMatrix,[ISNDs],[MatrixRow],[RankCorr])
Discrete Empirical	=DEMPIRICAL(Values,[USD],[Probabilities of Values])
Double Exponential	=DEXPONINV(Mu,Sigma,[USD])
Dirichlet	=DIRICHINV(Alphas,[USD],[MatrixRow])
Empirical	=EMP(Values,Probabilities,[USD],[NormTails])
Empirical	=EMPIRICAL(Values,Probabilities,[USD],[NormTails])
Exponential	=EXPONINV(Beta,[USD])
Extreme value	=EXTVALINV(Mu,Sigma,[USD])
Geometric	=GEOMINV(Prob,[USD])
GRK	=GRK(MinValue,MidPoint,MaxValue,[USD])
GRKS	=GRKS(MinValue,MidPoint,MaxValue,[USD],[LowerSD],[UpperSD])
Hotelling T Sq	=HOTELLTINV(P,Degrees_Freedom,[UniformRandomNumber])
Hypergeometric	=HYPERGEOMINV(n,N1,S1,[USD])
Inverse Gaussian	=INVGAUS(Mu,Sigma,[USD],[MaxIterations],[Precision])
Kernal Density	=KDEINV(DataRange,BandWidth,KernelEstimator,[USD],[MaxIter],[Prec])
Logistic	=LOGISTICINV(Mu,Sigma,[USD])
Log-Log	=LOGLOGINV(Mu,Sigma,[USD])
Log-Logistic	=LOGLOGISTICINV(Alpha,Beta,[USD])
Modified 2 Piece Normal	=MTPNORM(MinValue,MidPoint,MaxValue,[USD],[LowSD],[UpSD])
Multinomial	=MULTINOMINV(NumTrials,Probs,[USDs])
Multivariate Empirical	=MVEMPIRICAL(RandomValuesDataMatrix,[SND],[MatrixRow])
Multivariate Log Normal	=MVLOGNORM(MeanVector,CovMatrix,[SNDs],[MatrixRow],[Moments])
Multivariate Normal	=MVNORM(MeansVector,CovarianceMatrix,[SNDs],[MatrixRow])
Multivariate Students t	=MVTINV(Student t,CovarianceMatrix,[DegreesFreedom],[SNDs],[MatrixRow])
Negative Binomial	=NEGBINOMINV(k,Prob,[USD])
Normal	=NORM(Mean,StandardDeviation,[USD])
Pareto	=PARETO(Alpha,Beta,[UniformRandomNumber])
PERT	=PERTINV(A,B,C,[USD])
Power Normal	=PNORM(Mean,StandardDeviation,P,[USD])
Poisson	=POISSONINV(Lambda,[USD])
Random Sorting	=RANDSORT(InputRangeLocation,[RecalculationOff],[DataHorizontal])
RandomWalk	=RANDWALK(Mean,StandDev,USD,Distribution,InitialVal,Coefficient)
Semicircle	=SEMICIRCDIST(X,Center,Radius,[Cumulative or Density])
Truncated Empirical	=TEMPIRICAL(RandomValues,Probabilities,MinVal,MaxVal,[USD])
Truncated Gamma	=TGAMMAINV(Alpha,Beta,AbsoluteMin,AbsoluteMax,[USD])
Truncated Normal	=TNORM(Mean,StanDev,[Min],[Max],[USD],[StackTails])
Truncated Weibull	=TWEIBINV(Alpha,Beta,[Min],[Max],[USD])
2 Piece Normal	=TPNORM(Mean,StandardDeviation1,StandardDeviation2,[USD])
Triangle	=TRIANGLE(A,B,C,[USD])
Uniform	=UNIFORM(LowerValue,UpperValue,[USD])
Uncorrelated SNDs	=USND(CorrelationMatrixRange,CorrelatedNormalDeviatesRange)
Uncorrelated USDs	=UUSD(CorrelationMatrixRange,CorrelatedUniformDeviatesRange)
Weibull	=WEIBINV(Alpha,Beta,[USD])
Wilk's Lambda	=WILKSLINV(P,FirstDegrees of Freedom,SecondDegrees of Freedom)
Wishart	=WISHINV(CovarianceMatrix,Degrees of Freedom)

Native Excel probability distributions can be simulated in Simetar

Beta	=BETAINV(Uniform(),Alpha,Beta,Minimum,Maximum)
Chi-Squared	=CHIINV(Uniform(),Degrees of Freedom)
Gamma	=GAMMAINV(Uniform(),Alpha,Beta)
Log Normal	=LOGINV(Uniform(),Mean,StandardDeviation)
Students t	=TINV(Uniform(),Degrees of Freedom)
F	=FINV(Uniform(),Degrees of Freedom1, Degrees of Freedom2)

Simetar allows the user to specify the type of sampling procedure and the random number generator to use in generating random values. Three different random number generators are available: Mercene Twister, the Multiplicative Random Number Generator, or Excel's native generator. Two different random number sampling procedures are available : Latin hypercube and Monte Carlo. These random number generators are pseudo random and thus are suitable for conducting scenario and sensitivity analyses. The user can select the random number generator and the sampling method by selecting the General Settings Options  icon and choosing the desired options in the Default User Settings menu (Figure 7).

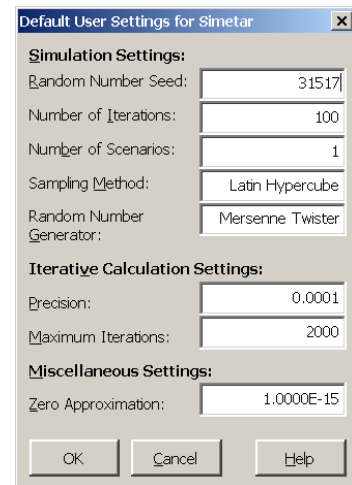



Figure 7. Setup Menu for User's Settings.

2.2 Simulation Engine in Simetar

The dialog box in Figure 8 for simulating a stochastic Excel simulation model is accessed by the  icon on the Simetar toolbar. Options specified in the dialog box are saved by selecting the *Save* or *SIMULATE* buttons.

The user must specify one or more *Output Variables* (KOVs) for the statistical analysis of simulated results. The summary statistics and each simulated value (in iteration order) for each KOV are saved in the *SimData* worksheet. An output KOV can be any cell in the spreadsheet. KOVs can be cells that contain random variables, intermediate calculations, and final answers.

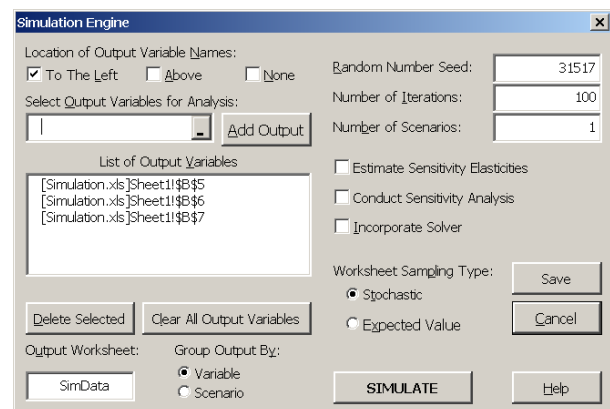


Figure 8. Simulation Dialog Box for Simetar.

Add variables to the *List of Output Variables* box by clicking in the *Select Output Variables for Analysis* window, highlighting the spreadsheet cell or cells to include, and clicking the *Add Output* box. Indicate where the variable's label is located, as in the cell *To The Left*, in the cell *Above*, or *None*. Several hundred output variables can be handled by Simetar. The sample menu in Figure 8 shows that the variables in B5, B6, and B7 are the output variables and their labels are *To The Left*. To delete an output variable or several variables, highlight the variables in the dialog box and click the *Delete Selected* button. Clicking on the *Clear All Output Variables* button will delete all of the output variables listed in the dialog box. Clicking on an output variable in the *List of Output Variables* box causes Excel to highlight the particular variable in the workbook. Simetar updates the location of KOVs in the Output Variable table if the spreadsheet is modified by adding rows or columns. Information in the Simulation Engine must be re-entered each time the workbook is opened.

After specifying the output variable(s) click the *SIMULATE* button and Simetar will simulate the workbook and save the simulated values for the output variables in the *SimData* worksheet or in the worksheet specified by the user. The statistics for each output variable are provided in rows 3-7 of *SimData* and the simulated values for each variable, by iteration, start in row 9 (Figure 9). After the 100 or more simulated values there are 10 rows of pre-programmed equations to calculate the probability of the output variable being less than a specified target. Type in a target

value in a row labeled “ x_i - value” and the probability of the KOV being less than or equal to the value will appear in the next row labeled “Prob ($X \leq x_i$).” For example, there is a 38.0 percent chance that receipts will be less than \$1,300 (see column D of Figure 9).

The simulated variables in the SimData worksheet always appear in the order they were added to the *List of Output Variables* (Figures 8 and 9). The rows of simulated values for the output variables correspond to the actual iterations as they were simulated, i.e., the iteration order is maintained across output variables in SimData. The simulated values of each iteration for all output variables are provided so the user can analyze the results using Simetar functions, (for hypothesis tests, charts for presenting simulation results, and ranking risky alternatives.

	A	B	C	D	E
1	Simetar Simulation Results for 100 Iterations © 2005.				
2	Variable	Sheet1!B4	Sheet1!B5	Sheet1!B6	
3	Mean	10.00694	149.9765	1508.8	
4	StDev	3.027181	29.02188	561.451	
5	CV	30.25081	19.35096	37.21177	
6	Min	2.434353	100.4193	360.1179	
7	Max	18.65196	199.1456	2871.398	
8	Iteration	Price	Production	Receipts	
9	1	10.08475	198.7316	2004.159	
10	2	8.584123	151.1181	1297.217	
11	3	10.6525	138.8259	1489.495	
12	
13	99	4.635541	196.4296	910.5577	
14	100	8.88908	195.6557	1739.199	

Figure 9. Example of Stochastic Results in the SimData

2.3 Specifying Options in the Simulation Engine

2.3.1 Variable Names. The user must specify the name for a KOV before it is added to the *List of Output Variables* box (Figure 8). The variable name will appear with the stochastic results in the SimData worksheet (Figure 9). There are three options for specifying the variable names. The first option is to use the text in the cell to the left of the KOV. The second option is to use the text in the cell above the KOV and the third option is to not specify a name for the KOV. The variable name can be a concatenation of the text in the cells to the left and above the KOV cell (Figure 10). The user must specify the location of the label before adding the variable to the List of Output Variables table.

	A	B	C	D
4	Labels	KOVs		
5	Price	8.46	=NORM(10,3)	
6	Production	101.50	=UNIFORM(100,200)	
7	Receipts	858.79	=B5*B6	

Figure 10. Labels for Key Output Variables in Cells to the Left.

2.3.2 Random Number Seed. The user may specify the *Random Number Seed*, in place of the default seed, 31517, to insure the same starting point for the pseudo random number generator from one run to the next (Figure 8). The default seed can be changed permanently in the Default User Settings menu (Figure 7).

2.3.3 Number of Iterations. The *Number of Iterations* to simulate the spreadsheet model can be set by the user (Figure 8). The default number of iterations can be changed in the Default Users Settings menu (Figure 7).

2.3.4 Output Worksheet. Output results for a simulation are stored in the SimData worksheet of the current (or a new) workbook using the specified *Output Location* (Figure 8).

2.3.5 Scenarios. The *Number of Scenarios* defaults to 1 in the menu box (Figure 8). If your model uses the =SCENARIO() function to simulate multiple scenarios, enter the number of scenarios. See Section 7.0 to learn more about the Scenario feature.

2.3.6 Conduct Sensitivity Elasticities Analysis. This option causes Simetar to simulate the spreadsheet model once for the base situation and once for each variable listed in the Sensitivity

Variable Input window (Figure 11). The elasticity is defined as the percentage change of the KOV to a one percent change in an exogenous variable. The larger the elasticity, the greater the sensitivity of the KOV to the exogenous variable. See Section 9.0 to learn more about this option.

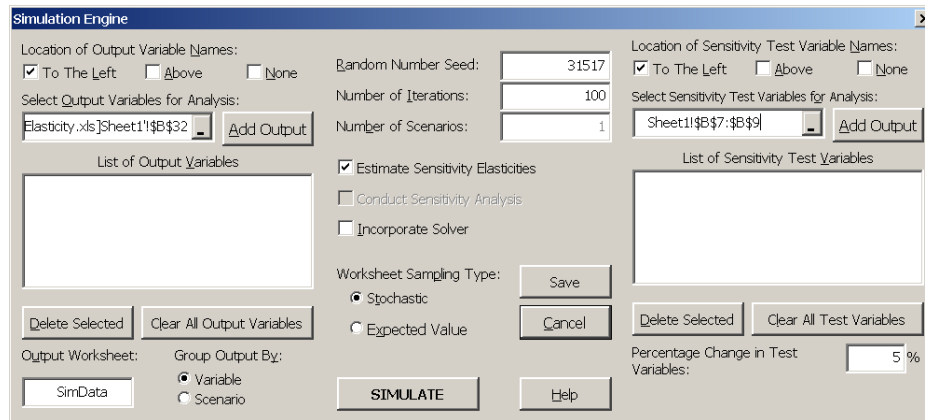




Figure 11. Simulation Menu Expanded to Estimate Sensitivity Elasticities for Variables in B6 – B8.

2.3.7 Conduct Sensitivity Analysis. Any Excel spreadsheet model can be simulated using the sensitivity analysis option. Numerous KOVs can be tested for percentage changes in one exogenous variable. Three percentage change levels, say, $\pm 5\%$, $\pm 10\%$, and $\pm 20\%$, can be specified by the user. See Section 5.0 to learn more about this option. Simetar also performs sensitivities by selecting the *Conduct Sensitivity Analysis* option. (See Section 8.0 for details on simulating sensitivity analyses.)

2.3.8 Incorporate Solver. Simetar can stochastically simulate a simultaneous equation or linear programming model by selecting *Incorporate Solver*. (See Section 10.0 for details on simulating with an optimizer.)

2.3.9 Expected Value. Once stochastic variables have been incorporated into an Excel simulation model, all of the values (cells) update every time the sheet calculates or F9 is pressed. This feature in Excel is very useful for testing if stochastic variables are working correctly and if they have been linked to the proper equations in the model. However, it is also very useful to have the stochastic values fixed at their means for equation verification. Clicking the Expected Value  icon sets all random variables to their means and un-clicking the icon causes Excel to calculate values for the stochastic variables. During simulation Simetar overrides the Expected Value button's setting and simulates stochastic values for all of the random variables.

2.4 User Defined Settings

The user may specify his/her preferred settings for the type of random number generator, sampling method, number of iterations, number of scenarios, random number seed, precision for MLE parameter estimation, and the maximum number of iterations for MLE and other iterative solution functions. The user defined settings are specified in the dialog box associated with the  icon (Figure 7).

3.0 Probability Distributions Simulated in Simetar

Simetar is capable of simulating univariate and multivariate random numbers from more than 50 probability distributions. Each probability distribution is described in detail in this section. Univariate probability distributions are treated first followed by multivariate probability distributions. Examples of how to simulate univariate the probability distributions are provided in Probability Distributions Demo.xls. Section numbers in the text are used to organize and identify the distributions in the demonstration workbook.

3.1. Uniform Probability Distribution

Uniformly distributed random numbers are the basis for all random numbers and are simulated by Simetar using the =UNIFORM() function. The function can be programmed three different ways:

= UNIFORM (Min, Max, [CUSD or USD])
 = UNIFORM (B8, B9)
 = UNIFORM ()

where: Min is the minimum value for the distribution or a cell reference,
 Max is the maximum value for the distribution or a cell reference, and
 CUSD is an optional input value reserved for a correlated USD (or uniform standard deviate) required for correlating non-normal distributions. See Section 3.9.2 for simulating CUSDs.

The =UNIFORM() function defaults to a uniform standard deviate (USD) distributed between 0 and 1 if it is programmed as =UNIFORM(). This form of the function is an essential input in the other Simetar random number generators, particularly for simulating the native Excel probability distribution functions. Three examples of the UNIFORM function are provided below and in 3.1.1 of Probability Distribution Demo.xls.

	A	B	C	D	E	F	G
8	3.1 Generate Random Numbers Using the Uniform Distribution						
9	Minimum		10.0				
10	Maximum		20.0				
11	Unifrom Standard Deviate U(0,1)		0.924		Random No.	Formula	
12	Specific Formula using actual numbers			Uniform (10,20)	14.906	=UNIFORM(10,20)	
13	General Formula using cell references			Uniform	16.371	=UNIFORM(C9,C10)	
14	General Formula using cell references and USD			Uniform w/ USD	19.241	=UNIFORM(C9,C10,C11)	

3.2 Normal Related Probability Distributions

3.2.1 Normal. A normally distributed random number is simulated using the =NORM() function. The =NORM() function defaults to a standard normal deviate (SND) generator when no parameters are provided, as =NORM(). A SND is a normally distributed random variable with a mean of zero and standard deviation of one. The function is programmed using one of the following forms of the command:

= NORM (Mean, Std Dev, [USD])
 = NORM (B35, B36, D13)
 = NORM (B35, B36)
 = NORM ()

where: Mean is the mean of the distribution (or a cell reference, as B35),
 Std Dev is the standard deviation of the distribution (or a cell reference as B36), and
 USD is an optional uniform standard deviate. When a USD is not provided, Simetar
 generates its own uniform standard deviate. This optional variable is included so
 Simetar can simulate multivariate normal distributions.

	A	B	C	D	E	F	G
17	3.2.1 Generate Random Numbers Using the Normal Distribution						
18	Mean		10.000				
19	Std. Deviation		3.000				
20	Uniform Standard Deviate (USD)		0.057		Random No.	Formula	
21	Specific Formula using actual numbers			Normal	10.641	=norm(10,3)	
22	General Formula using cell references			Normal	12.377	=norm(C18,C19)	
23	General Formula specifying the USD			Normal	5.271	=norm(C18,C19,C20)	

3.2.2 Truncated Normal. A truncated normal distribution uses the =TNORM() function. The function is programmed as follows:

= TNORM (Mean, Std Dev, [Min], [Max], [USD])
 = TNORM (B47, B48, B49, B50, D13)

where: Mean is the mean for the distribution entered as a number or stored in a cell as B47,
 Std Dev is the standard deviation for the distribution as B48,
 Min is the absolute minimum value as B49 and is optional,
 Max is the absolute maximum value as B50 and is optional, and
 USD is the optional uniform standard deviate generated by =UNIFORM().

To simulate a truncated normal with a truncated minimum, use the function as:

= TNORM (Mean, Std Dev, Min, , [USD])

To simulate a truncated normal distribution with a truncated maximum, use the function as:

= TNORM (Mean, Std Dev, , Max, [USD])

	A	B	C	D	E	F	G
37	Truncated Normal -- Truncate the Minimum only						
38	Mean		10.000				
39	Std Deviation		3.000				
40	Minimum		5.000		Random No.	Formula	
41	General Formula using cell references			Tnorm Min	8.197	=Tnorm(C38,C39,C40)	
42							
43	Truncated Normal -- Truncate the Maximum only						
44	Mean		10.000				
45	Std Deviation		3.000				
46	Maximum		12.500		Random No.	Formula	
47	General Formula using cell references			Tnorm Max	11.615	=Tnorm(C44,C45,,C46)	

3.2.3 Two-Piece Normal. The two-piece normal distribution combines half of the densities for two normal distributions with the same mean and possibly different standard deviations. The distribution is simulated as:

=TPNORM(Mean, SD Lower, SD Upper, [USD])

where: Mean is the middle value for the distribution,
SD Lower is the standard deviation for distribution less than the Mean,
SD Upper is the standard deviation for distribution greater than the Mean, and
USD is an optional uniform standard deviate.

3.2.4 Modified Two-Piece Normal. The two piece normal distribution is fully defined by specifying the minimum, middle point, the maximum and the standard deviations for the two sides. The =MTPNORM() is specified as:

=MTPNORM(Min, Mid, Max, [USD], [Lower SD], [Upper SD])

where: Min is the minimum value for the random variable on the number scale (default -1),
Mid is the middle value for the random variable (default 0),
Max is the maximum value for the random variable on the number scale (default 1),
USD is an optional uniform standard deviate,
Lower SD is the number of standard deviations in the lower tail (default of 2 means the minimum value is two standard deviations below the middle value), and
Upper SD is the number of standard deviations in the upper tail (default of 2 means the maximum value is two standard deviations above the middle value).

3.2.5 Student's-t (Excel's). The student's t-distribution is native to Excel but can be simulated using Simetar by providing a USD generated by =UNIFORM(). The probability distribution is simulated as:

=TINV (USD, Degrees of Freedom)

where: USD is a uniform standard deviate generated by =UNIFORM(), and
Degrees of Freedom is self explanatory.

	A	B	C	D	E	F
68	3.2.5 Generate Random Numbers Using the Student's t (Excel) Distribution					
69	Uniform Standard Deviate		0.213			
70	df Degrees of Freedom		30.000		Random Nos.	Formulas
71	Student's t-distribution with cell references			Student's t	1.272	=TINV(C69,C70)

3.2.6 F (Excel's). The F distribution, an Excel function, is simulated as:

=FINV (USD, Degrees of Freedom1, Degrees of Freedom 2)

where: USD is a uniform standard deviate generated by =UNIFORM(), and
Degrees of Freedom 1 and 2 are self explanatory.

3.2.7 Chi-Squared (Excel's). The chi-squared distribution, an Excel function, is simulated as:

=CHIINV (USD, Mean)

where: USD is a uniform standard deviate generated by =UNIFORM (), and
Mean is the average for the distribution.

3.2.8 Log Normal (Excel's). The log normal distribution, an Excel function, is used to simulate quantities like a normal distribution. The distribution is simulated as:

=LOGINV (USD, Mean, Std Dev)

where: USD is a uniform standard deviate generated by =UNIFORM (),
Mean is the average, and
Std Dev is the standard deviation for the distribution.

The simulated values from =LOGINV() are in “natural log” form so take the anti-log of the stochastic values using the Excel function =LN().

3.2.9 Power Normal. The power normal distribution is simulated in Simetar using the =PNORM() function as:

=PNORM(Mean, Sigma, Exp P, [USD])

where: Mean is a real number and indicates the central tendency parameter for the distribution,
Sigma is a number greater than zero that represents the variance for the distribution,
Exp P is a value greater than zero, the exponent parameter for the distribution, and
USD is an optional uniform standard deviate.

3.2.10 Inverse Gaussian. The inverse Gaussian distribution is simulated using an iterative solution procedure. The =INVGAUS() function is programmed as:

=INVGAUS (Mu, Sigma, [USD], [Max Iter], [Precision])

where: Mu is a positive real number representing the first parameter of the distribution,
Sigma is a number greater than zero that indicates the shape parameter for the distribution,
USD is an optional uniform standard deviate,
Max Iter is an optional maximum iterations used to find the stochastic value, and
Precision is an optional term to specify the precision of the answer.

3.3 Continuous Probability Distributions

3.3.1 Gamma (Excel's). The gamma distribution, an Excel function, can be used to simulate the length of time to complete a task. The distribution is specified as:

=GAMMAINV (USD, Alpha, Beta)

where: USD is a uniform standard deviate generated by =UNIFORM (),
Alpha is the first parameter for the gamma distribution, and
Beta is the second parameter for the gamma distribution.

3.3.2 Truncated Gamma. The gamma distribution can be truncated at the lower or upper end with the =TGAMMAINV() function. The function is used as:

=TGAMMAINV (Alpha, Beta, [Min], [Max], [USD])

where: Alpha is the first parameter for the gamma distribution and must be greater than zero,
Beta is the second parameter for the gamma distribution and must be greater than zero,
Min is the optional value for the absolute minimum ($0 < \min < \max$),
Max is the optional value for the absolute maximum ($\min < \max < \infty$), and
USD is an optional uniform standard deviate.

3.3.3 Exponential. The exponential distribution can be used to simulate times between independent events that occur at a constant rate, such as arrivals at a service center. The distribution is simulated as:

= EXPONINV (Beta, [USD])

where: Beta is the only parameter for the exponential distribution, and
USD is an optional uniform standard deviate.

3.3.4 Double Exponential. The double exponential distribution can be used to simulate times between independent events that occur at a constant rate, such as arrivals at a service center. The distribution is simulated as:

= DEXPONINV (Beta, [USD])

where: Beta is the only parameter for the double exponential distribution, and
USD is an optional uniform standard deviate generated by =UNIFORM().

3.3.5 Weibull. The Weibull distribution is often used to simulate reliability or lifetimes for machinery. The distribution is simulated as:

= WEIBINV (Alpha, Beta, [USD])

where: Alpha is the first parameter for the Weibull distribution and must be greater than zero,
Beta, the second parameter for the Weibull distribution, must be greater than zero, and
USD is an optional uniform standard deviate generated by =UNIFORM().

3.3.6 Truncated Weibull. The Weibull distribution can be simulated with a finite minimum and/or maximum as:

=TWEIBINV (Alpha, Beta, [Min], [Max], [USD])

where: Alpha is the first parameter of the Weibull distribution and must be greater than zero, Beta is the second parameter of the Weibull distribution and must be greater than zero, Min is the absolute minimum ($0 < \min < \max$), Max is the absolute maximum ($\min < \max < \infty$), and USD is an optional uniform standard deviate.

3.3.7 Cauchy. The Cauchy distribution is a symmetrical distribution about its parameter theta (θ). If median and sigma parameters are not provided the function defaults to a =CAUCHY(0,1) random variable. The distribution can be simulated in Simetar as:

=CAUCHY ([Median], [Sigma], [USD])

where: Median is an optional value for the mid point of the distribution, Sigma is an optional term to indicate the shape of the distribution, and USD is an optional uniform standard deviate.

3.3.8 Logistic. A logistic distribution can be simulated using the =LOGISTICINV() function as:

=LOGISTICINV(Mu, Sigma, [USD])

where: Mu is the first parameter for the logistic distribution and it must be a real value, Sigma is the second parameter for the distribution and must be greater than zero, and USD is an optional uniform standard deviate.

3.3.9 Log-Log. The log-log distribution is simulated as:

=LOGLOGINV (Mu, Sigma, [USD])

where: Mu is any real value indicating the position of the distribution on the number scale, Sigma is a value greater than zero indicating the scale parameter, and USD is an optional uniform standard deviate.

3.3.10 Log-Logistic. The log-logistic distribution is simulated as:

=LOGLOGISTICINV (Alpha, Beta, [USD])

where: Alpha is a value greater than zero which represents the shape parameter, Beta is the scale parameter and must be greater than zero, and USD is an optional uniform standard deviate.

3.3.11 Extreme Value. An extreme value distribution can be simulated as:

=EXTVALINV (Mu, Sigma, [USD])

where: Mu is the real value indicating the location parameter for the extreme value distribution.
Sigma any value greater than zero indicating the scale parameter of the distribution, and
USD is an optional uniform standard deviate generated by =UNIFORM().

3.3.12 Pareto. A Pareto distribution can be simulated using the =PARETO() function as:

=PARETO(Alpha, Beta, [USD])

where: Alpha is the first parameter for a Pareto distribution and it must be greater than zero,
Beta is the second parameter for the distribution and it must be greater than zero, and
USD is an optional uniform standard deviate.

3.4 Finite-Range Continuous Probability Distributions

3.4.1 Triangle. The triangle distribution is defined by the minimum, mode, and maximum. The distribution can be simulated as:

=TRIANGLE (Min, Mode, Max, [USD])
=TRIANGLE (A95, A96, A97)

where: Min is the minimum for the distribution,
Mode is the mode for the distribution,
Max is the maximum for the distribution, and
USD is an optional uniform standard deviate.

	A	B	C	D	E	F	G
205	3.4.1 Generate Random Numbers Using the Traingle Distribution						
206	Minimum		12.000				
207	Mode		18.000				
208	Maximum		27.000				
209	Uniform Standard Deviate		0.925		Random Nos.	Formulas	
210	Triangle Distribution			Triangle	23.134	=triangle(C206,C207,C208)	
211	Genearalized triangle			Triangle	23.823	=triangle(C206,C207,C208,C209)	

3.4.2 Beta (Excel's). The beta distribution, an Excel function, can be used to simulate the proportion of defective items in a shipment or time to complete a task. The distribution is simulated as:

=BETAINV (USD, Alpha, Beta, [Min], [Max])
=BETAINV (UNIFORM (), Alpha, Beta)

where: USD is a uniform standard deviate generated by =UNIFORM (),
Alpha is the first parameter for the distribution,
Beta is the second parameter for a beta,
Min is an optional value for truncating the minimum of the distribution, and
Max is an optional value for truncating the maximum of the distribution.

3.4.3 PERT Distribution. A PERT distribution can be simulated by Simetar using the =PERTINV() function as:

=PERTINV(Min, Middle, Max, [USD])

where: Min is a lower bound parameter,
Middle is a middle parameter with $\text{min} < \text{middle} < \text{max}$,
Max is an upper bound parameter, and
USD is an optional uniform standard deviate.

3.4.4 Cosine. The cosine distribution is simulated by Simetar using an iterative solution procedure. The =COSINV() function is programmed as:

=COSINV(Center, Radius, [USD], [Max Iter], [Precision])

where: Center is a real number that represents the first parameter for a cosine distribution,
Radius is a positive value that represents the second parameter,
USD is an optional uniform standard deviate,
Max Iter is the maximum number of iterations used to find the stochastic value, and
Precision is an optional term to specify the precision of the answer.

3.4.5 Semicircle. The semicircle distribution is simulated as:

=SEMICIRCINV(Center, Radius, [USD], Max Iter, Precision)

where: Center is a real number that indicates the first parameter of the distribution,
Radius is the second parameter for the distribution and must be greater than zero,
USD is an optional uniform standard deviate,
Max Iter is the maximum number of iterations to find the value ($\text{max} > 0$), and
Precision is a positive value to specify how precise the optimum answer should be. If an optimum answer is not found within the precision level in the maximum number of iterations, #VALUE1 error is returned.

3.5 Analogs to Finite Range Probability Distributions

3.5.1 GRK. The GRK distribution is an empirical substitute for the triangle distribution and is similar to a two piece normal distribution. The GRK distribution simulates values less than the minimum about two percent of the time. Values greater than the maximum are observed about two percent of the time. A GRK distribution can be simulated as:

= GRK (Min, Middle, Max, [USD])
= GRK (A95, A96, A97)

where: Min is the value for the minimum,
Middle is the value for the mid point of the distribution,
Max is the value (or cell) for the maximum, and
USD is an optional uniform standard deviate.

3.5.2 GRKS. The GRKS distribution is a continuous probability distribution for sampling from a minimum data population. Given a minimum, middle value and a maximum to describe the population the =GRKS() function is a continuous distribution substitute for the triangle distribution. The LSD and USD parameters indicate the number of standard deviations below and above the middle value that the distribution can extend. An LSD of 2 implies the minimum is at approximately the 2nd percentile and a LSD of 3 implies sampling with a minimum at approximately the 0.5 percentile. Program =GRKS() as follows:

=GRKS (Min, Middle, Max, [USD], [LSD], [USD])
 =GRKS (C250, C251, C252, C253, C259, C260)

where: Min is the value for the minimum,
 Middle is the value for the mid point of the distribution,
 Max is the value (or cell) for the maximum,
 USD is an optional uniform standard deviate,
 LSD optional number of standard deviations below the middle, as D97, and
 USD optional upper number of standard deviations above the maximum, as D98.

	A	B	C	D	E	F	G
248	3.5.1 Generate Random Numbers Using the GRK Distributions						
249	Parameters for the GRK Distribution						
250	Minimum		3.000				
251	Mode		5.000				
252	Maximum		16.000				
253	Uniform Standard Deviate		0.853		Random Nos.	Formulas	
254	Specific Formula using actual numbers			GRK	3.197	=grk(3,5,16,UNIFORM())	
255	General Formula using cell references			GRK	8.200	=grk(C250,C251,C252)	
256	General Formula with a USD specified			GRK	10.857	=grk(C250,C251,C252,C253)	
257	3.5.2 Generate Random Numbers Using the GRKS Distributions						
258	General formula without bounds			GRKS	4.564	=grks(3,5,16,UNIFORM())	
259	Lower no. of standard deviations		3.000	GRKS Min	4.748	=grks(3,5,16,UNIFORM(),C259)	
260	Upper no. of standard deviations		4.000	GRKS Max	7.889	=grks(C250,C251,C252,C253, C260)	
261				GRKS Min&Max	7.889	=grks(C250,C251,C252,C253,C259,C260)	

3.6 Discrete Probability Distributions

3.6.1 Bernoulli. A Bernoulli distribution can be used to simulate the occurrence of an event, such as a machine failure during a given time period. Simulate a Bernoulli distribution as:

=BERNOULLI (P)
 =BERNOULLI (A10)

where: P is the probability ($0 < P < 1$), of the variable or condition being true (or 1).

	A	B	C	D	E	F
265	3.6.1 Generate Random Numbers Using the Bernoulli Distribution					
266	Parameter for the Bernoulli Distribution					
267	Probability of a success or a 1		0.450		Random Nos.	Formulas
268	Specific Formula using actual numbers			Bernoulli	0.000	=bernoulli(0.45)
269	General Formula using cell references			Bernoulli	1.000	=bernoulli(C267)

3.6.2 Binomial. The binomial distribution is a discrete distribution for simulating the number of successes in N independent Bernoulli trials each having a probability P of success. Other applications are to simulate the number of units demanded in a given time period. Simulate the binomial distribution as:

=BINOMINV (N, P, [USD])

where: N is the number of trials,
 P is the probability of a positive success, and
 USD is an optional uniform standard deviate.

3.6.3 Negative Binomial. The negative binomial distribution simulates the number of failures before the Nth success in a sequence of independent Bernoulli trials each having a probability P of success. Simulate the negative binomial distribution as:

=NEGBINOMINV (N, P, [USD])

where: N is a positive integer representing the number of failures before the next success,
 P is the probability of success, and
 USD is an optional uniform standard deviate.

3.6.4 Multinomial. The multinomial probability distribution returns either an array of values or a scalar, depending upon how it is used. If the probabilities (Probs) are entered as an array the function returns an array, but if Probs is a scalar it returns a scalar. An example of the multinomial distribution in Step 3.6.4 of Probability Distributions Demo.xls demonstrates how the function can be used both ways.

=MULTINOMINV(No. Trials, Probs, [USD])

where: No. Trials is the sample size (integer greater than zero) used in the distribution,
 Probs is a vector of cell probabilities associated with each cell's random variable.
 Individual values are between zero and one and must sum to one. If a single value is entered for Probs the function returns a binomial random variable.
 USD is an optional univariate standard deviation.

3.6.5 Poisson. The Poisson distribution simulates the number of events that occur in an interval of time, such as arrivals at a service point. The distribution can also be used to simulate random quantities demanded during an interval of time. Simulate the Poisson distribution as:

=POISSONINV (L, [USD])

where: L, the only parameter for a Poisson, must be positive and is generally an integer, and
 USD is an optional uniform standard deviate.

3.6.6 Geometric. The geometric distribution simulates the number of failures before the first success in a sequence of independent Bernoulli trials each with a P probability. Also this distribution can simulate the number of items demanded in a given period. The geometric distribution is simulated as:

=GEOMINV (P, [USD])

where: P is the probability of each independent Bernoulli trial, and
 USD is an optional uniform standard deviate.

3.6.7 Hypergeometric. The Hypergeometric distribution is used to simulate the number of units that are acceptable in a sample of size K taken from a population of size N when it is known that M of the units in the population are acceptable. This is a sample without replacement problem made famous by the urn with N balls, m of which are green, (N-M are red), and a sample of K balls are drawn. The Hypergeometric function returns the number of red balls in the sample of K. Simulate the Hypergeometric distribution as:

=HYPERGEOMINV (N, M, K, [USD])

where: N is the population size,
M is the number of units in the population with the desired characteristic,
K is the sample size, and
USD is an optional uniform standard deviate

3.7 Sample Based Probability Distributions

3.7.1 Empirical. An empirical distribution can be simulated by Simetar using the =EMPIRICAL() or the =EMP() function. The function assumes a continuous distribution so it interpolates between the specified points on the distribution (S_i) using the cumulative distribution probabilities ($F(S_i)$). The most direct form of the function is =EMPIRICAL(S_i) or =EMP(S_i) which causes Simetar to calculate the $F(S_i)$ and USD values for the distribution. The function is programmed as follows:

= EMPIRICAL(S_i , $F(S_i)$, [USD], [Normal Tails])
= EMP(B75:B89, A75:A89, D13)

where: S_i represents an array of N sorted random values including the min and max,
 $F(S_i)$ cumulative probabilities for the S_i values, including the end points of zero and one,
USD is an optional uniform standard deviate generated by =UNIFORM(), and
Normal tails is an optional term to extend the tails of the distribution beyond the end of the data (enter a 1) or to truncate the distribution with the default value of 0.

Note: $i = 1$ to n for the S_i and $F(S_i)$ parameters denotes that these are ranges and not individual values.

	A	B	C	D	E	F	G
321	3.7.1 Generate Random Numbers Using the Empirical Distribution						
322	Parameters for an Empirical Distribution						
323							
324	P(x)	Price					
325	0	1.5030					
326	0.038462	1.5031					
327	0.115385	1.9440					
328	0.192308	2.0112					
329	0.269231	2.0743					
330	0.346154	2.2647					
331	0.423077	2.2847		Uniform Std Deviate			
332	0.500000	2.3649		0.834			
333	0.576923	2.3749					
334	0.653846	2.4551			Random Nos.	Formulas	
335	0.730769	2.5052			Four Forms of the Empirical Distribution using a Common USD		
336	0.807692	2.5453		Empirical 1	2.604	=empirical(B325:B339,A325:A339,D332)	
337	0.884615	2.7156		Empirical 2	2.604	=emp(B325:B339,A325:A339,D332)	
338	0.961538	3.2467		Empirical 3	2.604	=emp(B325:B339,,D332)	
339	1	3.2468		EMP Norm Tail	2.604	=emp(B325:B339,A325:A339,D332,1)	

3.7.2 Truncated Empirical. A truncated empirical distribution is the same as an empirical distribution but with a defined minimum and maximum. The distribution is simulated as:

=TEMPIRICAL((S_i, F(S_i), Min, Max, [USD])

where: S_i represents an array of N sorted random values including the min and max,
 F(S_i) cumulative probabilities for the S_i values, including the end points of zero and one,
 USD is an optional uniform standard deviate generated by =UNIFORM(),
 Min is the minimum for the distribution, and
 Max is the maximum for the distribution.

3.7.3 Discrete Empirical. When it is not appropriate to interpolate between the S_i points on the empirical distribution, then the data are said to be distributed discrete empirical. This distribution is applicable if the data can only take on set values. Each value is assumed to have an equal chance of being selected. The function is programmed in Simetar as follows:

=DEMPIRICAL (S_i, [USD])
 =DEMPIRICAL (B75:B89, D13)

where: S_i represents an array of n random values; the values do not have to be sorted, and
 USD is an optional uniform standard deviate.

	A	B	C	D	E	F	G	
347	3.7.3 Generate Random Numbers Using the Discrete Empirical Distribution							
348	Sorted Values for a Discrete Random Variable							
349	1							
350	5	USD	0.3009					
351	8				Random Nos.	Formulas		
352	9			Discrete Emp	9.000	=DEMPIRICAL(A349:A353)		
353	11			Discrete Emp	5.000	=DEMPIRICAL(A349:A353,C350)		

3.7.4 Kernel Density Estimated Random Variable. The =KDEINV() function uses Parzen type kernel density estimators to evaluate a smoothed value that represents a point on a cumulative distribution function (CDF). Eleven alternative kernel density estimators can be used to smooth an empirical distribution and simulate random values in Simetar. A graphical representation of the kernel density smoothed function can be developed using the smoothing option in the CDF chart tool (see Section 6.2 for the CDF chart function). The kernel density estimated random variable function is simulated as:

=KDEINV(Data Range,[BW], [KE], [USD], [Max Iter], [Precision])

where: Data range is the location for data series for the empirical distribution to simulate,
 BW is an optional bandwidth to use in estimating the influence of each data point on the CDF estimation,
 KE is an optional term to specify the kernel estimation type used to estimate the CDF. The KE types are: Gaussian (0 or 1), Uniform (2), Casinus (3), Triangle (4), Triweight (5), Epanechnikow (6), Quartic (7), Cauchy (8), Double Exponential (9), Histogram (10), and Parzen (11),
 USD is an optional uniform standard deviate,
 Max Iter is the maximum number of iterations to use to find the result, and

Precision is an optional term to specify how precise the final solution should be. If an optimal answer is not found #VALUE! will appear in the cell.

	A	B	C	D	E	F	G
353	3.7.4 Generate Random Numbers Using the Kernel Density Estimator Distribution						
354	Unsorted Data Range						
355	1.500	Bandwidth	0.1163	=BANDWIDTH(A355:A367)			
356	1.940	Kernel Estimator	0	"0" indicated Gaussian			
357	2.540	USD	0.0148				
358	2.360	Max Iterations	1000				
359	2.280	Precision	0.00001				
360	2.370						
361	2.070			Random Nos.	Formulas		
362	2.500		Kernel Density E	1.398670281	=KDEINV(A355:A367,C355,C356,C357,C358,C359)		
363	2.260						
364	3.240						
365	2.710						
366	2.450						
367	2.507						

3.7.5 Discrete Uniform. A discrete uniform random variable can take on only certain values, each with an equal probability. For example, a fair die can take on one of six values (1, 2, 3, 4, 5, 6) with an equal probability. To simulate a discrete uniform random variable use the =RANDSORT() function. For example, if the random values to define a distribution are 1, 2, 3, 4, 5, 6, and are stored in cells A1:A6, simulate a random value, by typing the following command in a cell:

=RANDSORT(A1:A6)

3.7.6 Random Sorting. The array form of the =RANDSORT() function can be used to simulate (sample) random draws of a list of names or objects or numbers without replacement. For example, if five names Jim, Joe, Sam, John, and Bill are to be randomly sorted (shuffled), enter the names in an array and use =RANDSORT() as an array function. Assume the five names are in A1:A5 and the random sample is to appear in B1:B5; type the following command in B1 after highlighting array B1:B5:

=RANDSORT(A1:A5)

Press *Control Shift Enter*, rather than Enter after typing the function, because this is the array form of the function. Press F9 to “resort” the data for a second iteration or sample.

	A	B	C	D	E	F
382	3.7.6 Random Sorting a List of Names Using the Array Function form of RANDSORT					
383	List of Names to be Sampled				Random Nos.	Formula
384	Ace		Random Shuffle of Objects	Jack	=randsort(A384:A387)	
385	King		Using the Array Function form:	Ace		
386	Queen			King		
387	Jack			Queen		
388			Draw a single name at random	Jack	=randsort(A384:A387)	

3.7.7 Bootstrapping (Random Sampling with Replacement). Bootstrap sampling techniques are used for advanced simulation problems and assume that past deviates or errors can be re-sampled an infinite number of times. This method of sampling can be accomplished using the =BOOTSTRAPPER() function which samples from a known distribution with replacement. An example of the function is provided below and in Step 3.7.7 in Probability Distributions Demo.xls.

=BOOTSTRAPPER (Array of Random Values, [Preserve Rows])
 =BOOTSTRAPPER (A27:A31, 1)

where: Array of random values is the location for the array of random values to be sampled during simulation, and
 Preserve Rows the optional term to retain the order of the values in rows, if the array of random variables is a matrix.

	A	B	C	D	E	F	G	H	I	J
392	3.7.7 Generate Random Numbers Using the Bootstrap Simulation									
393	Random Values for X				Random Nos.	Formulas				
394	1			Bootstrap	1	=BOOTSTRAPPER(\$A\$394:\$A\$399)				
395	2			Bootstrap	5	=BOOTSTRAPPER(\$A\$394:\$A\$399)				
396	3			Bootstrap	5	=BOOTSTRAPPER(\$A\$394:\$A\$399)				
397	4			Bootstrap	6	=BOOTSTRAPPER(\$A\$394:\$A\$399)				
398	5			Bootstrap	6	=BOOTSTRAPPER(\$A\$394:\$A\$399)				
399	6			Bootstrap	4	=BOOTSTRAPPER(\$A\$394:\$A\$399)				
400										
401	Bootstrap Simulation with a Matrix of Observations: Preserve the values across the row.									
402					Random Bootstrap Nos. Maintaining the Rows					
403	X1	X2	X3		X1	X2	X3	Formula		
404	1.0	23.4	4.5		5.0	55.5	1.5	=BOOTSTRAPPER(\$A\$404:\$C\$408,1)		
405	4.0	56.7	3.3		4.0	56.7	3.3	=BOOTSTRAPPER(\$A\$404:\$C\$408,1)		
406	8.0	12.9	6.9		1.0	23.4	4.5	=BOOTSTRAPPER(\$A\$404:\$C\$408,1)		
407	9.0	6.4	2.3		5.0	55.5	1.5	=BOOTSTRAPPER(\$A\$404:\$C\$408,1)		
408	5.0	55.5	1.5		4.0	56.7	3.3	=BOOTSTRAPPER(\$A\$404:\$C\$408,1)		

3.8 Time Series Probability Distributions

3.8.1 Random Walk. The =RANDWALK() function generates a random variable that is characteristic of a random walk. A random walk distribution for x_t is characterized as $X_t = X_{t-1} + \tilde{\epsilon}_t$ where $\tilde{\epsilon}_t$ is normally distributed. Simulating a variable for N iterations will result in a sample of length N. The function is used as:

=RANDWALK (Mean, Std Dev, [USD], [Distribution], [Initial Value], [Coefficient])

where: Mean is the expected value for the random variable,
 Std Dev is the standard deviation for the variable and is greater than zero,
 USD is an optional uniform standard deviate,
 Distribution is an optional code for the distribution for generating random changes as:
 normal (0 or 1), uniform (2), cosine (3), Cauchy (8), double exponential (9), logistic (12), extreme value (13), exponential (14), and log normal (15),
 Initial Value is an optional initial value to start the random sequence; the default is zero,
 Coefficient is an optional value on the lag variable as α in $X_t = \alpha X_{t-1} + \tilde{\epsilon}_t$.

	A	B	C	D	E	F	G	H
412	3.8 Generate Random Numbers Using the Random Walk							
413	Mean			10				
414	Standard Deviation			3				
415	USD		0.099617958					
416	Distribution	Gaussian		0				
417	Initial Value			4.5				
418	Coefficient			0				
419	Random Walk sequence				RandomWalk	6.148805478	=randwalk(C413,C414,C415,C416,C417,C418)	

3.9 Multivariate Distributions

3.9.1 Correlated Standard Normal Deviates. Correlated standard normal deviates (CSND's) are generated in Simetar using the =CSND() function. An array of correlated standard normal

deviates can be used to simulate multivariate normal (MVN) probability distributions in a two step procedure. An array of CSNDs is simulated as:

=CSND (Correlation Matrix Range, [Optional Range of Independent SNDs])
 =CSND (B152:G157)
 =CSND (B152:G157, B161:B166)

where: Correlation Matrix Range specifies the location of a non-singular NxN correlation matrix. Calculate the correlation matrix using the function described in Section 12.3. Optional Range of Independent SND's (ISND's) is an Nx1 array of SND's generated using =NORM() in N cells.

As an array function =CSND() must be used as follows: highlight the output location for N cells and type the command =CSND (correlation matrix location, optional range of ISNDs) and press the *Control Shift Enter* keys.

	A	B	C	D	E	F
422	3.9.1 Generate Correlated Standard Normal Deviates					
423	Correlation Matrix for 3 Random Variables					
424		Price 1	Price 2	Price 3		
425	Price 1	1	0.868849998	0.980957963		
426	Price 2	0	1	0.815781325		
427	Price 3	0	0	1		
428	CSND is an array function so highlight a 3x1 array for the three variables					
429					Random Nos.	Formulas
430				Price 1	1.37071	=csnd(B425:D427)
431				Price 2	1.05759	
432				Price 3	1.22886	

3.9.2 Correlated Uniform Standard Deviates. Correlated uniform standard deviates (CUSDs) are used to simulate multivariate non-normal (e.g., empirical) probability distributions in a two step process. An array of CUSDs is simulated as:

=CUSD (Correlation Matrix Range, [Optional Range of Independent SNDs])
 =CUSD (B152:G157)
 =CUSD (B152:G157, B161:B166)

where: Correlation Matrix Range specifies location of a non-singular NxN correlation matrix. Calculate the correlation matrix using the function described in Section 12.3. Optional Range of Independent SNDs is an Nx1 array of SNDs generated using =NORM() in N cells.

As an array function =CUSD() must be used as follows: highlight the output location for N cells and type the command =CSND (correlation matrix location, optional range of ISNDs), and press the *Control Shift Enter* keys.

	A	B	C	D	E	F
435	3.9.2 Generate Correlated Uniform Standard Deviates					
436	Correlation Matrix for 3 Random Variables					
437		Price 1	Price 2	Price 3		
438	Price 1	1	0.868849998	0.980957963		
439	Price 2	0	1	0.815781325		
440	Price 3	0	0	1		
441	CUSD is an array function so highlight a 3x1 array for the three variables					
442					Random Nos.	Formulas
443				Price 1	0.53728	=cusd(B438:D440)
444				Price 2	0.79301	
445				Price 3	0.46306	

3.9.3 Multivariate Normal (MVN) Distribution in One Step

Simetar provides a one step function for simulating a MVN distribution. The =MVNORM() function uses an Nx1 array of means and an NxN covariance matrix to generate correlated random values that are distributed multivariate normal. The array function is entered as follows:

=MVNORM (Means Vector, Covariance Matrix, [Array of ISNDs])

where: Means Vector is an Nx1 array of the averages to use for simulating MVN values, and Covariance Matrix is an NxN covariance matrix for the N random variables. Array of ISNDs is an optional Nx1 array of independent standard normal deviates generated with n cells of =NORM().

To use the array function, first highlight the number of cells equal to the number of means at the output location, second type =MVNORM (location for the array of means, location of covariance matrix), and press the *Control Shift Enter* keys.

	A	B	C	D	E	F	G
448	3.9.3 Multivariate Normal Distribution in One Step						
449	Covariance Matrix						
450		Price 1	Price 2	Price 3	Prod 1	Prod 2	Prod 3
451	Price 1	100.000	173.770	147.144	-8.980	-43.833	-190.314
452	Price 2		400.000	244.734	-9.523	-183.525	-246.381
453	Price 3			225.000	-12.938	-63.369	-304.910
454	Prod 1				9.000	39.829	63.377
455	Prod 2					625.000	425.693
456	Prod 3						1296.000
457				Array Function for			
458		Assumed Means		Correlated MVN Values	Formulas for random numbers		
459	Price 1	100.0		98.333	=mvnorm(B459:B464,B451:G456)		
460	Price 2	200.0		185.444	=mvnorm(B459:B464,B451:G456)		
461	Price 3	250.0		246.706	=mvnorm(B459:B464,B451:G456)		
462	Prod 1	25.0		22.737	=mvnorm(B459:B464,B451:G456)		
463	Prod 2	190.0		211.980	=mvnorm(B459:B464,B451:G456)		
464	Prod 3	260.0		252.160	=mvnorm(B459:B464,B451:G456)		

3.9.4 Multivariate Normal Distribution in Two Steps. A general formula for simulating a multivariate normal distribution is accomplished by first generating a vector of CSNDs and then using the CSNDs in the formula for a normal distribution. In step 1 an Nx1 array of CSNDs is generated using =CSND(), see Section 3.9.1. The example provided here is for a three variable model so N equals 3. Assume the non-singular covariance matrix is in A1:C3, the three means are in cells B7:B9, and the three standard deviations are in cells C7:C9.

Step 1: In A4:A6 = CSND (A1:C3)

Step 2: In A7 = B7 + C7 * A4
 In A8 = B8 + C8 * A5
 In A9 = B9 + C9 * A6

These three Excel statements can be repeated for N variables. The three random variables will be appropriately correlated within each period but will be independent across periods.

An alternative two step procedure for simulating a multivariate normal distribution uses a vector of CUSDs. In step 1 an Nx1 array of CUSDs is generated using =CUSD(), see Section 3.9.2. In step 2 use the =NORM() function to simulate the random values. The example provided here is

for a three variable model so N equals 3. Assume the non-singular correlation matrix is in A1:C3, the three means are in cells B7:B9, and the three standard deviations are in cells C7:C9.

Step 1: In A4:A6 = CUSD (A1:C3)

Step 2: In A7 = NORM(B7, C7, A4)
 In A8 = NORM(B8, C8, A5)
 In A9 = NORM(B9, C9, A6)

These three Simetar statements can be repeated for N variables. The three random variables will be appropriately correlated within each period but will be independent across periods.

3.9.5 Multivariate Empirical (MVE) in One Step

Simetar provides a one step function for simulating a MVE distribution. The =MVEMPIRICAL() function uses as input the MxN matrix of the M observations for the N random variables. The result is an Nx1 array of MVE correlated random values for the N variables. Program the function as:

=MVEMPIRICAL (Range for Random Variables ,,,, [Vector of Means], [Type])

where: Range for Random Variables is an MxN matrix of the M observed values for the N random variables,
 Vector of Means is an array of forecasted means for the N random variables, and
 Type is a option code for the type data transformation used to generate the forecasted means for the MVE: (0) for actual data, (1) for percent deviations from mean, (2) for percent deviations from trend, and (3) is for differences from the mean.

The =MVEMPIRICAL() function is an array function so highlight an Nx1 array at the output location and type the function, followed by pressing the *Control Shift Enter* keys. An example of using the one step =MVEMPIRICAL() function is provided below for a MVE distribution with 6 (N) variables and 13 (M) observations.

	A	B	C	D	E	F	G
496	3.9.5 Multivariate Empirical Distribution in One Step						
497	Historical Data for the Random Variables						
498		Var 1	Var 2	Var 3	Var 4	Var 5	Var 6
499	1	1.500	2.420	1.370	119.375	34.428	87.744
500	2	1.940	2.570	1.700	119.844	37.674	89.396
501	3	2.540	3.720	2.270	84.613	34.070	63.779
502	4	2.360	3.720	2.100	116.165	32.748	55.428
503	5	2.280	2.610	2.120	118.503	39.498	63.077
504	6	2.370	3.000	2.251	108.581	34.333	59.256
505	7	2.070	3.240	1.873	131.480	39.305	72.616
506	8	2.500	3.260	2.310	100.705	38.213	59.912
507	9	2.260	3.450	2.128	138.608	37.575	72.805
508	10	3.240	4.550	3.190	113.453	35.812	55.614
509	11	2.710	4.300	2.340	127.051	36.314	67.471
510	12	2.450	3.380	2.200	127.043	39.740	69.546
511	13	2.007	2.605	1.815	133.300	43.265	66.469
512							
513	Variables	One Step MVE	Formulas used for the random numbers			Vector of Means	
514	0.115384616	89.186	=MVEMP(B499:G511,...,F514:F519,1)			100.000	
515	0.192307696	154.426				200.000	
516	0.269230783	246.795				250.000	
517	0.346153855	27.924				25.000	
518	0.423076928	218.821				190.000	
519	0.5	285.841				260.000	

3.9.6 Multivariate Empirical (MVE) Distribution in Two Steps

Multivariate empirical distributions can be simulated in two steps using the =EMP() function and an array of correlated uniform standard deviates or CUSDs generated using the =CUSD() function described in Section 3.9.2. An example of the two step MVE is provided for a three variable model, assuming the correlation matrix is in H1:J3, the forecasted means are in cells A1:A3, the fractional deviations (S_i) from the mean are in cells C1:E12, and the three variables the probabilities for the deviates ($F(S_i)$) are in B1:B12.

Step 1:	In A14:A16	= CUSD (H1:J3)
Step 2:	In C14	= A1 + A1 * EMP(C1:C12, B1:B12, A14))
	In C15	= A2 + A2 * EMP (D1:D12, B1:B12, A15))
	In C16	= A3 + A3 * EMP (E1:E12, B1:B12, A16))

The values in cells C14:C16 are appropriately correlated based on the correlation matrix in cells H1:J3 and are distributed empirical about the respective forecasted means in cells A1:A3. The formulas in cells A14:A16 and C14:C16 can be repeated for as many periods (years) as the model simulates. Simulated MVE values will be correlated within each period but will be independent across periods.

3.9.7 Multivariate Mixed Distribution

Simetar can simulate a multivariate mixed distribution (MVM) which has correlated variables that are distributed differently. For example a MVM could include variables that are distributed uniform, empirical, normal, and beta. To simulate a MVM, use the =CUSD() function to simulate an Nx1 vector of correlated uniform standard deviates, one CUSD for each variable. Use each of the CUSDs in the appropriate Simetar function to simulate the random variables. Using an example of a four variable MVM with the variables distributed uniform, empirical, normal and beta, respectively, use the following functions:

Step 1:	=CUSD(Correlation Matrix Range)
Step 2:	=UNIFORM(Min, Max, CUSD ₁)
	=EMP(S_i , $F(S_i)$, CUSD ₂)
	=NORM(Mean, Std, Dev, CUSD ₃)
	=BETAINV(CUSD ₄ , Alpha, Beta, [Min], [Max])

where: CUSD_{*i*} values refer to the *i*th correlated uniform standard deviate simulated in the Nx1 CUSD array.

The simulated random variables will be appropriately correlated based on the correlation matrix.

3.9.8 Multivariate Log Normal. A log normally distributed series of random variables can be simulated multivariate using the =MVLOGNORM() array function. The function is used as:

=MVLOGNORM (Mean Vector, Covariance, [Array of ISNDs], [Matrix Row], [Moments])
=MVLOGNORM (A1:A4, B1:E4, F1:F4, 1, TRUE)

where: Means Vector is the location of the Nx1 vector of means. If the Moments switch is true, each mean must be greater than zero. If the Moments is false, mean are reals.
 Covariance is an NxN covariance matrix for the series,
 Array of ISND is an optional nx1 array of independent standard normal deviates generated with N cells of =NORM().
 Matrix Row is an optional term for the ith variable if the function is to return only the random value for the ith variable (Leaving this value blank makes the function return n values so treat it as an array function with *Control Shift Enter.*), and
 Moments is an optional switch to use the function two ways: if the term is TRUE ('1') the function is for the moments of a log normal vector, and FALSE ('0') indicates the moments are for a transformed normal distribution.

3.9.9 Multivariate Student's t. A distribution of N variables can be simulated multivariate Student's t using the =MVTINV() array function as:

```
=MVTINV (Means Vector, Covariance Matrix, [Array of ISND], [Matrix Row])
=MVTINV (A1:A4, B1:E4, F1:F4, 1,)
=MVTINV (A1:A4, B1:E4, F1:F4)
```

where: Means Vector is the location of the Nx1 vector of means,
 Covariance Matrix is the location of the NxN covariance matrix for the series,
 Array of ISND is an optional Nx1 array of N cells with =NORM() SNDs, and
 Matrix Row is the optional ith variable if only the random number for the ith series is to be simulated. Leaving this value blank makes the function return N values so treat it as an array function with *Control Shift Enter.*

3.9.10 Hotelling T-Squared. The Hotelling T^2 distribution is a multivariate analog to the univariate Student's t distribution. If \underline{x} is a Px1 random vector distributed as multivariate normal with a zero mean vector and an identity covariance matrix and W is a PxP random matrix distributed as Wishart with an identity covariance matrix and m degrees of freedom. And \underline{x} and W are independent, then the variable $T^2 = m \underline{x}^T W^{-1} \underline{x}$ is distributed as a Hotelling T^2 random variable. A special case is the Hotelling T^2 random variable with 1 and M degrees of freedom which is an F distribution with 1 and M degrees of freedom. The parameters for the Hotelling T^2 function, which produces a Hotelling T^2 random variable, are p and df = M. Simulate Hotelling T-Squared distribution as:

```
=HOTELLTINV(P, DF, [USD])
```

where: P is an integer indicating the dimension of the PxP covariance or identity matrix for a Wishart distribution,
 DF is the degrees of freedom or the number of observations in the MxP data matrix for a Wishart distribution, and
 USD is an optional uniform standard deviate.

3.9.11 Wishart. The Wishart distribution is a matrix generalization of the univariate chi square distribution. The Wishart array function produces a matrix of random values that are distributed Wishart. The distribution is derived from an MxP matrix X of normally distributed independent vectors with mean zero and covariance matrix C. The PxP matrix of $X'X$ has a Wishart

distribution and is simulated as:

=WISHINV(C, DF)

where: C is a PxP covariance matrix that is positive definite, and
DF is the degrees of freedom or the number of rows in an MxP data matrix of values used to calculate C.

The Wishart function is an array function so highlight a PxP block of cells, type the function, and end by pressing the *Control Shift Enter* keys.

3.9.12 Wilks' Lambda. If two independent random matrices, X and Y, are distributed as Wishart, both with a PxP identity covariance matrix and N1 and N2 degrees of freedom, respectively, then the scalar $|X|/|X+Y|$ has the Wilks' lambda distribution with P, N1, and N2 degrees of freedom. This distribution is found in several likelihood ratio tests in multivariate testing settings. Simulate Wilks' lambda distribution as:

=WILKSLINV(P, N1, N2)

where: P is an integer representing the dimension of the Wishart random matrix PxP
N1 is the integer value for the degrees of freedom in the random Wishart matrix X, and
N2 is the integer value for the degrees of freedom in the random Wishart matrix Y.

3.9.13 Dirichlet. A Dirichlet series of correlated random variables can be simulated using the Dirichlet array function as:

=DIRICHINV(Alpha Array, [Array of IUSD], [Matrix Row])

where: Alpha Array is the location of an Nx1 array of parameter values for the Dirichlet distribution; each value is greater than zero,
Array of IUSD is the location of an optional nx1 array of independent uniform standard deviates simulated =UNIFORM(), and
Matrix Row is the ith variable of the random series if the function is to return only the ith series. Leaving this value blank makes the function return n values so treat it as an array function with *Control Shift Enter*.

3.9.14 Uncorrelating Random Deviates (USD and SND). In advanced simulation applications it is useful to uncorrelate random values. Simetar provides a function to calculate the implicit independent deviates from a vector of CUSDs. The uncorrelated standard normal deviates function, =USND(), converts a vector of CSNDs to a vector of independent SNDs. The function is programmed as:

=USND (Correlation Matrix, CSND Array)

where: Correlation Matrix is the cell reference location for the correlation matrix used to generate the CSNDs, and
CSND Array is the cell reference location for the array of CSNDs to be converted to independent SNDs.

The uncorrelated uniform standard deviates function, =UUSD(), converts correlated uniform standard deviates (CUSDs) to uncorrelated USDs. The function is programmed as follows:

=UUSD (Correlation Matrix, CUSD Array)

3.10 Iteration Counter


For advanced simulation applications it is useful to use the iteration number to key a simulation model to perform certain calculations. For example a table lookup function can be used to draw values from a table where the rows correspond to the iterations for previously generated and tested random values. The iteration number function in Simetar is =ITERATION() and returns the iteration number from 1 to N, where N represents the number of iterations. As indicated in the example below, the function returns “1” until the workbook is simulated. Selecting the cell with =ITERATION() as a KOV for simulation will produce a series of values: 1, 2, 3, ... , 500 for a stochastic simulation with 500 iterations.

=ITERATION ()

	A	B	C	D	E
721	3.10 Iteration Counter in Simetar shows the iteration during simulation				
722	Count of present iteration				1 =ITERATION()

4.0 Parameter Estimation for Probability Distributions

4.1 Parametric Probability Distributions

A univariate parameter estimator in Simetar estimates the parameters for simulating a random variable for 16 parametric probability distributions. The univariate parameter estimator is activated by using the  icon. The Simetar menu for the univariate parameter estimator requires the user to specify the historical data series for the random variable and the method for estimating the parameters: method of moments or maximum likelihood estimator (Figure 12). If a variable is not consistent with a distribution, its parameter cells will be blank rather than contain a value.

Simetar also prepares the equations for simulating the random variable using the calculated distribution parameters in the Formulas column of the example above. The formulas in the Formulas column can be simulated to test how well the different assumed distributions simulate the random variable. The =CDFDEV() function can be used to calculate a test scalar to determine which distribution is best for simulating the random variable. See Section 5.7 for an explanation of =CDFDEV(). An example of the parameter estimation is provided in Parameter Estimation Demo.xls. The =CDFDEV() scalar in the example above indicates that the Beta distribution fits the data better than the other distributions tested.

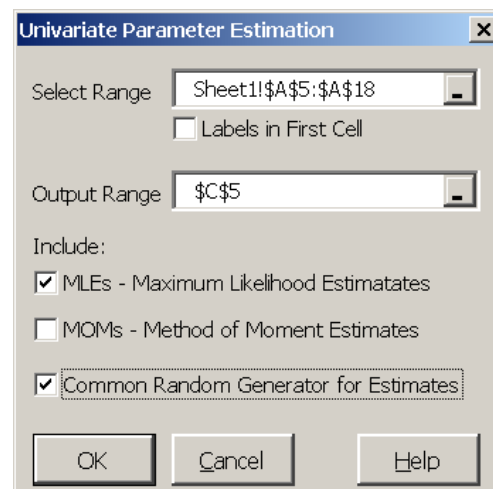



Figure 12. Univariate Parameter Estimator

	C	D	E	F	G	H	I	J	
5	Univariate Parameter Estimation for Wheat P at 12/18/2005 1:06:30 PM							Compare Simulated	
6			MLEs			Formulas	Results to History		
7	Distribution	Parameters	Parm. 1	Parm. 2	Distribution	MLE	CDFDEV	SCALER	
8	Beta	$\alpha, \beta; A \leq x \leq B, \alpha, \beta > 0$	0.739	1.004	Beta	3.206	1.962765		
9	Double Exponential	$\alpha, \beta; -\infty < x < \infty, -\infty < \alpha < \infty, \beta > 0$	3.260	0.513	Double Exponential	3.239	7.267778		
10	Exponential	$\mu, \sigma; -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$	2.420	0.874	Exponential	2.991	1.955898		
11	Gamma	$\alpha, \beta; 0 \leq x < \infty, \alpha, \beta > 0$	27.204	0.121	Gamma	3.222	3.335559		
12	Logistic	$\mu, \sigma; 0 \leq x < \infty, -\infty < \mu < \infty, \sigma > 0$	3.278	0.371	Logistic	3.248	5.38898		
13	Log-Log	$\mu, \sigma; -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$	2.990	0.530	Log-Log	3.154	2.718552		
14	Log-Logistic	$\mu, \sigma; 0 \leq x < \infty, -\infty < \mu < \infty, \sigma > 0$	8.493	3.226	Log-Logistic	3.196	3.863648		
15	Lognormal	$\mu, \sigma; 0 \leq x < \infty, -\infty < \mu < \infty, \sigma > 0$	1.174	0.192	Lognormal	3.203	3.096948		
16	Normal	$\mu, \sigma; -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$	3.294	0.639	Normal	3.262	4.147795		
17	Pareto	$\alpha, \beta; \alpha \leq x < \infty, \alpha, \beta > 0$	2.420	3.449	Pareto	2.925	1.958128		
18	Uniform	$a, b; a \leq x \leq b$	2.420	4.550	Uniform	3.442	1.939766		
19	Weibull	$\alpha, \beta; 0 \leq x < \infty, \alpha, \beta > 0$	5.480	1055.587	Weibull	3.296	5.300051		
20	Binomial	$n, p; x=0,1,2,\dots,n; 0 \leq p \leq 1$	4.000	0.712	Binomial	3.000	13.20388		
21	Geometric	$p; x=1,2,\dots; 0 \leq p \leq 1$	0.260		Geometric	3.000	7.140645		
22	Poisson	$\lambda; x=0,1,\dots; 0 \leq \lambda < \infty$	2.846		Poisson	3.000	13.20388		
23	Negative Binomial	$s, p; x=1,2,\dots; 0 \leq p \leq 1$			Negative Binomial				
24						Common USD			
25						0.48			

4.2 Empirical Probability Distributions

The parameters for an empirical probability distribution are estimated using a Simetar function activated by the  icon. The *Select Input Ranges* window indicates the data to be used for defining the probability distribution (Figure 13). Be sure to select the *Labels in First Cell* box when there is a name in the first cell (row or column) of the *Selected Input Ranges*. Four examples of using the Empirical Parameter estimation dialog box are provided in the Empirical Demo.xls workbook program.

The dialog box (Figure 13) allows the user to estimate the parameters for one empirical distribution or for numerous distributions at once. The only restriction for using this function is that all of the data series must have the same number of observations. The dialog box allows estimation of the parameters four different ways:

- Use actual data (no transformations) for the distribution,
- Convert the actual data to differences (residuals) from the mean prior to estimating the parameters,
- Convert the actual data to deviations (residuals divided by the mean) from the mean prior to estimating the parameters, and
- Convert the actual data to deviations (residuals divided by the trend values) from a linear trend line prior to estimating the parameters.

The empirical distribution parameter estimation output includes the random data (residuals from trend or mean),

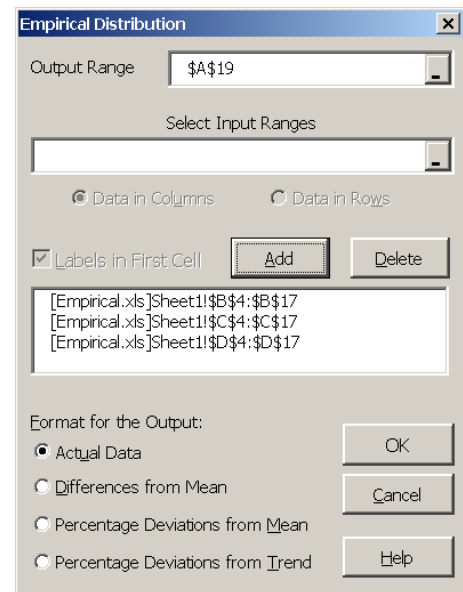



Figure 13. Parameter Estimation for the Empirical Distribution Dialog Box.

	A	B	C	D	E	F
56	Results for calculating the parameters for 3 Empirical distributions					
57	Sorted Deviations from Mean as a Percent of Mean					
58	F(x)	Corn P	Wheat P	Sorg P		
59	0	-0.35498	-0.26549	-0.46391		
60	0.0384615	-0.35495	-0.26547	-0.45386		
61	0.1153846	-0.16573	-0.21994	-0.33196		
62	0.1923077	-0.13563	-0.2078	-0.17104		
63	0.2692308	-0.10982	-0.2078	-0.11253		
64	0.3461539	-0.02812	-0.08942	-0.08815		
65	0.4230769	-0.01952	-0.01658	0.024006		
66	0.5	0.014886	-0.01051	0.033758		
67	0.5769231	0.019186	0.025916	0.072768		
68	0.6538461	0.053589	0.047163	0.097149		
69	0.7307692	0.075091	0.129115	0.106902		
70	0.8076923	0.092292	0.129115	0.126407		
71	0.8846154	0.165399	0.30516	0.141035		
72	0.9615384	0.393318	0.381041	0.555514		
73	1	0.393357	0.381079	0.555569		
74						
75	Simulate Random Values for 3 Empirical Distributions					
76	Stoch Dev	0.018298	-0.01615	-0.35955	=emp(D59:D73,A59:A73)	
77	Means	2.33	3.29	2.05	=D35	
78	Stoch Value	2.367935	3.241399	1.313415	=D77+D77*D76	

summary statistics, correlation matrix if more than one variable is specified, and the sorted random values with cumulative distribution probabilities. The sorted deviations required by =EMP() to simulate an empirical distribution are demonstrated in the example output to the right. Once the Empirical distribution parameters are estimated, they can be simulated using the =EMP (S_i, F(S_i)) function (see Section 3.7.1).

4.3 Multivariate Probability Distributions

Correlation and covariance matrices are both parameters for multivariate probability distributions. Correlation and covariance matrices can be calculated using the Correlation Matrix dialog box activated by the  icon. This Simetar dialog box calculates the upper right triangle correlation matrix of size NxN when the user specifies N variables. The first step to using the dialog box is to specify the location for placing the upper left hand corner of the generated correlation matrix by indicating the *Output Range* in the menu (Figure 14). Next, specify whether the data to correlate are in columns or rows. The first cell of each column (or row) indicated in the *Selected Arrays* box should have a label so the output matrix is easier to read.

The Correlation Matrix dialog box calculates either the Pearson's (standard) correlation coefficient matrix or the rank correlation matrix. The default is the Pearson's correlation coefficient matrix. The rank correlation coefficient matrix is calculated when the *Rank Correlation* radio button is selected.

The statistical significance of each correlation coefficient can be tested by Simetar. Student's-t values for the correlation coefficients greater than the t-critical value indicate whether the correlation coefficient is statistically different from zero and are displayed in bold. See Complete Correlation Demo.xls for examples of using the correlation matrix dialog box.

A covariance matrix can be calculated using the Correlation Matrix dialog box (Figure 14). The upper triangle covariance matrix is calculated by selecting the *Covariance Matrix* radio button after specifying the arrays to include in the matrix. The *Full Symmetric* covariance matrix is calculated by selecting this option in the dialog box and the *Covariance Matrix*. See the demonstration program Complete Correlation Demo.xls for examples of estimating covariance matrices.

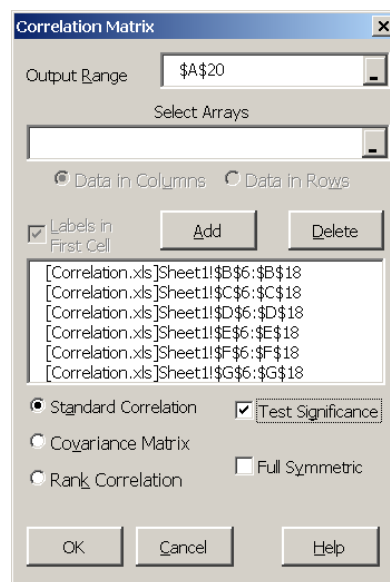


Figure 14. Correlation Matrix Dialog Box.


	A	B	C	D	E	F	G
20	Correlation Matrix						
21		Corn P	Wheat P	Sorg P	Corn Y	Wheat Y	Sorg Y
22	Corn P	1	0.87	0.98	-0.30	-0.18	-0.53
23	Wheat P		1	0.82	-0.16	-0.37	-0.34
24	Sorg P			1	-0.29	-0.17	-0.56
25	Corn Y				1	0.53	0.59
26	Wheat Y					1	0.47
27	Sorg Y						1
28							
29	Correlation Coefficient t-values.						
30	Significance	95%		t-critical	2.20		
31		Corn P	Wheat P	Sorg P	Corn Y	Wheat Y	Sorg Y
32	Corn P		5.83	16.79	1.04	0.60	2.08
33	Wheat P			4.68	0.54	1.31	1.21
34	Sorg P				1.00	0.58	2.27
35	Corn Y					2.08	2.40
36	Wheat Y						1.78
37	Sorg Y						
38	Bold values indicate statistical significance at the specified level.						
39							
40	Covariance Matrix						
41		Corn P	Wheat P	Sorg P	Corn Y	Wheat Y	Sorg Y
42	Corn P	0.180533	0.222644	0.160682	-1.68098	-0.20039	-1.20096
43	Wheat P		0.408639	0.213254	-1.42204	-0.66108	-1.24053
44	Sorg P			0.16711	-1.64783	-0.1966	-1.30985
45	Corn Y				194.9708	20.82524	46.47593
46	Wheat Y					7.887421	7.534503
47	Sorg Y						32.17119

4.4 GRKS Probability Distribution

Parameters for the GRKS probability distribution can be estimated using the dialog box in Figure 15. The GRKS distribution dialog box is accessed via the toolbar Simetar drop down menu – GRKS Distribution. The GRKS pdf is defined by three values: *Minimum*, *Middle Value*, and *Maximum*. Simetar places the parameters on the worksheet starting in the designated *Output Range*. The parameters are presented as values and their associated probabilities (see GRKS Distribution Demo.xls). Simetar also generates a chart of the distribution and that displays how the shape of the distribution changes as the minimum, middle, and maximum values change. Test this feature by changing the three parameters and observing their affects on the GRKS distribution figure. The GRKS pdf parameters can be simulated using the =GRKS() function in Section 3.5.2.

Figure 15. Parameter Estimation for the GRK Distribution.

5.0 Statistical Tests for Model Validation

Model validation must be done prior to application of a simulation model for decision making. Validation can utilize graphs, such as PDFs and CDFs, but statistical testing of the simulated distributions is required to determine whether the stochastic variables in the model are statistically from the same distribution as the historical data. To facilitate the validation process several hypothesis tests have been included in Simetar. The tests are organized using 5 tabs in the Hypothesis Testing for Data dialog box opened by the  icon (Figure 16). Examples of the validation tests described in this section are available in Hypothesis Tests Demo.xls.

5.1 Univariate Distribution Tests for Model Validation

The means and variances for two distributions (or series) can be compared by using the *Compare Two Series* tab for the Hypothesis Testing dialog box (Figure 16). The mean and variance tests are univariate as they only test the difference between two variables. This type of hypothesis testing is useful in validation for comparing the simulated distribution to the historical distribution. The null hypotheses are that the simulated mean equals the historical mean and the simulated variance equals the historical variance. As demonstrated in the example below, it is useful to statistically test if the simulated data have the same mean and variance as the historical data series.

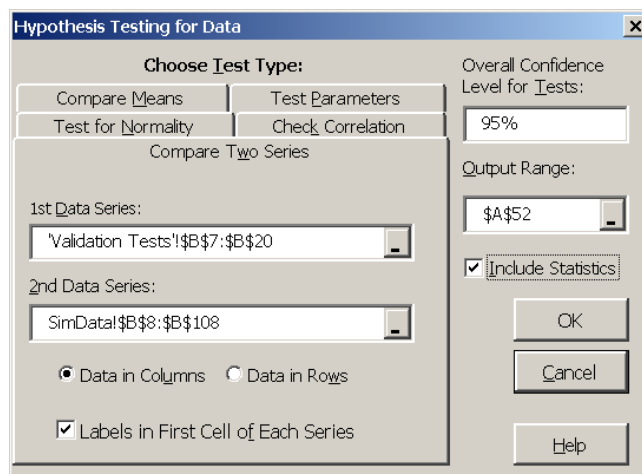


Figure 16. Univariate and Multivariate Distribution Tests.

The statistical tests are performed when the *Compare Two Series* tab in Figure 16 is selected and you specify the two distributions (data series) to compare. A two sample, Student-t test is used to allow comparison of means from distributions with an un-equal number of observations (see example below). See Step 4 in Hypothesis Tests Demo.xls for an example of comparing two distributions.

	A	B	C	D	E	F	G	H
52	Step 3. Compare Two Distributions Using Student's-t and F Tests Compare Historical Corn P with the Simulated Corn P							
53	Distribution Comparison of Corn P & Sim Corn P							
54	Confidence Level		95.0000%					
55		Test Value	Critical Value	P-Value				
56	2 Sample t Test	-0.02	2.49	0.982	Fail to Reject the Ho that the Means are Equal			
57	F Test	1.03	1.85	0.428	Fail to Reject the Ho that the Variances are Equal			

5.2 Multivariate Distribution Tests for Model Validation

Means and variances for multivariate (MV) probability distributions can be statistically tested against the distribution's historical data in one step by selecting the *Compare Two Series* tab in the Hypothesis Testing for Data dialog box and specifying matrices as the input Figure 17.

The first MV test uses the two-sample Hotelling T^2 test which tests whether two data matrices (the historical data $M \times N$ and the simulation results $P \times N$) statistically have equivalent mean vectors and covariance matrices. Assume historical data are arranged in an $M \times N$ matrix and the simulated data are in a $P \times N$ matrix, where P is the number of iterations, then the means can be tested with the Hotelling T^2 test procedure. The Hotelling T^2 test is analogous to a Student's-t test of two means in a two-sample univariate case.

The second MV test calculated for this statistical test, Box's M , tests the equality of the covariance matrices with dimensions $M \times M$ and $P \times N$, respectively, using a large sample likelihood ratio testing procedure. The Box's M test of homogeneity of covariances is used to test whether the covariance matrices of two or more data series, with n columns each, are equal. The assumptions under this test are that the data matrices are MV normal and that the sample is large enough for the asymptotic, or central Chi-Squared, distribution under the null hypothesis to be used.

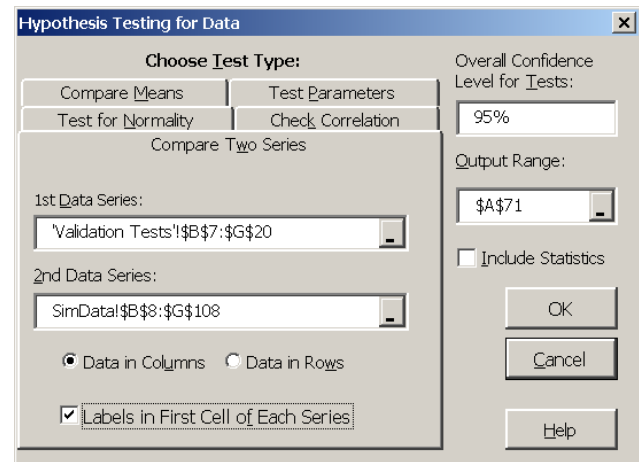


Figure 17. Multivariate Hypothesis Tests for Six Variables.

	A	B	C	D	E	F	G	H	I	J	K	L
69	Step 4. Compare Simulated Results to the Input Distribution											
70	Distribution Comparison of Two Data Series											
71	Distribution Comparison of Corn P & Sim Corn P											
72	Confidence Level		95.0000%									
73		Test Value	Critical Value	P-Value								
74	2 Sample Hotelling T^2 Test	0.00	13.73	1.000	Fail to Reject the Ho that the Mean Vectors are Equal							
75	Box's M Test	1.99	32.67	1.000	Fail to Reject the Ho that the Covariance Matrices are Equivalent*							
76	Complete Homogeneity Test	2.12	40.11	1.000	Fail to Reject the Ho that the Mean Vectors and Covariance Matrices are Equivalent, Respectively*							
77					*Based on asymptotic distribution							

The third MV test is the Complete Homogeneity test. This statistical test simultaneously tests the mean vectors and the covariance matrices for two distributions. The historical data's mean vector and covariance matrix are tested against the simulated sample's mean vector and covariance matrix. If the test fails to reject that the means and covariance are statistically equal, then one can assume that the multivariate distribution in the historical series is being simulated appropriately. An example of this test is provided above and in Step 4 of Hypothesis Tests Demo.xls.

5.3 Test Correlation

Another multivariate distribution validation test in Simetar is a test to compare the correlation matrix implicit in the simulated output to the input (assumed) correlation matrix. This test is useful for validating multivariate probability distributions, particularly the non-normal multivariate distributions. Selecting the *Check Correlation*

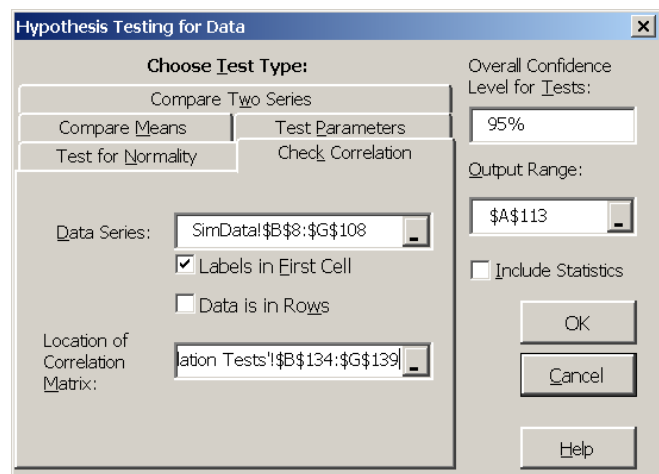


Figure 18. Test Correlation of MV Distribution Simulation Results.

tab in the Hypothesis Testing for Data dialog box (Figure 18) calculates the Student's-t test statistics for comparing the corresponding correlation coefficients in two matrices. The dialog box requires information for the location of the simulated series (a PxN matrix) and the location of the NxN correlation matrix used to simulate the multivariate distribution (or the correlation matrix implicit in the historical data for the distribution).

The confidence level for the resulting Student's-t test defaults to a value greater than 0.95 but can be changed by the user after the test has been performed. An example of this test is provided in Step 6 of Hypothesis Tests Demo.xls. If a correlation coefficient for two simulated variables is statistically different from the respective historical correlation coefficient, the Student's t-test statistic will exceed the Critical Value and its respective statistic will be displayed as a bold value. If the test shows several bold values check the formulas used to simulate the multivariate distribution to insure the distribution is modeled correctly.

	A	B	C	D	E	F
112	Step 6. Compare Simulated Correlation Matrix to Input Correlation Matrix					
113	Test Correlation Coefficients					
114	Confidence Level		99.6586%			
115	Critical Value		3.00			
116						
117		Sim Wheat P	Sim Sorg P	Sim Corn Y	Sim Wheat Y	Sim Sorg Y
118	Sim Corn P	0.31	0.03	0.39	0.79	0.21
119	Sim Wheat P		0.42	0.50	0.48	0.57
120	Sim Sorg P			0.38	0.94	0.28
121	Sim Corn Y				0.67	1.19
122	Sim Wheat Y					0.57


5.4 Test Mean and Standard Deviation

The mean and standard deviation for any data series (e.g., simulated data) can be compared to a specified mean and standard deviation using the *Test Parameters* tab in Figure 19. The Student's-t test is used to compare the user specified mean to the observed mean of any distribution (or series) as demonstrated in Figure 19. A Chi-Squared test is used to test a user specified standard deviation against the standard deviation for any distribution. The null hypothesis is that the statistic for the series equals the user's specified values. An example of testing the historical data for a variable against a specified mean and standard deviation is provided below and in Hypothesis Tests Demo.xls.

Figure 19. Test Mean and Standard Deviation for a Univariate Distribution.

	A	B	C	D	E	F	G	H	I	J	K
87	Step 5. Compare Simulated Results to an Assumed Mean and Standard Deviation										
88	Assumed Mean	2.325									
89	Assumed Std Dev	0.400									
90	Test of Hypothesis for Parameters for Corn P										
91	Confidence Level		95.0000%								
92		Given Value	Test Value	Critical Value	P-Value						
93	t-Test	2.325	0.00	2.58	1.00	Fail to Reject the Ho that the Mean is Equal to 2.325					
94	Chi-Square Test	0.400	13.01	LB: 4.40	0.74	Fail to Reject the Ho that the Standard Deviation is Equal to 0.4					
95				UB: 23.34							

5.5 Univariate Tests for Normality

Five different tests for normality can be performed by selecting the *Test for Normality* tab after the  icon is selected (Figure 20). The normality tests are: Kolmogorov-Smirnoff, Chi-Squared, Cramer-von Mises, Anderson-Darling, and Shapiro-Wilks. The Chi-Squared test requires the number of bins (or intervals); 20 or more intervals appear to work for most data series. In addition to the normality tests this option calculates the skewness and kurtosis, relative to a normal distribution (not shown in the example below). See an example of these normality tests in Hypothesis Tests Demo.xls.

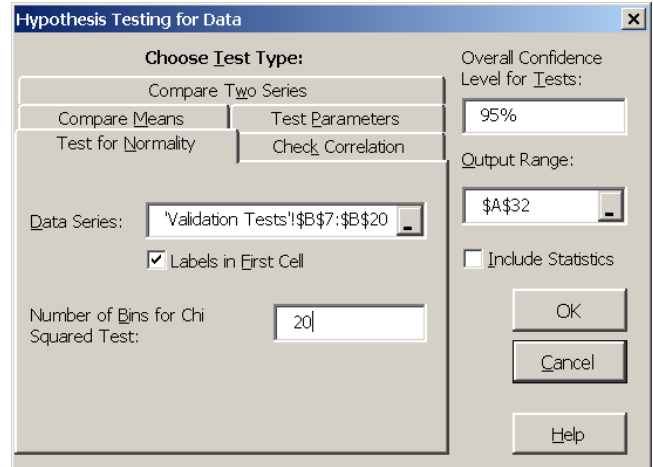


Figure 20. Univariate Normality Test.

	A	B	C	D	E	F	G	H
31	Step 2. Tests for Normality							
32	Test for Normality of Distribution for Corn P							
33	Confidence Level							95.0000%
34	Procedure	Test Value	p-Value					
35	S-W	0.962956115	0.798551994	Fail to Reject the Ho that the Distribution is Normally Distributed*				
36	A-D	0.303060068	0.523780533	Fail to Reject the Ho that the Distribution is Normally Distributed*				
37	CvM	0.04591582	0.544799272	Fail to Reject the Ho that the Distribution is Normally Distributed*				
38	K-S	0.149172142	NA	Consult Critical Value Table				
39	Chi-Squared	13.61538462	0.805633117	Fail to Reject the Ho that the Distribution is Normally Distributed*				
40				*Based on approximate p-values				

5.6 Multivariate Tests for Normality

A multivariate distribution test for normality can be performed on any data matrix of PxN. The MV normality test can be performed by specifying a PxN matrix in the *Data Series* box for the *Test for Normality* tab in the Hypothesis Testing dialog box (Figure 21). The MV normality tests are: skewness criterion, kurtosis criterion, and Chi-Squared quantile correlation. Simetar reports the test statistics, critical value, and p-value for the first two tests and the test statistic for the third test. The null hypothesis is that the data matrix is distributed MV normal. See the example output for this test below and in Hypothesis Tests Demo.xls.

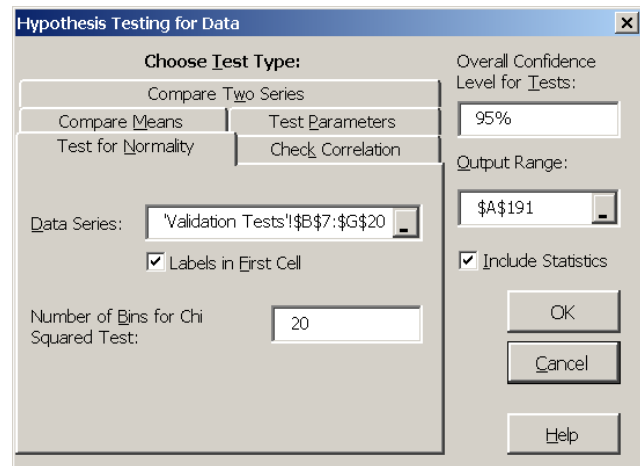


Figure 21. Multivariate Normality Tests Dialog Box.

	A	B	C	D	E	F	G	H	I	J
190	Step 9. Multivariate Test for Normality									
191	Test for Normality of a Multivariate Distribution									
192	Confidence Level									
193	Procedure	Test Value	Critical Value	p-Value						
194	Skewness Criterion	43.65956655	74.4683241	0.88502126	Fail to Reject the Ho that the Data Are Multivariate Normally Distributed*					
195	Kurtosis Criterion	-1.196897667	1.959963985	0.115673222	Fail to Reject the Ho that the Data Are Multivariate Normally Distributed*					
196	Chi-Squared Quantile Correlation	0.940571025			*Based on asymptotic p-values					

5.7 Compare Means (ANOVA)

The Hypothesis Testing for Data dialog box includes a means test (ANOVA) capability (Figure 22). Selecting the *Compare Means* tab in the Hypothesis Testing for Data dialog box produces a menu for specifying the two series to compare. For this test the user must specify the two distributions (or series) using the *Select Data Series* to compare window and the A button to list the series in the window at the bottom. The confidence level defaults to 0.95 and must be specified before clicking the *OK* button. The results of the ANOVA test are the sum of squares, mean square error, F-statistic and its p-value. A sample ANOVA test is demonstrated below and is provided in Hypothesis Tests Demo.xls.

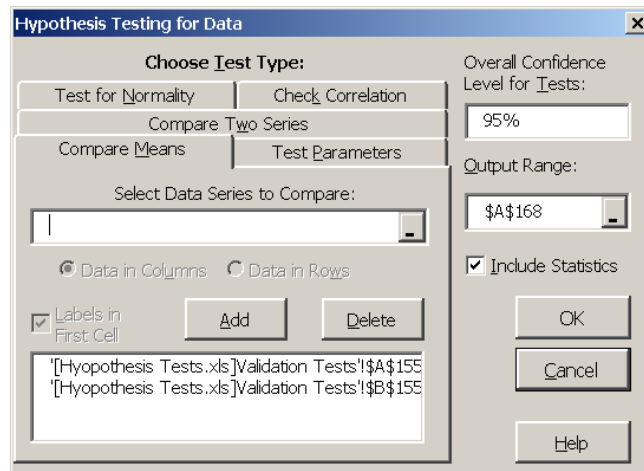


Figure 22. Compare Means Test Dialog Box

	A	B	C	D	E	F
167	Step 8. Compare Means for Two Series in other words, an ANOVA test					
168	Tests to Compare Multiple Means					
169	ANOVA Table					
170	Source of	Sum of	Degrees of	Mean		
171	Variation	Squares	Freedom	Square	F statistic	p-value
172	Treatments		6	1	6	19.51642633
173	Error	6.763533333		22	0.307433333	
174	Total	12.76353333		23		0.0002175

5.8 Compare Two Cumulative Distribution Functions (CDFs)

A scalar measure to compare the difference between two cumulative distribution functions (CDFs) is calculated by the =CDFDEV() Simetar function. The function calculates the sum of the squared differences between two CDFs with an added penalty for differences in the tails. The scalar is calculated for two CDFs, F(x) and G(x) as:

$$\text{CDFDEV} = \sum_{i=1}^N (F(x_{(i)}) - G(x_{(i)}))^2 + w_i$$

where: w_i is a penalty function that applies more weight to deviations in the tails than values around the mean.

If the G(x) distribution is the same as the F(x) distribution, then the CDFDEV value equals zero. The CDFDEV measure is programmed to compare a historical series Nx1 to a simulated series Px1 as follows:

$$=\text{CDFDEV}(\text{Range for Historical Series}, \text{Range for Simulated Series})$$

where: Range for Historical Series is the location for the historical data, such as B1:B10, and Range for Simulated Series is the location for the simulated values, such as B9:B109.


The =CDFDEV() function is useful when testing the ability of different assumed probability distributions to simulate a random variable. In this case, the =CDFDEV() measure is calculated using the simulated values for each of the alternative probability distributions. The probability distribution associated with the lowest =CDFDEV() scalar is the “best” distribution for simulating the random variable. See Parameter Estimation Demo.xls for an example.

Compare Ability of Alternative Distributions to Fit a Series		
Distributions Tested	CDFDEV()	Formula
Beta	0.0163	=cdfdev(Sheet1!\$A\$6:\$A\$18, SimData!B9:B108)
Double Exponential	2.2610	=cdfdev(Sheet1!\$A\$6:\$A\$18, SimData!C9:C108)
Exponential	1.1321	=cdfdev(Sheet1!\$A\$6:\$A\$18, SimData!D9:D108)
Gamma	0.2233	=cdfdev(Sheet1!\$A\$6:\$A\$18, SimData!E9:E108)
Logistic	1.2496	=cdfdev(Sheet1!\$A\$6:\$A\$18, SimData!F9:F108)
Log-Log	0.3239	=cdfdev(Sheet1!\$A\$6:\$A\$18, SimData!G9:G108)

6.0 Graphical Tools for Analyzing Simulation Results

Simetar provides nine graphics tools for displaying the results of stochastic simulations and for analysis of data. These graphics tools utilize the charting capabilities of Excel so all charts and graphs can be edited and enhanced using standard Excel charting tools. Simetar charts and graphs are developed using menus which allow the user to easily specify the data, titles, and labels for charts that are used frequently for simulation. An example of Simetar’s charts is provided in Charts Demo.xls.

6.1 Line Graph

Any series of numbers can be graphed on an X-Y axis as a line graph using this option. The icon to access line graphs is . The Line Graph menu (Figure 23) requires that you specify the values for the X axis (such as, years) and the Y values (such as, prices) in the *X and Y-Range* boxes. Labels for these variables are optional and are entered in the *Y and X-Axis Label* boxes. The *Chart Title* is optional. You may include a label in the first cells (row or column) indicated for each Y variable, if you select the box for *Series Labels in First Cell*.

The chart can have more than one line by using the *Add Y’s* button and indicating multiple Y series in the *Select Y-Axis Range*, one at a time or all at once if the variables are contiguous. Once the graph is drawn by Excel, it can be edited using Excel chart commands.

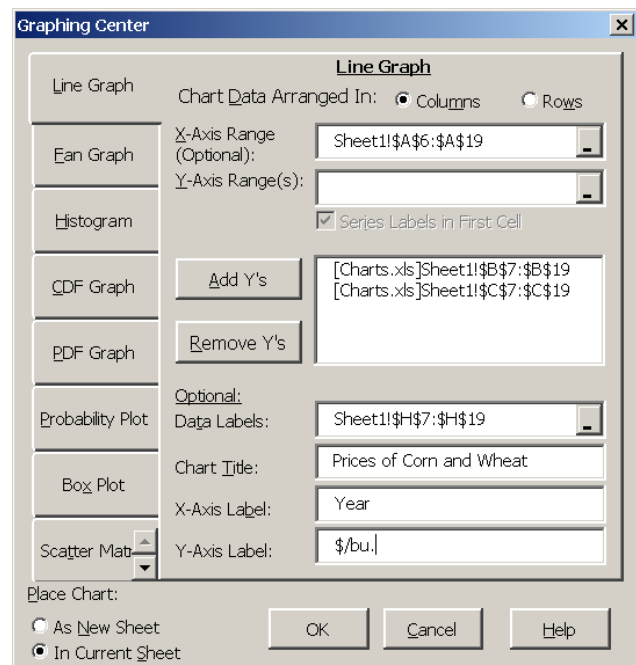
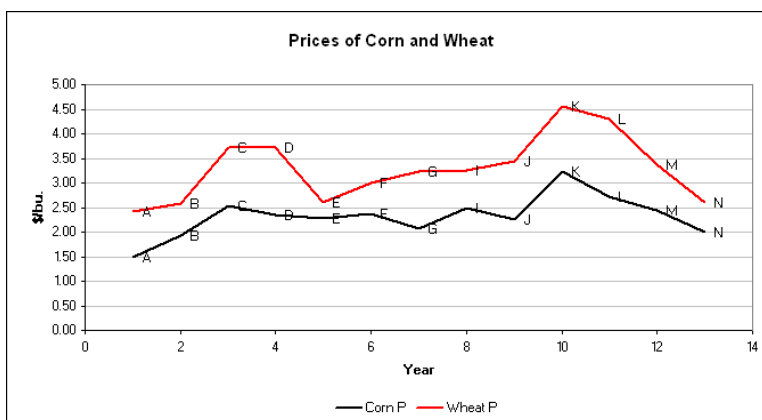



Figure 23. Dialog Box for Developing a Line Chart.

The Line Graph dialog box allows the user to label the points on line graphs. For example, a price/quantity chart can be developed with year labels on the individual data points to show

years when structural changes took place. To use this option indicate the column or row of labels in the *Data Labels* box, being sure to have the same number of labels as there are rows (or columns) of data to graph. The result of the chart specified in Figure 18 is presented below and in Charts Demo.xls.



6.2 CDF Graph

Cumulative distribution function (CDF) charts of individual or multiple variables (simulated values) can be developed using Simetar. CDF graphs are initiated by selecting the  icon. Identify the variables to graph by highlighting the column(s), after first clicking in the *Select Range to Graph* box (Figure 24). Include names in the first cell of the variable range, so the chart will include names for the individual lines. (Be sure the variable names begin with a letter.) The chart can be placed on the current worksheet or in a new chart sheet. Use Excel's chart commands to format the scale for the X axis and to make changes to the title.

CDF graphs developed using Simetar are dynamic so when the values referenced for the chart change, the CDF graph is automatically updated by Excel. This feature is particularly useful for simulation. Each time the simulation results are updated in SimData, the CDF graphs will be updated.

The *Smoothing* option in the CDF menu utilizes kernel density functions to smooth the observed values and develop smoothed CDF charts. In addition to the CDF charts, the output for this option includes a text box with a drop down menu to allow the user to select the kernel. The default kernel is the Gaussian, but ten more are provided. The kernel smoothed CDF for a historical series depicts the probability distribution Simetar would use if the series was simulated using =KDEINV(), see Section 3.7.4.

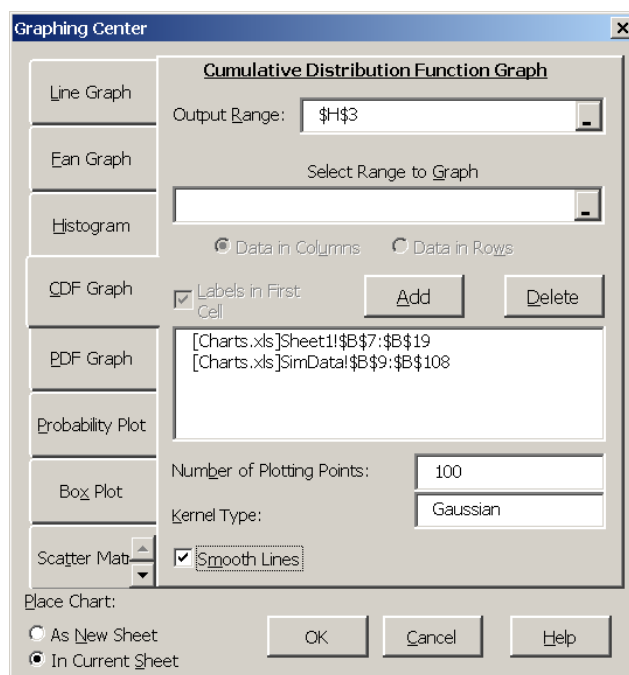
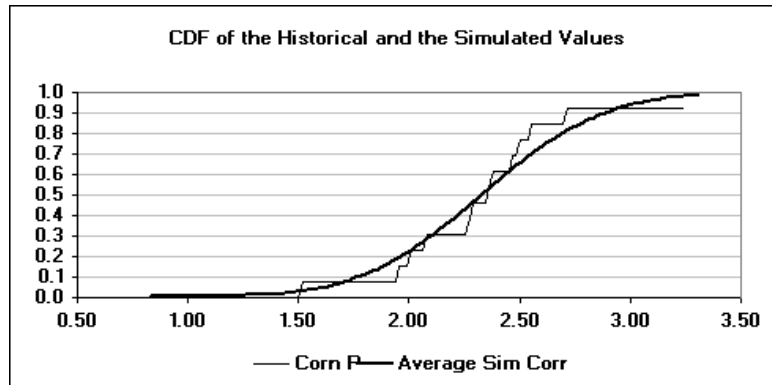



Figure 24. CDF Chart Dialog Box.

The CDF graph option is useful for comparing simulated values of a random variable to the variable's historical data. This is possible in Simetar even though the two series have a different number of observations. See the example below and in Charts Demo.xls.



6.3 PDF Graph

Probability distribution function (PDF) graphs of individual or multiple variables can be estimated using the  icon. Identify the variables to include in the PDF graph by selecting the variables in the *Select Range to Graph* box and the *Add* button if the variables are not in continuous columns (or rows) (Figure 25). The PDF graph function uses kernel estimators to smooth the data rather than just using line segments to connect the dots. Eleven kernels are available to develop the PDF graphs: Gaussian, Cauchy, Cosinus, Double Exp., Epanechnikov, Histogram, Parzen, Quartic, Triangle, Triweight, and Uniform. Once the graph is drawn you can change the kernel by editing the output range in the worksheet.

If the data series have names in the first cell indicate this on the menu, otherwise unselect the *Labels in First Cell* option. Multiple PDFs can appear on the same axis so the simulated values and their historical values can both be graphed on the same axis. This feature is possible because the data series being graphed do not have to be the same lengths.

PDF graphs developed using Simetar are dynamic so when the values in the *Selected Range to Graph*, change the graph is instantly updated. This feature is useful when displaying simulation results using PDFs. The mean of the variables in a PDF is included in the chart. Confidence intervals at the alpha equal 5 percent level can be added by selecting the *Plot Quantiles*. The quantiles can be redrawn by changing the Alpha equal 0.9 to 0.10 in the seventh row of the PDF Graph output table. The title can be changed by editing the first line of the PDF Graph output.

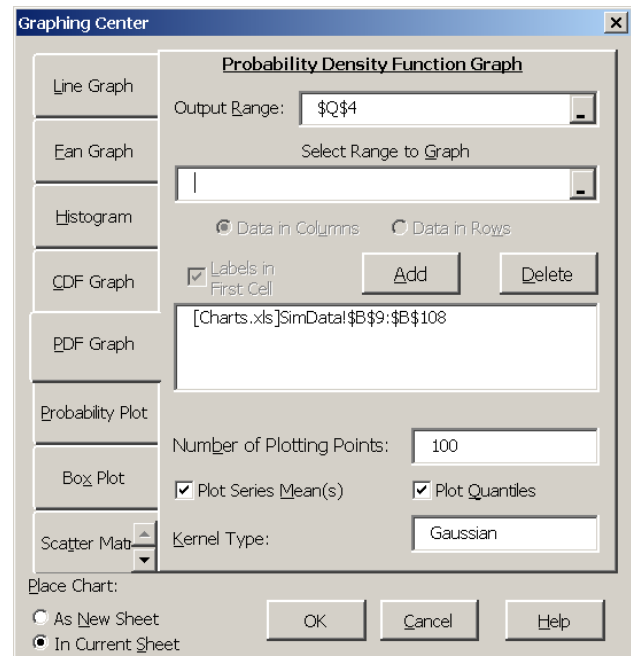
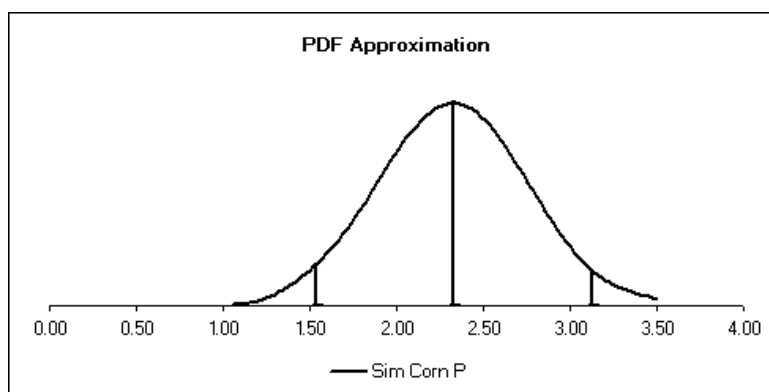



Figure 25. PDF Chart Dialog Box.

See the example below of a PDF chart developed for a simulated series in Charts Demo.xls.



6.4 Histograms

Histograms of individual variables (simulated output) can be developed using the Simetar menu. The histogram icon  activates this option. Indicate the variable to graph by clicking the *Select Range to Graph* box in the dialog box (Figure 26) and highlighting the variable in the worksheet. Specify the *Number of Bins* (intervals) and select *OK*. The more bins the smoother the histogram. The maximum number of bins is the number of observations minus one. Experiment with the number of bins to find the number which best suits the data. An added feature of the histogram option in Simetar is to display data as a cumulative distribution with the bins growing in height from zero to one as the X value gets large.

The screenshot shows the "Graphing Center" dialog box with the "Histogram" tab selected. The "Output Range" is set to "\$Z\$4". The "Select Range to Graph" is set to "SimData!\$B\$8:\$B\$108". The "Number of Bins" is set to 25. The "Histogram Title" is "Histogram of Simulated Corn Prices". The "X-Axis Label" is "Simulated Corn Price". The "Chart Type" is set to "PDF". The "Place Chart" options are "As New Sheet" and "In Current Sheet", with "In Current Sheet" selected. Buttons for "OK", "Cancel", and "Help" are visible at the bottom.

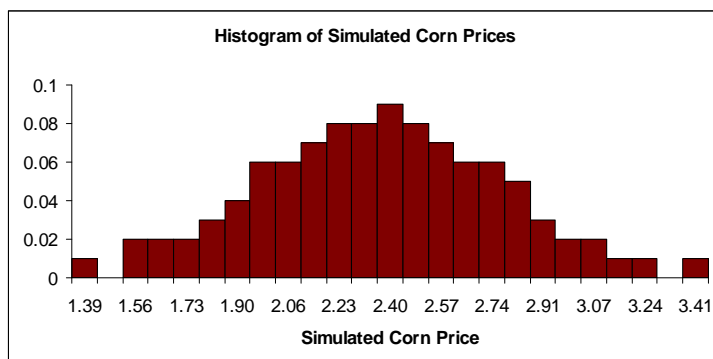



Figure 26. Histogram Dialog Box.

6.5 Fan Graph

A Fan Graph consists of multiple lines in the Y axis for multiple scenarios (or multiple years of one variable) graphed in the X axis. The variables graphed in the X axis can be successive years for a simulated output variable. Alternatively, the variables on the X axis can be the same simulated variable but for different scenarios. The purpose of a Fan Graph is to show the effect of risk on a variable over time or across scenarios.

A Fan Graph showing the simulated mean and percentiles or confidence interval lines about the mean can be developed using the  icon in Simetar. The range of variables to be graphed on the X axis must be specified in the *Select Ranges to Graph* box (Figure 27). The variables (scenarios or years) must be specified in the order they appear in the graph. For example, if the graph is for 10 years of a probabilistic forecast, specify the 10 variables across the, say, 500 iterations as the selected range to graph. If the variables are not contiguous, they can be specified one at a time using the *Add* box. The Fan Graph dialog box (Figure 27) provides boxes to specify up to six percentile or confidence lines about the mean. The individual lines to add to the Fan Graph must be specified as fractions, such as 0.05 and 0.95 would result in a graph with 3 lines: the mean, the 5 percentile and the 95 percentile lines. Once the Fan Graph has been developed, you can dynamically change the graph by editing the percentile values in the output table. For example, if the 5% and 95% lines need to be changed to 1% and 99%, simply change the 0.05 to 0.01 and the 0.95 to 0.99 in the Fan Graph output table. Changing the percentile causes Excel to re-draw the graph. An example of a fan graph developed to show the relative risk between three distributions is provided below.

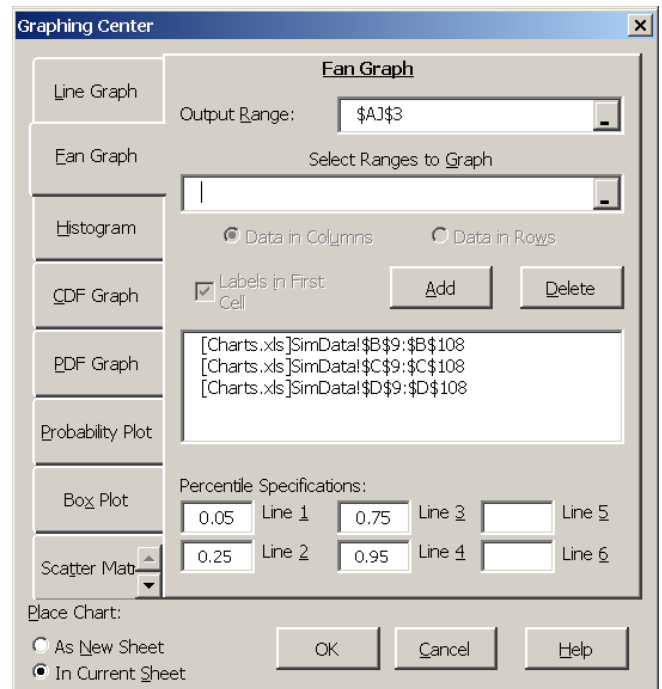
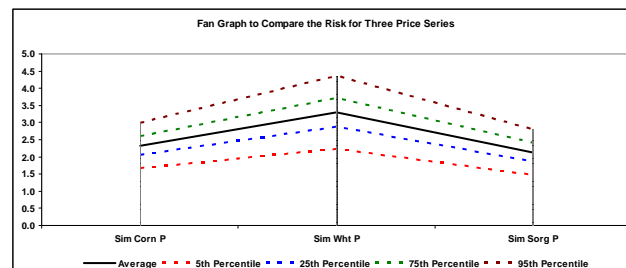

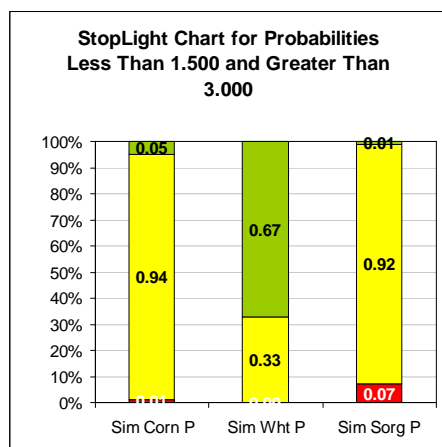


Figure 27. Fan Graph Dialog Box.




6.6 StopLight Chart

The StopLight chart compares the target probabilities for one or more risky alternatives and is activated by selecting the  icon. The user must specify two probability targets (*Lower Target* and an *Upper Target*) for the StopLight and the alternative scenarios to compare (Figure 28). The StopLight function calculates the probabilities of: (a) exceeding the upper target (green), (b) being less than the lower target (red), and (c) observing values between the targets (yellow). An example is provided below.



6.7 Probability Plots

Three types of probability plots can be generated by selecting the probability plot icon . The probability plot function develops Normal Probability (or NP), Quantile–Quantile (or Q–Q) Plots and Probability–Probability (or P–P) Plots (Figure 29). See Charts Demo.xls for an example of all three types of probability plots.

The Normal Plot is a method for checking how close to normal a random variable is distributed. A Normal Plot compares the ordered data to the standard normal distribution's percentiles. If a variable is normally distributed the sorted data values will be entirely on a straight line with the only deviations from the line due to sampling error.

A Quantile-Quantile (Q-Q) Plot can be used to compare two distributions. If the two random variables have the same distribution, their paired observations lie on a 45° line. If the two random

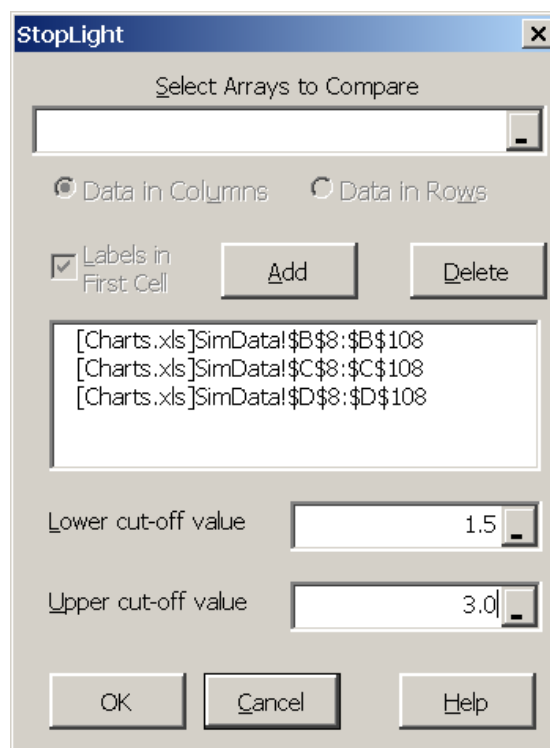


Figure 28. StopLight Dialog Box.

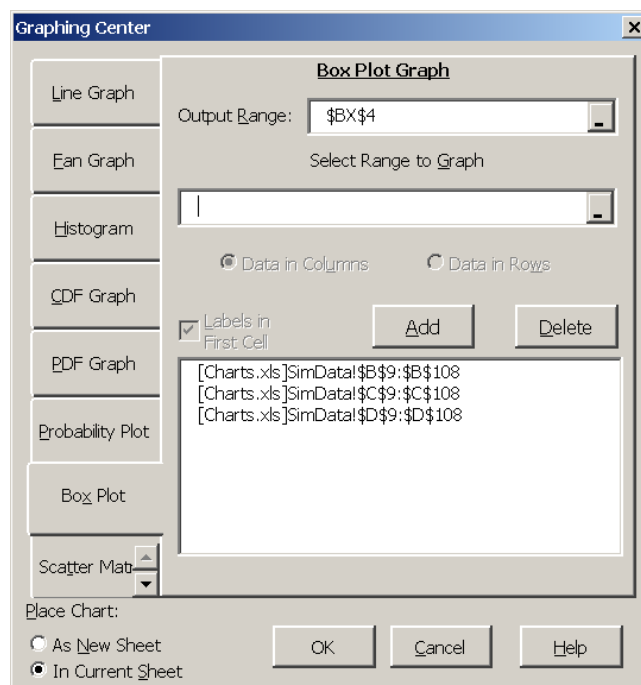
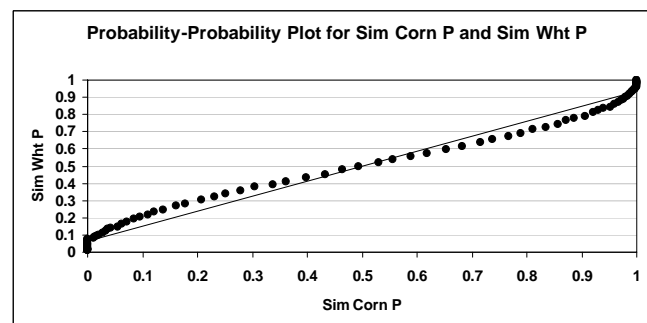



Figure 30. Box Plot Dialog Box.



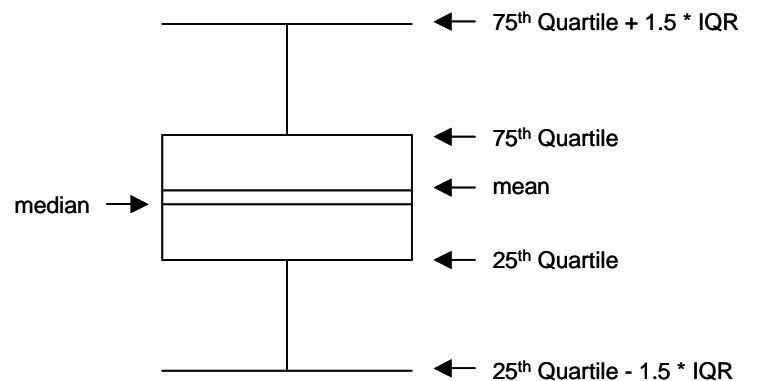
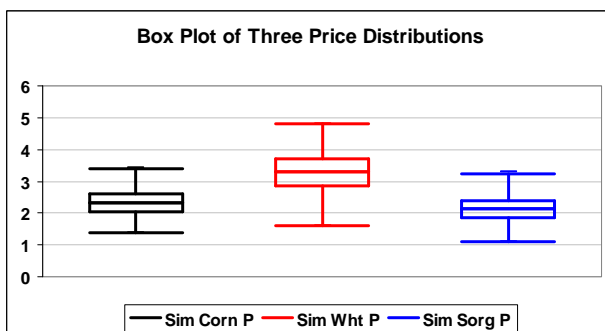
variables are in the same family of distributions, their paired observations tend to be linear although they may not lie on the 45° line. A P-P Plot consists of a graph of the percentiles for the sorted values of two variables graphed on one axis. If the two random variables have the same distribution (shape) the observations for a P-P Plot will be on a 45° line.

6.8 Box Plots


Box plots of one or more variables can be prepared by selecting the  icon. The Box Plot dialog box (Figure 30) indicates the information required for this function. The Box Plot is a quartile summary of a random variable in graphical form that indicates whether a variable is skewed to the left or right. The names and values of the Box Plot are best defined in a chart:

where: $IQR = [75^{\text{th}} \text{ Quartile} - 25^{\text{th}} \text{ Quartile}]$

Fifty percent of the observed values fall within the box (25th to 75th quartile). If the distribution is skewed to the right then the bottom line segment is longer than the top line segment, and vice versa if the distribution is skewed left. Values that lie outside the extreme lines are likely to be outliers. The median and mean will show up as one line for symmetrical distributions.



6.9 Scatter Matrix Graph

A scatter matrix of multiple univariate data series can be created using the scatter matrix icon  (Figure 31). The scatter matrix is an array of individual graphs of several univariate data series. Each series is plotted against each of the other series, one at a time, like a correlation matrix (see the example below). The graphs show the linear relationships between individual series and can be useful in multiple regression to determine collinearity and for identifying linear relationships between variables for a multivariate probability distribution. See Charts Demo.xls for an example of a scatter matrix.

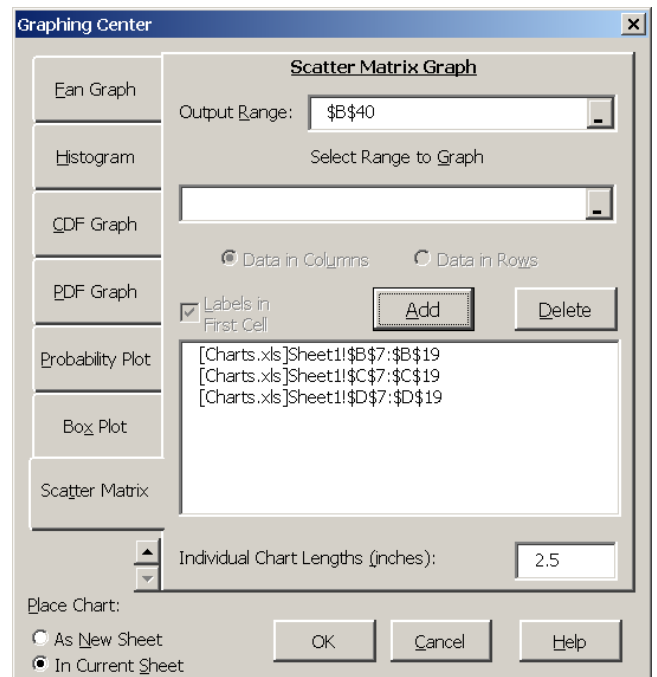
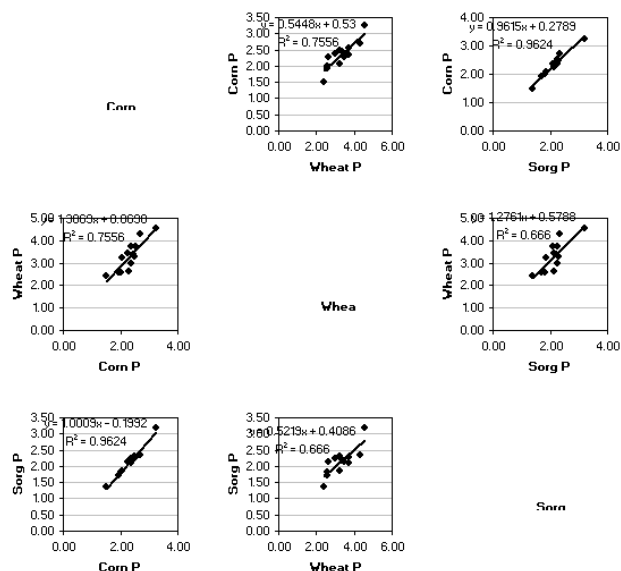


Figure 31. Scatter Matrix Dialog Box.



7.0 Scenario Analysis

Simulation models are most useful when used to simulate alternative scenarios. Scenario analysis involves specifying different values for several exogenous or management control variables and simulating the model for the different scenarios. The Simetar Simulation Engine dialog box (Figure 32) provides an input field for entering the *Number of Scenarios*. When the number of scenarios exceeds 1, Simetar executes the =SCENARIO() functions in the model.

A separate =SCENARIO() function must be specified for each variable to be systematically changed for the alternative scenarios. The =SCENARIO() function specifies the values the variable can take on for each scenario. For example, simulating three input variables for example, Hours Product; for five scenarios (see the example below) is programmed using three =SCENARIO() functions as follows:

In B21:D25 enter the values for 3 variables for the exogenous variables
 In B27 = SCENARIO (B21:B25)
 In C27 = SCENARIO (C21:C25)
 In D27 = SCENARIO (D21:D25)

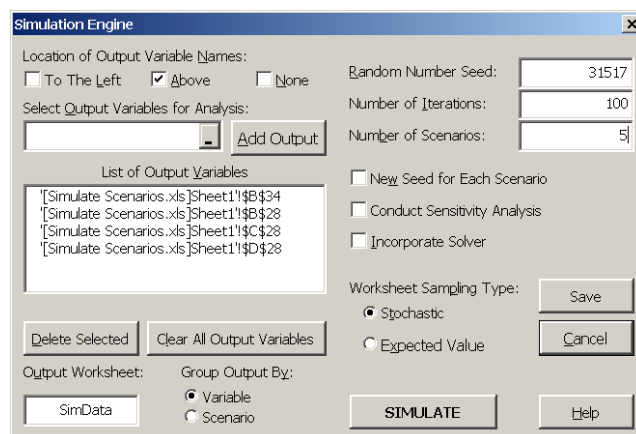


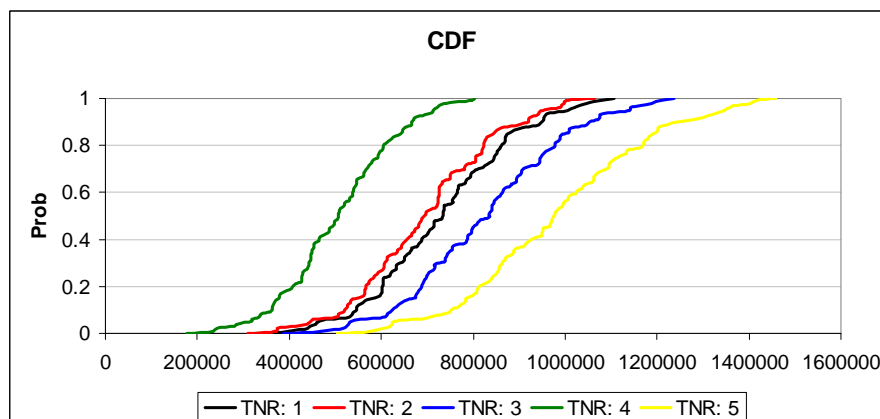
Figure 32. Scenario Analysis Dialog Box.

The values for the first scenario in cells B21:D21 appear in the =SCENARIO() after the functions have been entered. During simulation the subsequent scenario values of Hours Product_{ij} (values in rows 22-25) are used, when the Simulation dialog box (Figure 32) is set to simulate 5 scenarios. If the *Number of Scenarios* cell in Figure 32 is set to 1, only the values for the first scenario are used in simulation. The cells containing the =SCENARIO() function must be used in the equations of the model for the multiple scenario option to work. For example B27 is used in B30 and B32 below. See Simulate Scenarios Demo.xls for the example provided below.

	A	B	C	D	E	F	
19	Scenario Analysis Table contains the alternative values for each						
20	of the variables across the n scenarios.						
21		Hours Product 1	Hours Product 2	Hours Product 3			
22	Scenario 1	1400	1700	2660			
23	Scenario 2	1200	3060	1500			
24	Scenario 3	2000	1000	2760			
25	Scenario 4	0	2500	3260			
26	Scenario 5	3000	2760	0			
27		Product 1	Product 2	Product 3	Formulas for Product 1		
28	Active Scenario (Hrs)	1400	1700	2660	=SCENARIO(B22:B26)		
29	Stoch Price	2.229	3.119	1.963	=TRANPOSE(B10:B12)		
30	Stoch Quantity/Hour	87.232	33.395	62.889	=TRANPOSE(B13:B15)		
31	Receipts	272,266	177,052	328,303	=B29*B30*B28		
32	Resource Cost/Hour	25	35	48	Constants		
33	Net Return	237,266	117,552	200,623	=B31-B32*B28		
34	TNR	555,441	Key Output variable				

The results of a scenario simulation can be reported to SimData two ways using the *Group Output* option in the Simulation Dialog Box (Figure 32). Grouping the results by *Variable* causes Simetar to present the results in SimData as: Scenario 1-M for Variable 1, then Scenario 1-M for Variable 2, and so on for K output variables. Grouping the results by *Scenario* causes Simetar to present the results as: Variables 1-K for Scenario 1, then Variables 1-K for Scenario 2, and so on for M scenarios. Both formats have their own advantage, use the one which best suits your purpose. It is recommended when using the Scenario option that the *List of Output Variables* include the cells associated with the =SCENARIO() functions. This will facilitate verifying that the values in the Scenario Table were appropriately used in the simulation.

The benefit of using the =SCENARIO function is that Simetar runs the model multiple times using exactly the same random deviates (risk) for each scenario. Thus the analysis guarantees that each scenario was simulated using the same risk and the only difference is due to the differences in the scenario variables. The results can be presented as charts and used in risk ranking analyses.



8.0 Sensitivity Analysis

When the *Conduct Sensitivity Analysis* option in the Simulation Engine dialog box is selected, the Simulation Engine dialog box expands to add the sensitivity options in Figure 33. Simetar systematically manipulates one exogenous variable at a time to quantify the sensitivity of the output variables. The *Select Input Variable to Manipulate* cell can refer to any cell in the Excel workbook. The variable to manipulate can be either a constant or a formula. In either case, Simetar uses the initial value as the base and simulates the model using fractional deviations about the base value.

Figure 33. Simulation Sensitivity Dialog Box.

The range of test values for the manipulated input variable are specified using the three *Sensitivity Ranges*. If you are interested in testing the effects of +/- 5, 10, and 15 percent changes in the selected input variable, type these values in the *Sensitivity Range* boxes and simulate the model. If further investigation shows that the ranges could be +/- 3, 6, and 9 percent, then type in these values and re-simulate the model.

Results of sensitivity analyses are summarized in the SimData worksheet. The results are presented, by output variable, in the following order: the Base value for the Input Variable to Manipulate (or IVM) is $1.0 * IVM$, the smallest IMV (say, $0.85 * IVM$), the next larger IMV (say, $0.9 * IVM$), and so on until the seventh value which is the largest IMV tested (say, $1.15 * IVM$). This organization of results facilitates direct comparison of the impacts of the IMV on each of the Output Variables using a Fan Graph.

It is recommended that when sensitivity analyses are being simulated, the list of Output Variables in the Simulation Engine should include the Input Variable to Manipulate. Using this convention, one can easily verify that the Input variable indeed took on the intended values.

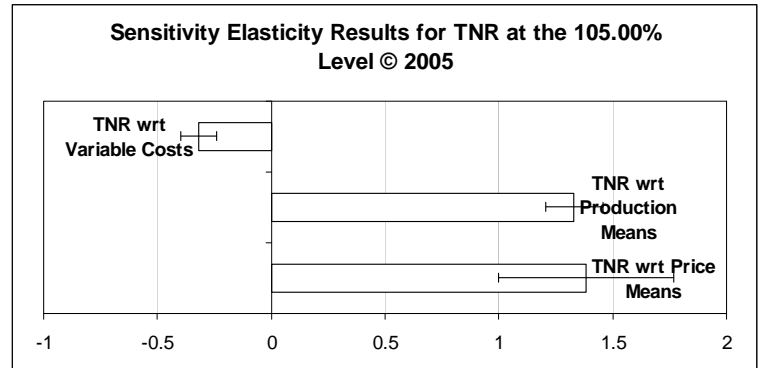
9.0 Sensitivity Elasticity Analysis

The sensitivity of a key output variable (KOV) in a simulation model to changes in several exogenous variables can be measured using sensitivity elasticities (SE_i). A SE_i is like an elasticity, but it quantifies the average percentage change in a KOV to a one percent change in the exogenous variable X_i . Simetar calculates SE_i

values by simulating the model for the base value of each X_i exogenous variable to be tested. Next Simetar changes one X_i at a time

by a specified percentage change and simulates the model. The SE_i values are calculated for each X_i value across all iterations and the mean and standard deviation of the SE_i are reported in worksheet SEDATA.

A chart of the SE_i values is provided so the analyst can see which X_i variable has the greatest impact on the KOV. The standard deviation for the SE_i 's is displayed in the SE chart as well. To simulate SE values for stochastic simulation model in Excel, select the *Calculate Sensitivity Elasticities* button in the Simulation Engine (Figure 34). This action causes Simetar to expand the Simulation Engine menu to include the inputs for SEs. Select the one KOV to be used for the analysis and select the exogenous variables for which SE's are to be estimated. Specify the percentage change to use for estimating the SE's; 5 percent is usually adequate for this purpose. Simulate the model and review the simulated results in the SimData and SEData worksheets. Edit the SE chart using the Excel chart commands. An example Sensitivity Elasticity chart is presented in Simulate Sensitivity Elasticities Demo.xls.



10.0 Simulating and Optimization

Stochastic simulation and optimization of a model is complicated because it requires iteratively simulating random shocks to the equations and then optimizing the system. For example, in a two equation supply and demand model with stochastic shocks we would solve for the price that makes demand equal supply or:

$$Q_S = a + b \text{ Price} + cX + (\text{Std Dev} * \text{SND})$$

$$Q_D = a + b \text{ Price} + cY + (\text{Std Dev} * \text{SND})$$

$$E_S = Q_S - Q_D$$

If the stochastic shock is zero ($\text{SND} = 0.0$) we simply use Excel's Solver (Figure 35) to solve for the price where ending stocks (E_S) equals zero. See the Sim Solve Demo.xls for an example. Sim

Solve Demo.xls demonstrates how a simultaneous equation system can be simulated using the *Incorporate Solver* option in Simetar (Figure 36).

The first step in simulating a stochastic simultaneous equation model is to use Excel's Solver (**Tools > Solver**) to specify the change variable (price, in the example) and the target

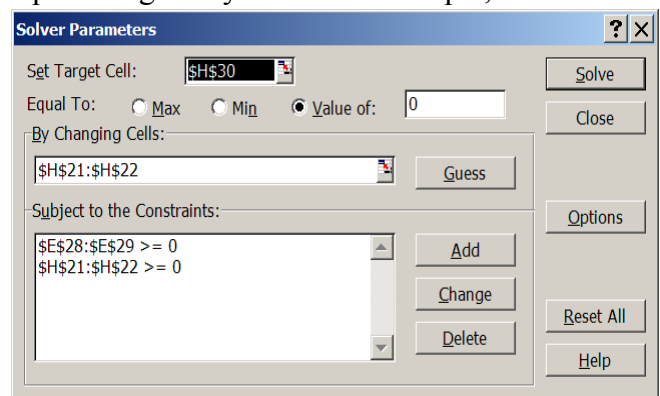


Figure 35. Excel's Solver Dialog Box for Solving an Optimal Control Problem.

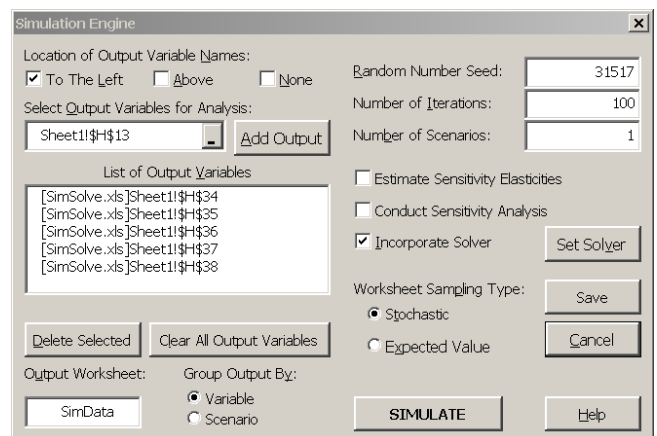


Figure 36. Sim-Solver Dialog Box.

variable (stocks or E_S , in the example). An example of Excel's Solver dialog box is provided in Figure 35. While the spreadsheet is set to Expected Value, solve the model using Solver, after specifying the Solver parameters, and then open the Simetar Simulation Engine. In the Simulation Engine dialog box select the *Incorporate Solver* option and specify the output variables and simulate the model as usual (Figure 36). It is recommended that the *Output Variables* include the control variable and the target value which Solver is programmed to optimize. This pair of output variables allows one to check Solver's results for each iteration. As should be expected the *Incorporate Solver* option is slow. The reason being that Excel is solving an optimal control problem 100 or more times. Sim-Solver option works well for small models but will not be efficient for large simulation models with numerous (10 or more) simultaneous equations. See the example in Sim Solve Demo.xls.

11.0 Numerical Methods for Ranking Risky Alternatives

The results of a Simetar simulation are written to the SimData worksheet. The results can be analyzed many different ways to help the decision maker determine the most preferred alternative.

Functions in Simetar to facilitate analysis of simulation results are described in this section.

11.1 Stochastic Dominance (SD)

11.1.1 First Degree Stochastic Dominance. First degree SD is the least discriminating stochastic dominance method for ranking risky alternatives. However, if the CDFs for the risky alternatives do not cross, this is the preferred method for ranking alternatives.




First degree SD can be accessed in Simetar by selecting the  icon. Select the *1st and 2nd Degree Dominance Table* option Simetar will develop first degree stochastic dominance table (Figure 37). The Stochastic Dominance dialog box (Figure 35) requires the analyst enter the location for the simulated values of the risk alternatives (or scenarios) specify the risk aversion coefficients (RACs). The first degree SD table will be placed in the SD1 spreadsheet. See the example below and in Stochastic Dominance Demo.xls.

Figure 37. Stochastic Dominance Dialog Box.

11.1.2 Second Degree Stochastic Dominance. Second degree SD assumes the decision maker is risk averse so the RACs must be positive. The  icon causes Simetar to open the Stochastic Dominance menu (Figure 37) which asks for the simulated values for the risky alternatives and the RACs (say -0.0001 and 0.0001). By selecting the *1st and 2nd Degree Dominance Table* option Simetar will develop a second degree SD output table in the SD1 worksheet. The results for a second degree SD analysis are generally inconclusive. See the example below and in Stochastic Dominance Demo.xls.


First Degree Dominance Table					
	TNR: 1	TNR: 2	TNR: 3	TNR: 4	TNR: 5
TNR: 1 FDD:		FDD		FDD	
TNR: 2 FDD:				FDD	
TNR: 3 FDD:	FDD	FDD		FDD	
TNR: 4 FDD:					
TNR: 5 FDD:	FDD	FDD	FDD	FDD	
					The table says that TNR:1 is second degree dominate over the 2 and 4 risky alternatives.
					The table says that TNR:5 is first degree dominate over the risky alternatives.
Second Degree Dominance Table					
	TNR: 1	TNR: 2	TNR: 3	TNR: 4	TNR: 5
TNR: 1 SDD:				TNR: 4	
TNR: 2 SDD:		TNR: 2		TNR: 4	
TNR: 3 SDD:	TNR: 1	TNR: 2		TNR: 4	
TNR: 4 SDD:					
TNR: 5 SDD:	TNR: 1	TNR: 2	TNR: 3	TNR: 4	
					The table says that TNR:3 is second degree dominate over the 1, 2, and 4 risky alternatives.
					The table says that TNR:5 is second degree dominate over the risky alternatives.

11.1.3 Generalized Stochastic Dominance with Respect to a Function (SDRF). The SDRF option is initiated by selecting the  icon which opens the dialog box depicted in Figure 37. When specifying the simulation results in *Select Arrays to Compare*, be sure to highlight the label in row one and all of the rows (simulated values) and columns (scenarios or alternatives) to compare. Use the Add button to add scenarios that are not adjacent to the first scenario added in the *Select Array* window. All of the scenarios must have the same number of observations. The SDRF comparison of risky alternatives uses the *Lower and Upper Risk Aversion Coefficients* (RACs) the user specifies in the dialog boxes (Figure 37). The lower RAC must be less than the upper RAC. No scaling takes place with the user’s RAC values. If a RAC is too large in absolute terms (relative to the series to analyze), the STODOM ranking results will show “#VALUE!” rather than ranking each scenario. This result comes about because an exponent overflow is caused by excessively large RACs.

The SDRF results table are written to worksheet SDRF1 (see the example below). The SDRF results table is dynamic so the user can systematically change the RACs in the stochastic dominance results table and observe the effect on scenario rankings. When the SDRF table uses simulation results in the SimData worksheet, the SDRF table will be updated automatically each time Simetar simulates the model.

	A	B	C	D	E	F	G
1	Analysis of Stochastic Dominance with Respect to a Function (SDRF)						
2	© 2005						
3							
4		Efficient Set Based on SDRF at				Efficient Set Based on SDRF at	
5		Lower RAC	0			Upper RAC	0.00001
6		Name	Level of Preference			Name	Level of Preference
7	1	TNR: 5	Most Preferred		1	TNR: 5	Most Preferred
8	2	TNR: 3	2nd Most Preferred		2	TNR: 3	2nd Most Preferred
9	3	TNR: 1	3rd Most Preferred		3	TNR: 1	3rd Most Preferred
10	4	TNR: 2	4th Most Preferred		4	TNR: 2	4th Most Preferred
11	5	TNR: 4	Least Preferred		5	TNR: 4	Least Preferred

11.2 Stochastic Efficiency with Respect to a Function (SERF)

SERF is a new procedure for ranking risky alternatives based on their certainty equivalents (CE) for alternative absolute risk aversion coefficients (ARACs). The CEs for risky alternatives are calculated and the results are presented in a table and a chart by selecting the SERF option in the Simetar toolbar, . The SERF icon opens the SERF Analyzer dialog box (Figure 38). The SERF table and chart are placed in a worksheet named SERFTb11. The CE values in the table and chart are dynamic so the lower and/or upper ARACs and the utility function can be changed after the dialog box has been run.

The SERF procedure defaults to the Exponential Utility Function, yet six more utility functions are available in cell D4 of SERFTb11. The SERF table values and chart can be calculated assuming a Power Utility Function by typing a “2” in place of the “1” in cell D4. The rule for ranking risky alternatives is that at any given ARAC value, the

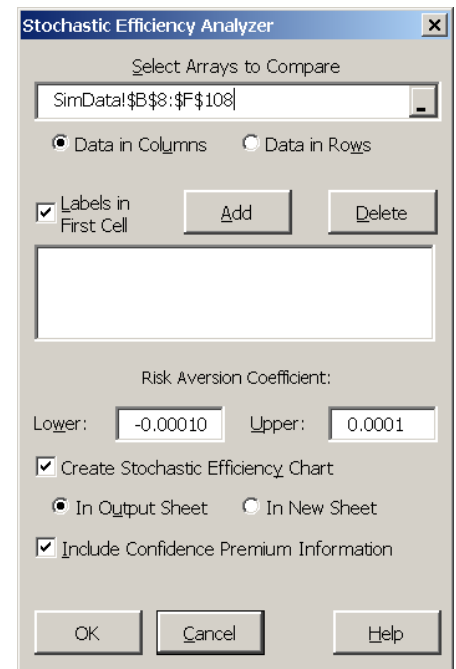
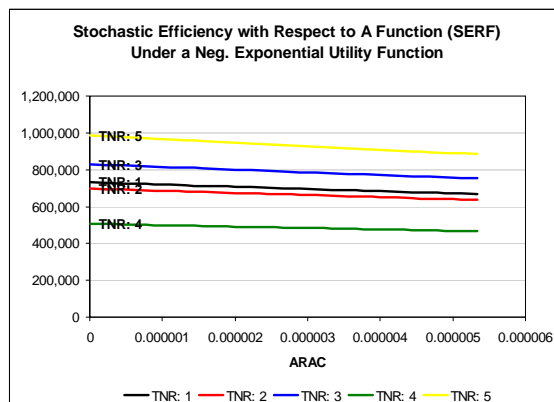
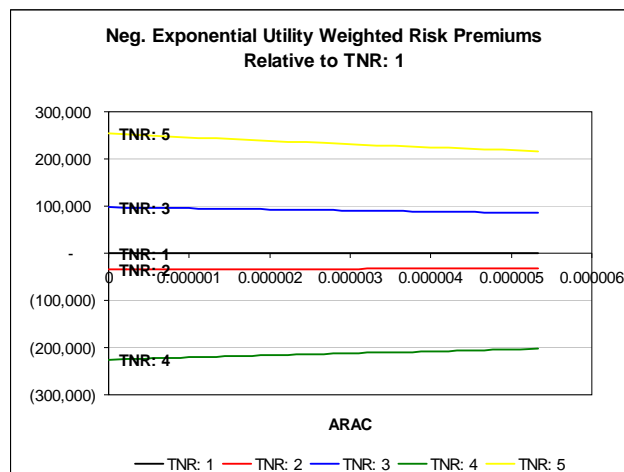


Figure 38. SERF Dialog Box.

preferred alternative is the one which is the highest on the Y (or CE) axis. An example of the SERF analysis is available in SERF Analysis Demo.xls.

11.3 Risk Premiums

The confidence premium (or the conviction level) with which a decision maker would prefer one alternative over another is visually displayed in the SERF Chart as the vertical distance between the CE lines at each RAC. The SERF analysis also produces a certainty equivalents risk premium (RP) table and chart in the SERFTbl1 worksheet. The RP table compares the absolute differences in the CE's for a base alternative with the other alternatives across RAC values. A chart of the RP's displays the relative position of each alternative to the base over the range of the RACs. The user can change the lower and upper RACs and the alternative designated as the base.



An example of the RP analysis is presented here and in SERF Analysis Demo.xls. The dynamic nature of the SERF option will degrade execution time if the model is re-simulated. If this is a problem, delete the SERFTbl1 Worksheet before re-simulating the model.

11.4 Target Probabilities for Ranking Risky Alternatives

The probability of a variable taking on a value less than or equal to a specified target value for a simulated distribution can be calculated using the =EDF() function in Simetar. Risky alternatives can be ranked with respect to their probabilities exceeding target values. The =EDF() function is programmed as follows:

= EDF (Array Location, Target Value)
 = EDF (B8:B108, B110)

where: Array Location is the location for the distribution (simulation results) to analyze, and Target Value is the location for the target value or an actual number.

An example of how the =EDF() function can be used is to first simulate net returns for a business. The probabilities of observing net returns less than particular target values are calculated using =EDF(). Alternative target values for net returns can be specified by the decision maker. See the Stoplight chart in Section 6.6 for a graphical means of calculating and displaying target probabilities. An sample table of EDF values is presented below from the Simulate Scenarios Demo.xls workbook.

	A	B	C	D	E	F	G	H	I
85	Demonstrate the Quantile and EDF Functions for Analyzing Scenarios of Net Returns.								
86	Prob X<x	TNR: 1	TNR: 2	TNR: 3	TNR: 4	TNR: 5			
87	600,000	17.7%	27.3%	7.8%	77.8%	2.8%	=EDF(SimData!F\$9:F\$108,Sheet1!\$A67)		
88	800,000	88.9%	72.8%	45.1%	99.8%	17.2%	=EDF(SimData!F\$9:F\$108,Sheet1!\$A68)		
89	1,000,000	94.9%	98.2%	85.4%	100.0%	56.4%	=EDF(SimData!F\$9:F\$108,Sheet1!\$A69)		
70									
71	Quantiles	TNR: 1	TNR: 2	TNR: 3	TNR: 4	TNR: 5	=SimData!F8		
72	0.100	543,591	519,786	621,780	360,289	751,556	=QUANTILE(SimData!F\$9:F\$108,Sheet1!\$A72)		
73	0.250	618,722	586,552	700,798	428,431	846,876	=QUANTILE(SimData!F\$9:F\$108,Sheet1!\$A73)		
74	0.500	733,556	693,150	834,415	506,923	975,918	=QUANTILE(SimData!F\$9:F\$108,Sheet1!\$A74)		
75	0.750	846,658	807,150	948,356	592,738	1,117,963	=QUANTILE(SimData!F\$9:F\$108,Sheet1!\$A75)		
76	0.990	1,092,817	1,033,887	1,223,517	799,216	1,437,351	=QUANTILE(SimData!F\$9:F\$108,Sheet1!\$A76)		

11.5 Target Quantiles for Ranking Risky Alternatives

Instead of ranking risky alternatives based on their probability of exceeding a target, some decision makers want to know the target value which has a particular probability of being true, or the quantile for their KOV. This method can be implemented by calculating the value of the key output variable at, say, the 25 percentile. The =QUANTILE() function returns the value of a series that is associated with a specified probability. If =QUANTILE() is given a series of values, such as, [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] and asked to locate the 35th quantile, then the function returns the value of 3.5 as the 35th quantile value. The array of values to evaluate does not have to be sorted from low to high. An sample table of QUANTILE values is presented above from the Simulate Scenarios Demo.xls workbook. The function is used as:

```
=QUANTILE (Array Location, Percentile)
=QUANTILE(B9:B108,0.56)
```

where: Array Location is the cell reference for the distribution to be evaluated, and Percentile is the percentile to evaluate and is a fraction, such as 0.56.

12.0 Tools for Data Analysis and Manipulation

The Simetar functions developed to facilitate data analysis and manipulation are described in this section. All of the Simetar functions in this section are dynamic so if the historical data for a model or its stochastic variables change, the parameters are automatically updated. This feature is particularly useful when developing simulation models that can use different input data from one application to another. Another feature of Simetar functions is that the formulas are cell locked so the formulas can generally be copied and pasted or dragged to new locations to speed up the data analysis process

12.1 Matrix Operations

Most data in an Excel workbook can be thought of as a matrix. Thirty-three Simetar functions that facilitate the manipulation and analysis of data matrices can be accessed by clicking the **M** icon (Figure 39). The Simetar functions are programmed in C++ and therefore not constrained to Excel's restrictions on array size. The matrix functions are in alphabetical order in the Matrix Operations dialog box:

- Center Matrix of a Specified Dimension
- Choleski Factorization of a Matrix
- Cofactor of a Square Matrix
- Column Vector to a Diagonal Matrix
- Column Vector to a Matrix
- Column Vector to a Toeplitz Matrix
- Concatenate Two Matrices
- Determinant of a Square Matrix
- Eigenvalues of a Square Matrix
- Eigenvectors of a Square, Symmetric Matrix
- Equicorrelation Matrix of a Specified Dimension
- Exponential Power of a Matrix
- Factor a Square, Symmetric Matrix
- Generalized Inverse of a Matrix
- Inner Product of Two Matrices

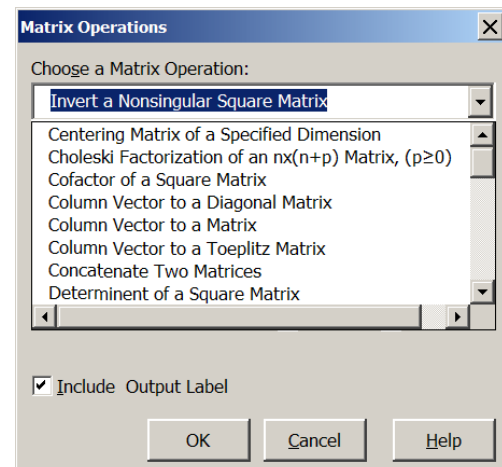


Figure 39. Matrix Operation Menu.

Invert a Nonsingular Square Matrix
 Kronecker Multiply Two Matrices
 Mahalanobis Distance of Two Data Matrices
 Matrix of 1s
 Matrix to a Vector
 Multiply Two Matrices
 Norm of a Matrix
 Orthogonalize a Matrix
 Rank of a Matrix
 Reduced Row Echelon Form of a Matrix
 Reverse a Column or Row of Values
 Row Echelon Form of a Matrix
 Sequence of Numbers
 Sort a Matrix by a Specified Column
 Sweep a Square Matrix on a Diagonal Element
 Trace of a Square Matrix
 Transpose a Matrix
 Wishart Matrix of Random Variables

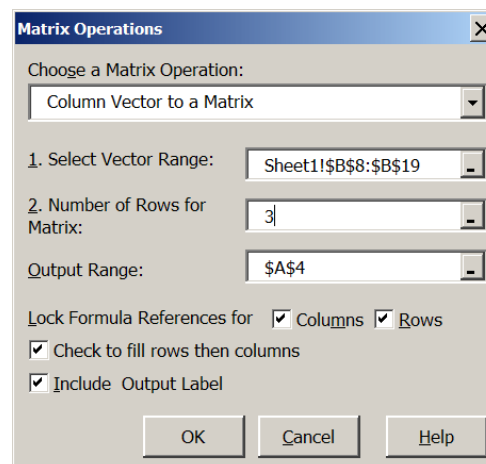


Figure 40. Dialog Box for Changing a Vector to a Matrix.

The most frequently used matrix functions are described in detail in this section. The Simetar Matrix and array functions are dynamic so changes made to the data are automatically observed in the output functions. For example, changes to the input data will change the associated correlation matrix, the Choleski decomposition matrix of the correlation matrix, and subsequent calculations for parameter estimation and stochastic simulation. The matrix functions described in Section 12.0 are demonstrated in the Excel workbook Matrix Operation Functions.xls.

12.1.1 Column Vector to a Matrix. The Matrix Operations dialog box accessed by selecting the **M** icon contains a function to *Change a Column Vector to a Matrix* (Figure 40). The function is dynamic so changes in the original vector are observed in the matrix.

12.1.2 Reverse a Column or Row of Values. A vector of values can be reversed by selecting the *Reverse a Column or Row of Values* in the Matrix Operations menu. The function outputs the data as a column if a column of input is provided and as a row if the input is in a row.

12.1.3 Convert a Matrix to a Vector. The task of converting a matrix of weekly, monthly, or quarterly data to a vector for time series analysis is simplified with the *Matrix to a Vector* function. To use this function indicate the matrix to operate on and the output location for the vector.

12.1.4 Sort a Matrix. An array or a matrix can be sorted in Simetar using the *Sort a Matrix by a Specified Column* in the Matrix Operations menu. The user must specify the *Column to Sort By* as well as the location for the matrix. The sort is dynamic so as the values in the original data matrix change, the values in the sorted matrix will be updated.

12.1.5 Factor a Square Matrix. Simetar can factor a covariance or correlation matrix for simulating a multivariate probability distribution by either the Square Root method or the Choleski method. Both of these

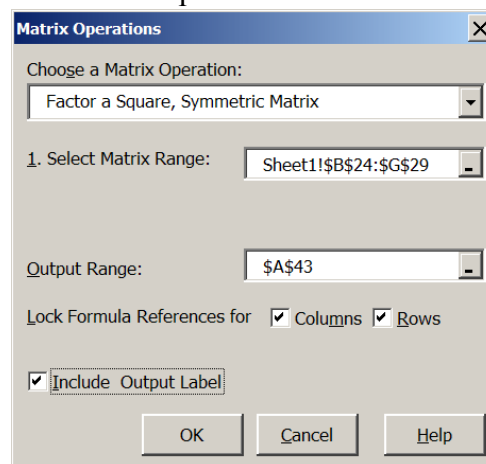


Figure 41. Factor a Square Matrix Dialog Box.

methods are accessed via the  icon for matrix functions (Figure 41).

12.1.6 Transpose a Matrix (Excel). A matrix can be transposed by selecting the *Transpose a Matrix* option in the Matrix Operations dialog box, specifying the matrix to transpose and the upper-left hand cell to anchor the output matrix. This procedure simplifies Excel's transpose function by eliminating the need to block the area for the transposed matrix and avoids array size limitations in Excel.

12.1.7 Generalized Inverse of a Rectangular Matrix. The *Generalized Inverse of a Matrix* function in the Matrix Operations dialog box uses Simetar's function. Select this option and specify the input matrix (highlight only the numbers) and the output range for the upper left hand value, then select OK. The inverse of the input matrix will appear in the worksheet without row/column names. Copy and paste in the names if needed.

12.1.8 Invert a Nonsingular Square Matrix (Excel). The *Invert a Nonsingular Square Matrix* option in the Matrix Operations dialog box is demonstrated in Figure 42. (Simetar uses Excel's function but provides an easy to use menu.) Select this option and then specify the input matrix (highlight only the numbers) and the output range for the upper left hand value, then click OK. The inverse of the input matrix will appear in the worksheet without row/column names. Copy and paste in the names if needed.

12.1.9 Multiply Two Matrices (Excel). Excel's matrix multiplication, MMULT, function is made easier by selecting the *Multiply Two Matrices* option in the Matrix Operations dialog box. An additional feature is that Simetar's matrix multiplication will handle larger matrices than the Excel function MMULT.

12.1.10 Concatenate Two Matrices. A new matrix of data can be developed by concatenating the data from two locations in the workbook. The *Concatenate Two Matrices* option in the Matrix Operations menu requires as input the location of the two input arrays or matrices and the output location.

12.1.11 Convert a Vector to a Diagonal Matrix. In simulation it is useful to convert a vector of standard deviations to a diagonal matrix. The Simetar function =MDIAG() can be used to convert an array to a diagonal matrix using the *Column Vector to a Matrix* option in the Matrix Operations dialog box.

12.1.12 Find the Determinant of a Square Matrix. The determinant of a square matrix can be calculated by selecting the *Determinant of a Square Matrix* option in the Matrix Operations dialog box. The Excel function =MDETERM (square matrix) is used for this calculation.

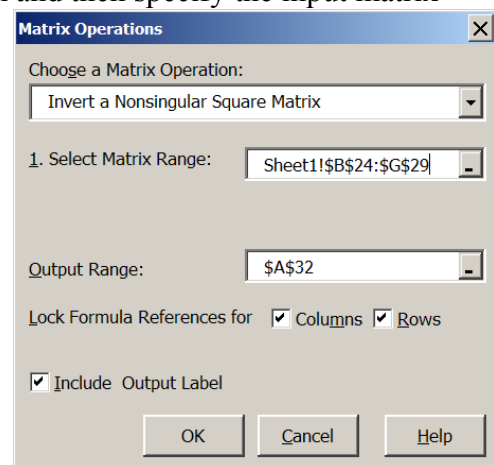



Figure 42. Invert a Square Matrix Dialog Box.

12.2 Data Manipulation

Data often comes in the wrong format or orientation. Data may be in an array when we need it in a matrix or vice versa. Sometimes we need to reverse the order of the data or concatenate arrays from different places in the worksheet. Functions to make these data manipulations easy have been included in Simetar and can be accessed by selecting the  icon. Additional data manipulation functions are also presented in this section.

12.2.1 Create an Identity Matrix. An indemnity matrix of dimension $N \times N$ can be generated using the =MIDEN() function in Simetar. The format for the function is =MIDEN (dimension) where dimension is a scalar to specify the number of rows in the square identity matrix.

11.2.2 Create a Sequence of Numbers. A sequence of numbers in an array can be created using the =SEQ() function. The =SEQ() returns a column of numbers that follow any sequence you specify. The function is programmed as:

=SEQ(No. of Values, Starting Value, Interval or Increment)

where: No. of Values is the number of cells to be highlighted,
Starting Value is the first value in the sequence, and
Interval or Increment is the interval between each value.

For example the sequence of number for 10, 20, 30, ..., 200 is generated by programming the function as =SEQ(20,10,10) and a sequence of 2, 4, 6, ..., 20 is generated by programming the function as =SEQ(10, 2, 2).

11.2.3 Create a Matrix of Ones. In statistics a J matrix is an array or matrix with a 1.0 in each cell. The Simetar function =MJ() is used to create a J matrix. To create a 10x1 array of 1.0s highlight 10 cells in column and type the function =MJ(10). To create a 10x10 matrix of 1.0s, highlight a 10 cells in 10 columns and type =MJ(10,10). Be sure to hit the Control Shift, Enter keys after typing the =MJ() function as it is an array function.

11.2.4 Create a Centering Matrix. The =MCENTER() array function that creates an $N \times N$ centering matrix when n is specified as the dimension.

11.2.5 Create an Equicorrelation Matrix. The =MEQCORR() array function generates an $N \times N$ equicorrelation matrix using any specified correlation coefficient. The =MEQCORR() function is an array function so you must highlight the cells for the square equicorrelation matrix and end the function by hitting *Control Shift Enter*. The function is programmed as =MEQCORR(Rho), where Rho is the correlation coefficient.

11.2.6 Create a Toeplitz Matrix. The =MTOEP() array function creates a square symmetric Toeplitz matrix given a column or row of data. To create a Toeplitz matrix of an array in A1:A4, highlight a 4X4 array and type the function as =MTOEP(A1:A4). Be sure to press *Control Shift Enter* as this is an array function.

12.3 Box-Cox Transformation

The =BOXCOX() function can be used to transform the data for a skewed distribution to make it approximately normally distributed. The function uses a user specified exponent to transform the data. The =BOXCOXEXP() function is provided to estimate an appropriate exponent. The format for the Box Cox transformation functions are:

=BOXCOX(Data Array, Power Value, [Shift to Plus])

where: Data Array refers to the location of the Nx1 data series to be transformed,
Power Value is the exponent for the transformation, and
Shift to Plus is an optional term if the data are to be shifted to positive values enter 'TRUE or 1', otherwise enter 'FALSE or 0'.

The =BOXCOX() function is an array function so highlight the appropriate number of cells and type the function and press *Control Shift Enter*. See Data Analysis Demo.xls for an example. Once a model has been estimated using a Box-Cox transformation, the =UNBOXCOX() function can be used to transform the forecast values back to original data. The reverse Box-Cox transformer function is:

=UNBOXCOX (Data Array, Power Value, Original Data Array, [Shift to Plus])

where: Data Array is the location for the Nx1 array transform back to the original data,
Power Value is the exponent for the transformation,
Original Data Array is the location for the original data Nx1 array, and
Shift to Plus is an optional term if the data are to be shifted to positive values enter 'TRUE or 1', otherwise enter 'FALSE or 0'.

The maximum likelihood estimation of the Box-Cox transformation exponent function can be calculated using the following function:

=BOXCOXEXP(Data Array, [Shift to Plus], [Lower], [Upper], [Max Iter])

where: Data Array refers to the location of the data n-1 array to be transformed,
Shift to Plus is an optional term if the data are to be shifted to positive values,
Lower is an optional minimum for the search routine, -2 is the default,
Upper is an optional maximum for the search routine, +2 is the default, and
Max Iter is an optional parameter for the search routine.

12.4 Workbook Documentation

12.4.1 Delete Numbers in a Cell. When a cell has both numbers and text, to extract only the text, use the =DELNUM() function. See Data Analysis Demo.xls for an example. If cell A1 contains the string "1013 Sycamore Street" and we want the text in cell A2, then in A2 type:

=DELNUM(A1)



12.4.2 Delete Text in a Cell. Often times the numbers in a cell are needed even though the cell

contains both numbers and text. For example, the worksheet may have an address in a cell as “1013 Sycamore Street” and we want the number without the text. Rather than re-typing the numbers to a new cell or editing the existing cell use the =DELTEXT() function. See Data Analysis Demo.xls for an example. Say the cell A1 has the string “1013 Sycamore Street” and you want just the number to appear in cell B1, then in B1 type:

```
=DELTEXT (A1)
```

12.4.3 View Cell Formulas. To show the formula typed in a particular cell use =VFORMULA(). An advantage of using this function is that you can both see the formula for a cell, say B24, and you can see the value in B24. The =VFORMULA() function is dynamic and changes (updates itself) as rows and columns are added to or removed from the worksheet. The Simetar function to view the formula in cell B24 can be typed into any cell (say, C24) as follows:

```
= VFORMULA (B24)
```

12.4.4 View All Formulas. In the process of writing and documenting simulation models in Excel we often write formulas that need to be printed. Simetar provides a function to easily view every cell in the worksheet as a formula, and then switch the worksheet back to values. This function can be accessed by clicking the  icon in the Simetar toolbar. Click the  icon a second time and the worksheet will return to the normal view.

12.4.5 Workbook and Worksheet Name. Functions in Simetar have been provided to dynamically show the name of the workbook or the worksheet in a cell. These functions are useful for documenting a model. The workbook name is shown in any cell that contains the following command:

```
=WBNAME( )
```

The worksheet name is shown in any cell that contains the following command:


```
=WSNAME( )
```

If you rename the workbook or the worksheet, the function updates the text in the cell after pressing F5.

13.0 Regression Analysis

Simple and multiple regression (ordinary least square (OLS), Probit, Logit, GLS, Ridge, 2SLS, and GLS) capabilities are included in Simetar to facilitate estimating parameters for simulation models. Not only are the regression coefficients (beta-hats) useful, but in simulation the residuals are used to quantify the unexplained risk for a random variable. The regression functions in Simetar take advantage of Excel’s ability to recalculate all cells when a related value is changed. Thus when an observed X or Y value is changed the betas are recalculated. Also, multiple regression models can be instantly re-estimated for different combinations of the X variables by using restriction switches to ignore individual variables.

13.1 Simple Regression

The parameters for a simple OLS regression are calculated when you select the  icon. The simple regression icon opens the dialog box depicted in Figure 43 so the X and Y variables can be specified. The intercept (\hat{a}) and slope (\hat{b}) parameters for the equation:

$$\hat{Y} = \hat{a} + \hat{b} X$$

are estimated and placed in the worksheet starting where the *Output Range* specifies. The names of the estimated parameters appear in the column to the left of the parameters. The R^2 , F-Ratio, Student's *t*-test statistics, and residuals are calculated if you select the appropriate boxes.

	A	B	C
74	Simple Regression of Y = a + b X		
75		Corn P	
76	Intercept	1.98	
77	Slope	0.049231	
78	R-Square	0.211367	
79	F-Ratio	2.948195	
80	Prob(F)	0.11396	
81	S.E.	0.028672	
82	T-Test	1.717031	

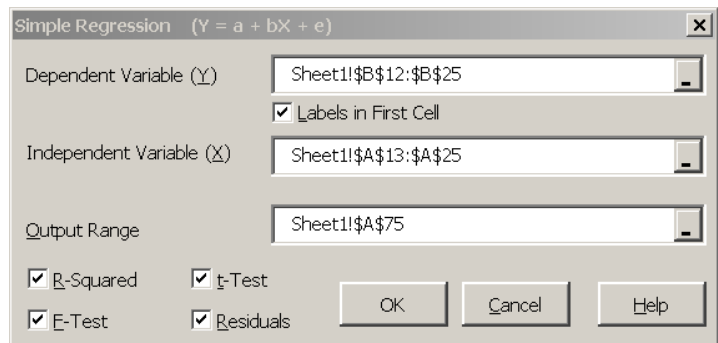



Figure 43. Simple Regression Dialog Box.

Be sure that X and Y have the same number of observations when you specify their ranges in the Simple Regression dialog box. This Simetar function is useful for checking the presence of a trend in a random variable Y. In this case, create a column of X values that increment from 1, 2, 3, ..., N and then use Simetar to estimate the regression parameters. A feature to this function is that the coordinates for the X variable are cell reference locked (fixed) so the formula cells can be copied and pasted across the spreadsheet to estimate simple regressions for numerous Y's using a common X or trend variable. An example of the simple regression function in Simetar is provided below and in the Data Analysis Demo.xls workbook.

13.2 Multiple Regression

The Multiple Regression option is accessed through the  icon. Multiple regression estimates the least squares \hat{a} and \hat{b}_i parameters for:

$$\hat{Y} = \hat{a} + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \dots + \hat{b}_n X_n$$

The Multiple Regression dialog box (Figure 44) allows the user to specify the Y and X variables, and the type of output for seven different multiple regression models.

A sample output for a multiple regression is provided below to show the format for the first part of the results. The name of an X variable and its beta are in bold if the variable is statistically

significant at the indicated one minus alpha level (e.g., X_1 , X_2 , X_3 , and X_4 in the example). Standard errors for the betas, the t-test statistics and the probability (p) value of the t-statistics are provided for each explanatory variable. The elasticity at the mean for each independent variable as well as the partial and semi-partial correlations for these variables is provided as well. The variance inflation factor is reported for each X variable to indicate the degree of multicollinearity of X_i to other variables in the model. See Multiple Regression Demo.xls for the example presented in this section.

The Restriction row in the parameter block of output values allows the user to interactively experiment with various combinations of X variables. After the initial parameter estimation the Restriction coefficients are all blank, meaning that every X variable is included in the unrestricted model. The user can interactively drop and re-include a variable by changing its restriction coefficient from blank to 0. Compare the results in the first example to those in the second example where X_5 was restricted out of the model. The exclusion of X_5 improves the F –test (61.5 vs. 79.2).

Three test statistics (F, R^2 and \bar{R}^2) for the Unrestricted Model are provided and remain fixed while testing alternative specifications of the model’s variables. This is done to facilitate the comparison to the original unrestricted model to the restricted models. If you type a non-zero number in the restriction row, the value becomes the beta-hat coefficient for a restricted regression.

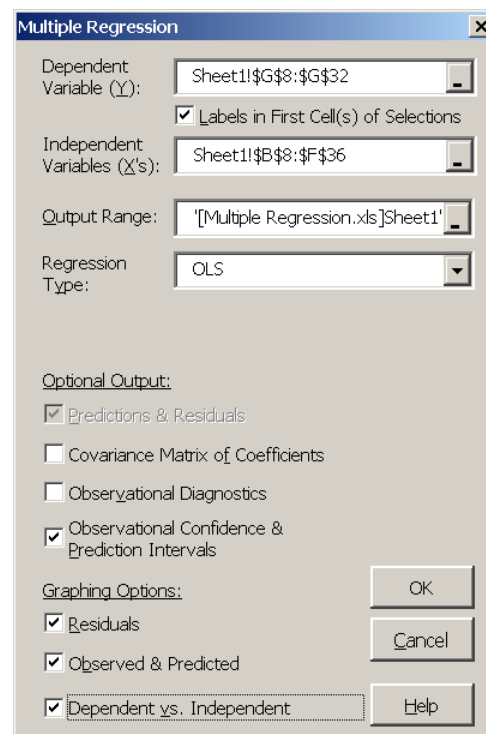


Figure 44. Multiple Regression Dialog Box.

	A	B	C	D	E	F	G
40	OLS Regression Statistics for Y, 11/18/2005 10:53:10 AM						
41	F-test	88.156	Prob(F)	0.000	Unrestricted Model		
42	MSE ^{1/2}	0.011	CV Regr	0.204	F-test	88.156	
43	R ²	0.961	Durbin-Watson	1.768	R ²	0.961	
44	RBar ²	0.950	Rho	-0.006	RBar ²	0.950	
45	Akaike Infor	-8.883	Goldfeld-Qua	1.618	Akaike Infor	-8.883	
46	Schwarz Infr	-8.638		Schwarz Infor	-8.638		
47		95% Intercept	X1	X2	X3	X4	X5
48	Beta	0.417	-0.156	0.298	-0.014	0.611	-0.013
49	S.E.	0.776	0.042	0.109	0.003	0.097	0.047
50	t-test	0.537	-3.734	2.730	-4.302	6.307	-0.268
51	Prob(t)	0.598	0.002	0.014	0.000	0.000	0.792
52	Elasticity at Mean		-0.077	0.436	-0.035	0.612	-0.014
53	Variance Inflation Factor		18.155	67.613	98.155	4.997	4.105
54	Partial Correlation		-0.661	0.541	-0.712	0.830	-0.063
55	Semipartial Correlation		-0.174313628	0.127443618	-0.200861965	0.294466122	-0.012500054
56	Restriction						

	A	B	C	D	E	F	G
40	OLS Regression Statistics for Y, 11/18/2005 10:53:10 AM						
41	F-test	115.837	Prob(F)	0.000	Unrestricted Model		
42	MSE ^{1/2}	0.011	CV Regr	0.199	F-test	88.156	
43	R ²	0.961	Durbin-Watson	1.764	R ²	0.961	
44	RBar ²	0.952	Rho	0.004	RBar ²	0.950	
45	Akaike Infor	-8.863	Goldfeld-Qua	1.378	Akaike Infor	-8.883	
46	Schwarz Infr	-8.766		Schwarz Infor	-8.638		
47		95% Intercept	X1	X2	X3	X4	X5
48	Beta	0.370	-0.153	0.288	-0.014	0.620	0.000
49	S.E.	0.737	0.039	0.099	0.003	0.089	0.000
50	t-test	0.502	-3.910	2.898	-4.547	6.946	0.000
51	Prob(t)	0.622	0.001	0.010	0.000	0.000	1.000
52	Elasticity at Mean		-0.076	0.421	-0.034	0.621	0.000
53	Variance Inflation Factor		16.708	58.635	88.872	4.454	4.105
54	Partial Correlation		-0.668	0.554	-0.722	0.847	0.000
55	Semipartial Correlation		-0.178026492	0.131961938	-0.207052398	0.316282667	0
56	Restriction						0

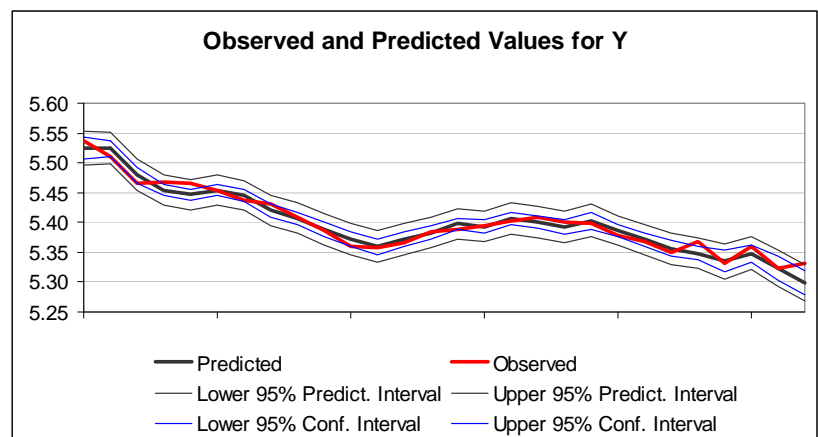
In addition to the ability to exclude and re-include variables in the model, Simetar’s multiple regression function allows the analyst to make corrections to the data for the actual observations of the X and Y values, without having to re-run the regression. The Simetar multiple regression routine is not limited in the number of exogenous variables that can be included in the model. Regression

	A	B	C	D	E	F	G	H	I	J
66	S.D. Resids	0.009581383	MAPE	0.14581097	Lower	Upper	Lower	Upper	Scatter Plot of A	
67	Actual Y	Predicted Y	Residuals	SE Mean Pre	SE Predicted	95% Conf. Int	95% Conf. Int	95% Predict.	195% Pred	X1
68	5.536	5.524	0.012	0.008	0.014	5.507	5.542	5.485	5.594	2.850
69	5.511	5.524	-0.013	0.008	0.013	5.511	5.537	5.497	5.551	2.934
70	5.497	5.479	-0.014	0.008	0.013	5.498	5.492	5.453	5.505	3.011
71	5.485	5.447	0.038	0.004	0.012	5.446	5.463	5.429	5.478	2.918
72	5.452	5.454	-0.002	0.004	0.012	5.445	5.463	5.422	5.471	2.982
73	5.430	5.445	-0.007	0.005	0.012	5.435	5.455	5.420	5.479	2.890
74	5.431	5.421	0.010	0.005	0.012	5.410	5.431	5.395	5.446	2.910
75	5.409	5.407	0.002	0.005	0.012	5.397	5.417	5.382	5.433	2.854
76	5.398	5.398	-0.002	0.006	0.012	5.376	5.400	5.362	5.414	2.820
77	5.359	5.372	-0.013	0.008	0.012	5.360	5.395	5.346	5.399	2.754
78	5.358	5.359	-0.001	0.006	0.013	5.347	5.371	5.333	5.385	2.721
79	5.387	5.372	-0.005	0.006	0.013	5.359	5.395	5.345	5.398	2.863
80	5.385	5.383	0.002	0.005	0.012	5.372	5.394	5.357	5.409	2.599
81	5.388	5.388	-0.010	0.004	0.012	5.388	5.407	5.373	5.423	2.559
82	5.394	5.393	0.001	0.005	0.012	5.383	5.404	5.368	5.419	2.547
83	5.403	5.408	-0.003	0.005	0.012	5.398	5.417	5.381	5.432	2.485
84	5.408	5.400	0.008	0.005	0.012	5.389	5.412	5.375	5.426	2.612
85	5.401	5.393	0.009	0.006	0.012	5.380	5.405	5.366	5.419	2.526
86	5.398	5.403	-0.005	0.007	0.013	5.389	5.417	5.376	5.430	2.340
87	5.378	5.386	-0.008	0.004	0.012	5.377	5.386	5.361	5.411	2.386
88	5.380	5.371	-0.004	0.005	0.012	5.381	5.392	5.348	5.397	2.338
89	5.349	5.356	-0.007	0.006	0.013	5.343	5.369	5.330	5.383	2.338
90	5.360	5.348	0.020	0.006	0.012	5.337	5.360	5.322	5.374	2.269
91	5.367	5.335	0.000	0.000	0.014	5.317	5.353	5.305	5.364	2.374
92	5.346	5.348	-0.002	0.007	0.013	5.333	5.362	5.320	5.375	2.250
93	5.306	5.323	-0.000	0.000	0.014	5.304	5.343	5.293	5.353	2.337
94	5.261	5.288	-0.000	0.000	0.014	5.278	5.310	5.260	5.328	2.328

models with 5000 observations and 250 X variables can be estimated with Simetar.

If the analyst specifies more observations for the X variables than for the Y variable, Simetar will forecast the Y values. The forecast values in the “Predicted Y” column of the output uses the betas for the regression and the additional Xs. Probabilistic forecasts of the Y variable are provided as bold values in the Actual Y column of the output. For the example, there are five extra X values indicated for the regression dialog box (Figure 44) so Simetar calculated the deterministic forecast values in column B and the probabilistic forecast values in column A, starting in row 91 (see the output above). Probabilistic forecasts are calculated assuming normality, the mean equals the deterministic forecast, and the standard deviation is the standard error of the predicted Y in column E for the example. Press F9 to make Excel simulate the probabilistic forecasts. The probabilistic forecasts can be used in a stochastic simulation model.

Residuals for the regression are also included in the example output. The residuals for the regression are calculated as $\hat{\epsilon}_i = Y_i - \hat{Y}_i$ for each observation i and represent the unexplained risk for the dependent Y variable. The standard error for the mean predicted value (SE mean predicted) is provided for each observation i . In addition the SE of the Predicted Y for each observation is provided in column E of the example output. As indicated in the example output, the SE Predicted Values increase as the forecasted period gets longer. Prediction and confidence intervals for the model are provided in the table (above) and graphically (below) for the alpha equal 5 percent level. The alpha level can be changed by changing the value in line 47 of the output example from 95% to, say, 90% or 99%.




The observed and predicted Y values can be viewed graphically along with the confidence and prediction intervals. For the example program five more Xs than Ys were used to estimate the model, as a result the last five values in the Observed line to the right are the probabilistic forecast values and will change each time the F9 function key is pressed.

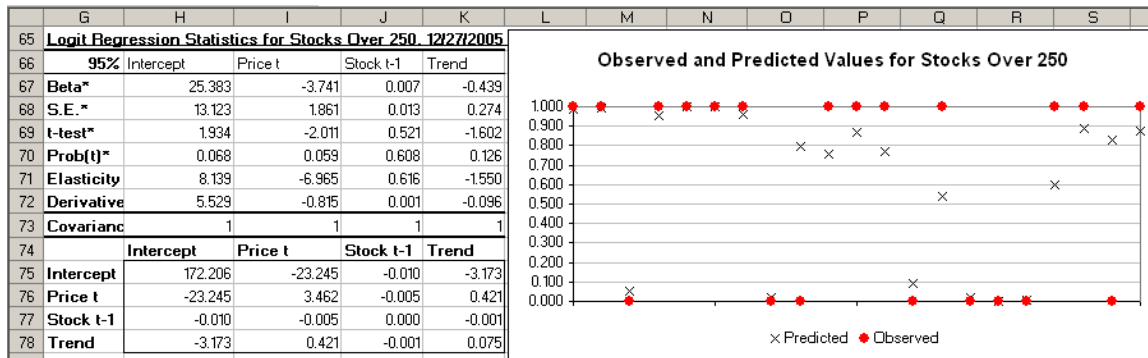
The covariance matrix for the betas is an optional output for multiple regressions. The beta covariance matrix is used in simulation when the model is assumed to have stochastic betas. The beta covariance matrix is provided when specified as an option in the multiple regression dialog box (Figure 44).

If requested in the regression dialog box (Figure 44), observational diagnostics are calculated and reported for the unrestricted model (see the example to the right). The column of 1's in the DFBetas Restriction column indicate that the unrestricted model was fit using all of the observed data.

If you change a DFBetas Restriction to 0 for a particular row the model is instantly updated using a dummy variable to ignore the effects for that row of X's and Y. The rule for excluding an observation is if its Studentized Residual is greater than 2 (is bold). This is the case for observation 24 in the sample output. Setting the Restriction value to 0 for observation 24 causes the F statistic to increase from 88 to 107, given that X_5 has not been excluded from the model. The R^2 increases to 96.1 from 96.9 (see Multiple Regression Demo.xls). This result suggests that observation 24 is either an outlier or should be handled with a dummy variable. A priori justification should be used when handling observations in this manner.

A		B		C		D		E		F		G		H		I		J		K		L		M		
95		Observational Diagnostics																								
96		DFBetas																								
97		Restriction	Observation	Intercept	X1	X2	X3	X4	X5	Cutoff Values																
										0.408	0.408	0.408	0.408	0.408	0.408	0.500	2.000	1.750	1.000	0.250						


13.3.2 Logit Analysis. The LOGIT function estimates a logistic regression given dependent and independent variables. Logit regression models can be estimated by using the multiple regression icon  and selecting the Logit option in the menu, see Figure 44 for the menu. The LOGIT function allows for independent variables to be restricted from the complete model. In addition, individual observations can be restricted from the regression. The LOGIT function uses an iteratively re-weighted least squares technique to estimate the model parameters. A sample Logit output for Simetar is presented below from the Probit and Logit Demo.xls.



14.0 Cyclical Analysis and Exponential Forecasting

Functions to facilitate analysis of seasonal and cyclical data are included in Simetar. Seasonal indices and moving average analysis of cyclical data are described in this section. Three different procedures for developing exponential forecasts included in Simetar are described as well.

14.1 Seasonal Index

A seasonal index of any array can be calculated by Simetar using the Forecasting and Cyclical Data icon  and clicking on the *Seasonal Indexing* tab. The Seasonal or Cyclical Indexing dialog box (Figure 45) allows the user to specify the data series to analyze and the number of periods in the cycle, (say, 4 or 8 or 12). A sample output table is presented below and in Seasonal Analysis Demo.xls.

When the input data are months and the *Number of Periods in the Cycle* is 12 the result will be a 12 month seasonal index. The quarterly index in the example below is developed from five years of quarterly sales to calculate a seasonal sales index.

A seasonal index can be calculated one of two ways, namely: simple average or centered moving average. The simple average index is a more reliable indicator of the seasonal pattern if the data has no trend. If the data series has an underlying trend the *Centered Moving Average* will remove a portion of the variability caused by the trend. The Seasonal and Cyclical Indexing

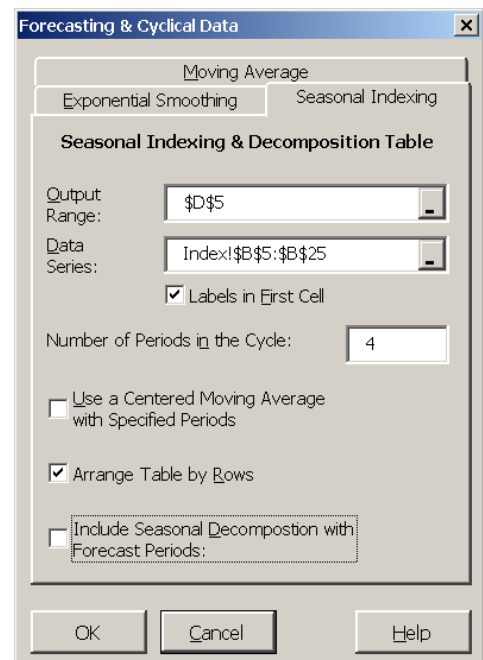



Figure 45. Seasonal or Cyclical Indexing Dialog Box.

dialog box (Figure 45) assumes the user wants a simple average index.

	D	E	F	G	H	I	J	K	L	M
5	Cyclical Indices for Sales (\$1,000s), 11/18/2005 12:28:26 PM									
6		1	2	3	4	SUM	AVERAGE	ST DEV	INDEX	FRAC. CO
7	1	25.14	15.12	21.41	30.14	91.81	22.9525	5.481165	0.884	0.177
8	2	26.71	18.91	24.15	30.21	99.98	24.995	4.119427	0.962	0.192
9	3	26.12	17.45	23.36	35.61	102.54	25.635	6.555641	0.987	0.197
10	4	27.3	14.89	31.28	38.99	112.46	28.115	8.715614	1.082	0.216
11	5	24.91	16.76	29.67	41.45	112.79	28.1975	8.936178	1.085	0.217
12	SUM	130.18	83.13	129.87	176.4	519.58				
13	AVERAGE	26.036	16.626	25.974	35.28		25.979			
14	ST DEV	0.908815	1.49635	3.815865	4.562157			7.287482		
15	INDEX	1.002	0.640	1.000	1.358				4.0 \ 5.0	
16	FRAC. CONT. INDEX	0.251	0.160	0.250	0.340					1
17	INDEX LCI	0.934	0.464	0.712	1.105					
18	INDEX UCI	1.071	0.816	1.288	1.611		Conf. Level		95%	1.959964

14.2 Seasonal Decomposition Forecasting

A Seasonal Decomposition forecast of a data series can be calculated by Simetar using the Forecasting and Cyclical Data icon  and clicking on the *Seasonal Indexing* tab (Figure 46). After indicating where the data series is located and the number of periods in the cycle, click on the last box in the menu to *Include Seasonal Decomposition with Forecast Periods*. This will cause Simetar to calculate the parameters for a seasonal decomposition forecast for the number of periods indicated in the last window of the dialog box, four for the example presented below.

The output for the seasonal decomposition forecast contains two switches that allow the user to alter the type of decomposition model that best fits the data series being forecasted. The options are Additive and Cycle (see the example output below). The default value for the ADDITIVE option, “TRUE”, is for an additive model which assumes the seasonal component is additive. If the seasonal effects are multiplicative, use the “FALSE” setting for the ADDITIVE option. The second option, CYCLE, defaults to “TRUE” assuming the series has an underlying cycle. If a cycle is not present change this option to “FALSE”.

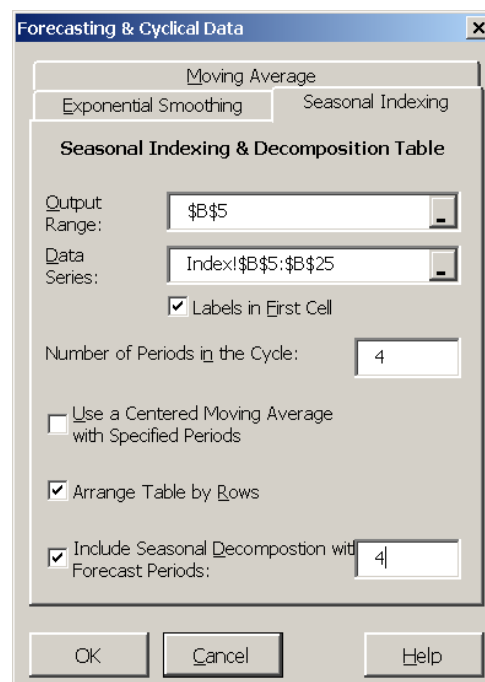



Figure 46. Seasonal Decomposition Forecasting.

The user's requested forecast values are presented in the charts, the trend component forecast is the series of dashes on the linear trend line. The cyclical and seasonal forecasts are the dashes on their respective lines. The composite forecast is the dashes on the actual data line (Sales in the example). The values for these forecast components are indicated in the table after the historical values, the last four values for the example below and in Seasonal Analysis Demo.xls.

14.3 Moving Average Forecast

A moving average of any series can be calculated by selecting the forecasting icon  and selecting the *Moving Average* tab (Figure 47). The Moving Average dialog box requires information on the number of periods to include in the moving average and the number of periods to forecast.

Once Simetar has completed the analysis you can change the number of periods for the moving average using the sliding scale to observe how the number of periods affects the goodness of fit measures. The MAPE, WAPE, Thiel U2, RMSE, and MAE are included in the output so you can experiment with different moving average lengths and observe the affects on forecast error. A graph of the historical and predicted values is provided as well. The example of a moving average forecast below comes from the Moving Average Demo.xls workbook.

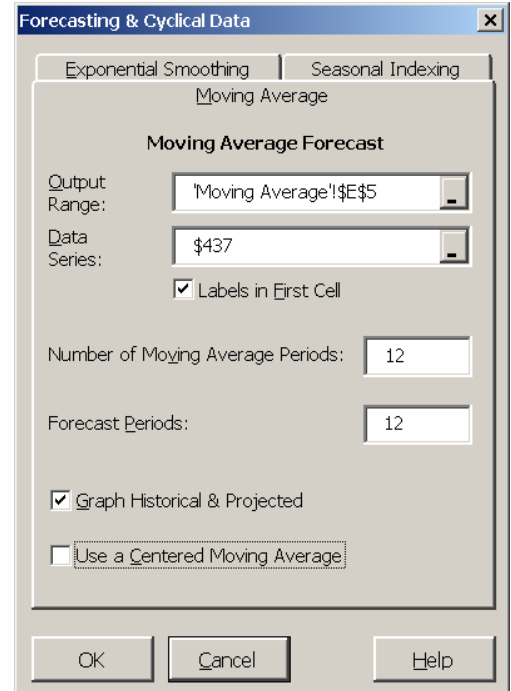
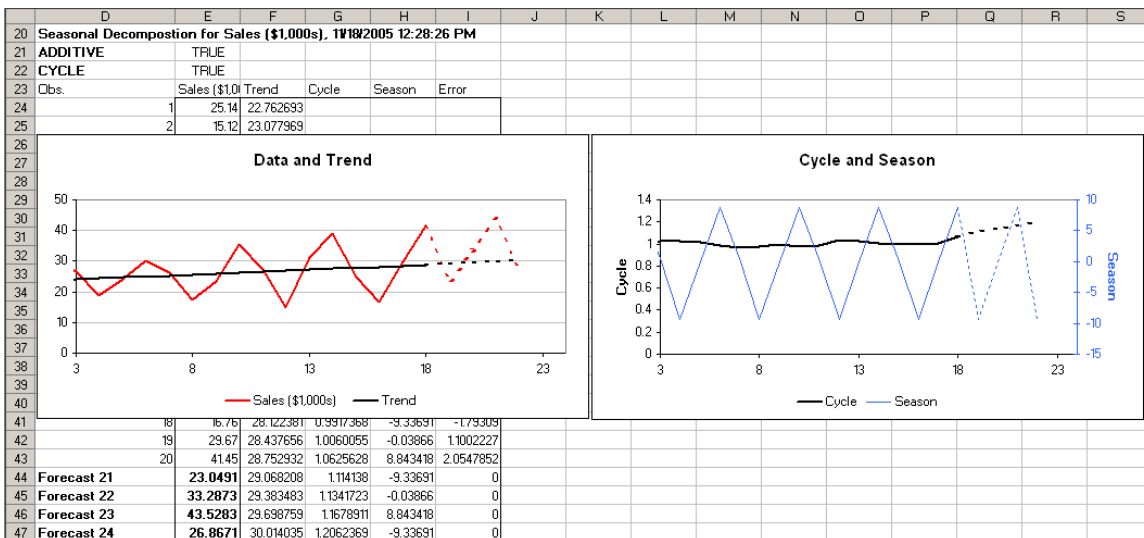
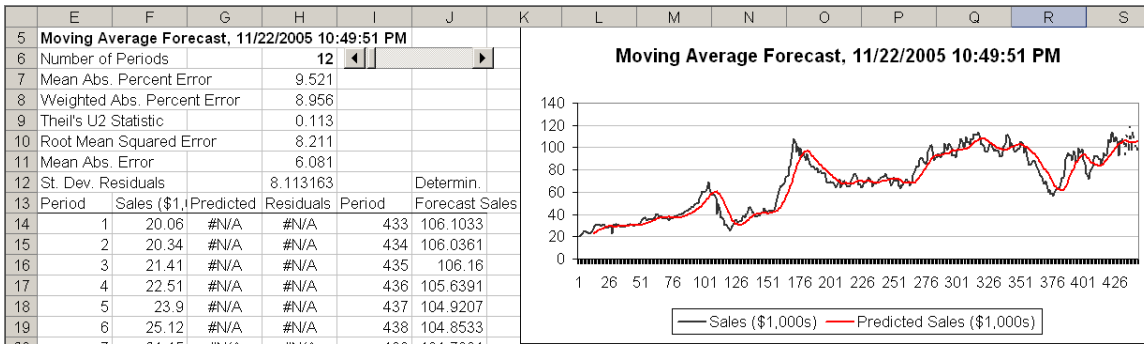



Figure 47. Moving Average Forecast Dialog Box.



14.4 Exponential Smoothing Forecast

An exponential smoothing forecast for any data series can be developed using the forecasting icon  and selecting the *Exponential Smoothing* tab (Figure 48). **Before running the Exponential Smoothing option, open Solver to make Excel activate Solver** in the worksheet where you want the forecast model to appear. Solver can be opened and closed by clicking on **Tools > Solver > Close**

Simetar provides three different exponential smoothing estimator/forecasts tools:

- Single exponential smoothing estimates one parameter alpha (*Dampening Factor*).
- Double exponential smoothing or Holt's method estimates parameters for two parameters alpha and beta (*Optional Trend Factor*).
- Triple exponential smoothing or Holt-Winter's method estimates three parameters alpha, beta, and gamma (*Optional Seasonal Factor*).

Additionally, Simetar estimates the parameters for the exponential smoothing model with different assumptions about the trend and seasonal component. The options are:

- Holt Method Trend with
 - No trend
 - Dampened additive trend
 - Dampened multiplicative trend
- Holt Winters Seasonal with
 - No seasonal component
 - Additive seasonal component
 - Multiplicative seasonal component

These alternative specifications are effected by changing the Trend Method and the Trend Method options from 0 to 1 or 2 in the output. Re-run Solver after changing any option.

Simetar estimates and forecasts the requested model based on the non-zero initial guesses the user provides in the dialog box or by using SOLVER to optimize the parameters by selecting parameters that minimize the MAPE (Figure 48). Probabilistic forecasts of the exponential smoothing model can be observed by setting the Stochastic Forecast option to "TRUE". The probabilistic forecast values appear at the bottom of the second column of the results. See Exponential Smoothing Demo.xls for the example presented below.

After Simetar estimates the initial model, you can experiment with alternative parameters by using the slide scales for the Level Smoothing Constant, the Trend Smoothing Constant, the Season Smoothing Constant, and the Dampening Parameter, to see what they do to the MAPE, RMSE, MAE, etc.

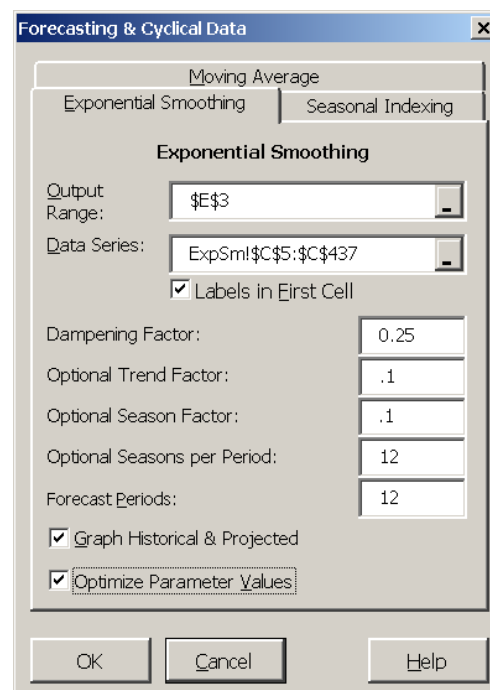
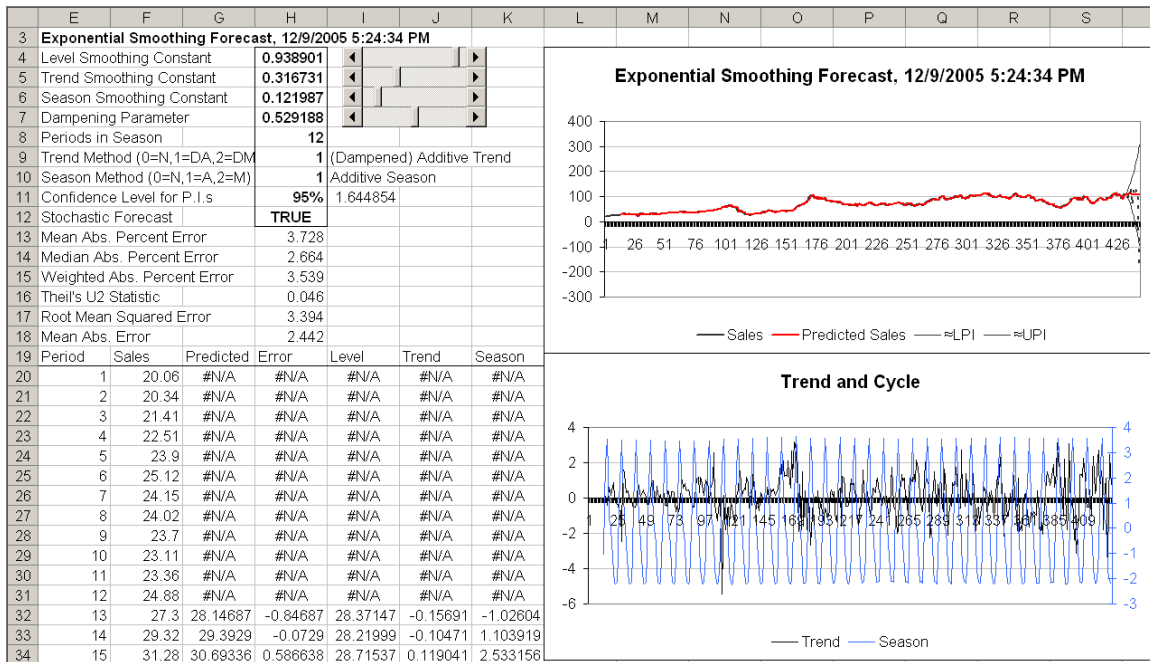


Figure 48. Exponential Smoothing Forecast Dialog Box.



14.5 Measuring Forecast Errors

Five functions are included in Simetar for quantifying forecast errors. The functions are found in most statistics books so the equations are not presented here. An example of the five forecast error statistics is available below and in Forecast Errors Demo.xls.

- Mean Absolute Percent Error function is:
=MAPE (Array of Residuals, Array of History)
- Weighted Absolute Percent function is:
=WAPE (Array of Residuals, Array of History)
- Mean Absolute Error function is:
=MAE (Array of Residuals)
- Root Mean Square Error function is:
=RMSE (Array of Residuals)
- Theil U2 statistic function is:
=THEILU2 (Array of Residuals, Array of History, Change)

where: Array of Residuals is the cell reference for the array of errors or residuals,
Array of History is the cell reference for the array of historical data that was used to generate the residuals, and
Change is an optional term to indicate if the statistic is to be calculated in the given levels of the data or as a function of the changes in forecast. FALSE returns the statistic based on levels; TRUE returns the statistic based on changes. The default value is FALSE.

	E	F	G	H	I	J	K
5	Forecast Error Statistics			Values	Formulas		
6	Mean Abs. Percent Error			5.226	=MAPE(D6:D29,B6:B29)		
7	Weighted Abs. Percent Error			5.295	=WAPE(D6:D29,B6:B29)		
8	Theil's U2 Statistic			0.064	=THEILU2(D6:D29,B6:B29)		
9	Root Mean Squared Error			1.714	=RMSE(D6:D29)		
10	Mean Abs. Error			1.400	=MAE(D6:D29)		
11	Mean			26.43083	=AVERAGE(B6:B29)		
12	Std Dev from Mean			3.657701	=STDEVP(B6:B29)		
13	Std Dev of Residuals Trend			1.714291	=STDEVP(D6:D29)		

15.0 Time Series Analysis and Forecasting

Functions for estimating and forecasting time series models in Simetar are presented in this section. Functions used to test for stationarity and number of lags are described first, followed by a general autoregressive model menu for estimating autoregressive (AR) and vector autoregressive (VAR) models. The time series analysis functions facilitate parameter estimation and forecasting with both AR and VAR models to aid in developing probabilistic forecasts for simulation. The time series capabilities of Simetar are demonstrated in Time Series Demo.xls.

15.1 Tests for Stationarity

Time series models should only be estimated for data series that are stationary. A series can generally be made stationary by differencing. An accepted test for determining if a series is stationary is the Dickey-Fuller test. The Dickey-Fuller Test can be calculated using the Simetar function =DF(). The =DF() function allows the user to test for alternative combinations of differences in an efficient manner to find the combination of adjustments necessary to make a series stationary. The equation used to calculate the DF statistic is:

$$\underbrace{\Delta Y_t = B_0 + B_1 Y_{t-1}}_{\text{Dickey-Fuller Test}} + \underbrace{B_3 T_t + \sum_{i=1}^n \sigma_i \Delta Y_{t-i}}_{\text{Augmented Dickey-Fuller Test}}$$

where: ΔY_t is the first difference of the data series Y ,
 B_0 is the intercept,
 B_1 is the slope parameter estimated for the lagged Y variable (Y_{t-1}),
 B_3 is the slope parameters estimated for the trend variable (T), and
 σ_i is the parameter for the ΔY_{t-i} for different lengths of higher order lags (i), such as first, second, third, ... order lags.

The Dickey-Fuller Test uses the first two components of the above equation and tests for the presence of nonstationarity, in the absence of trend. The Augmented Dickey-Fuller Test includes the third and/or the fourth components of the equation to test for the presence of a trend in the series and for higher order differences. The Simetar function to calculate the Dickey-Fuller Tests on a series of data is:

$$=DF(\text{Y Values Range}, [\text{Time Trend}], [\text{No. of Lag Diffs}], [\text{No. of Diff.]})$$

where: Y Values Range is the location of the data series to be tested (this is all that is necessary for the basic Dickey-Fuller Test),
 Time Trend is a true or false switch to indicate whether a trend is to be included in the Augmented Test: “False or 0” for no trend and “True or 1” for a trend,
 No. of Lag Diffs is the number of higher order lags to use for the Augmented Test, usually 0, (this is the value for n in the ΔY_{t-i} summation), and
 No. of Diff is the number of differences for the original data series Y. This parameter can be used to test for nonstationarity of a specified number of differences, say 2.

Examples of using the =DF() function are provided below and in Time Series Demo.xls to demonstrate how it can be used. The basic Dickey-Fuller Test is entered as:

=DF(Y Values Range)

The Augmented Dickey-Fuller Test that includes a trend is entered as :

=DF(Y Values Range, 1)

The Augmented Dickey-Fuller Test that has no trend and tests for the presence of a second order autocorrelation lag is entered as:

=DF(Y Values Range, 0, 2)

The Augmented Dickey-Fuller Test that includes trend and tests for the presence of a second order autocorrelation lag is entered as:

=DF(Y Values Range, 1, 2)

The null hypothesis for the Dickey-Fuller Tests is: H_0 : data series is nonstationary. The critical test statistic for the Dickey-Fuller Test, based on large sample theory, is approximately -2.9 at the 5% level. The null hypothesis is rejected if the DF statistic is less than the -2.9 critical value. The Dickey-Fuller test demonstrated above is in the Tests worksheet of Time Series Demo.xls workbook. The Dickey-Fuller tests for the data are reported for alternative lags, differences, and trend show how the function can help identify the combination of differences, trend, and lags necessary to make the raw data series stationary.

	C	D	E	F	G	H	I	J
6	Demonstrate the use of Dickey-Fuller Test							
7	=DF(Data Range, [Time Trend True or False], [No. of Lags], [No. of Differences])							
8	Dickey-Fuller Test of the Witchita Data							
9	No. Diff	No. Lags	Trend	=DF() Statistic	Formula for DF()			
10	0	0	0	-2.6038	=DF(\$A\$7:\$A\$91,E10,D10,C10)			
11	1	0	0	-12.146	=DF(\$A\$7:\$A\$91,E11,D11,C11)			
12	2	0	0	-16.188	=DF(\$A\$7:\$A\$91,E12,D12,C12)			
13	3	0	0	-18.86	=DF(\$A\$7:\$A\$91,E13,D13,C13)			
14	0	1	0	-1.7454	=DF(\$A\$7:\$A\$91,E14,D14,C14)			
15	1	2	0	-6.3545	=DF(\$A\$7:\$A\$91,E15,D15,C15)			
16	2	3	0	-8.0561	=DF(\$A\$7:\$A\$91,E16,D16,C16)			
17	3	4	0	-10.776	=DF(\$A\$7:\$A\$91,E17,D17,C17)			
18	0	0	1	-4.1588	=DF(\$A\$7:\$A\$91,E18,D18,C18)			
19	1	0	1	-12.111	=DF(\$A\$7:\$A\$91,E19,D19,C19)			
20	Critical Value is -2.90							
21	Reject "H ₀ : The Series is Nonstationary" if DF is less than the Critical Value							

15.2 Number of Lags

For time series analysis it is necessary to determine the optimal number of lags for the AR model after determining the number of differences necessary to make the series stationary. The =ARLAG() function in Simetar suggests the optimal number of lags to use for the AR model. The =ARLAG() function returns the number of lags that minimizes the Schwarz criterion given a particular number of differences. The function is programmed as:

=ARLAG (Y Values Range, [Constant], [No. of Diff])

where: Y Values Range is the range of the time series data to be evaluated,

Constant is an optional term if the AR model is expected to have a constant term (true or 1) or has no constant (false or 0). The default is to use a constant term (true) if the value is omitted, and

No. of Diff is the optional number of differences of the original data series Y assumed to make the series stationary.

The =ARLAG() function bases its suggestion for the number of lags on the Schwarz criterion test. The test statistic for the Schwarz criterion can be calculated using the following Simetar function

=ARSCHWARZ (Y Values Range, [Constant], [No. of Diff])

where: All parameters are defined the same as the ARLAG function.

A table for implementing the =ARLAG() and =ARSCHWARZ() functions is demonstrated above. In Excel these functions are dynamic, so you can change the number of differences or the presence of a constant and observe the change in the test statistics. An example of how the =ARLAG() and the =ARSCHWARZ() functions are used is provided in the Tests worksheet of Time Series Demo.xls workbook. Both tests are demonstrated for 1-4 differences, with and without the constant term. Use the =ARSCHWARZ() function to test alternative differences and select the lag structure that minimizes the Schwarz test statistic.

15.3 Sample Autocorrelation Coefficients

In time series modeling it is useful to estimate the sample autocorrelation coefficients and the sample partial autocorrelation coefficients. These coefficients are calculated using the Simetar functions =AUTOCORR() and =PAUTOCORR(). The functions are programmed as:

=AUTOCORR (Y Values Range, No. of Lags, No. of Diff)

and

=PAUTOCORR (Y Values Range, No. of Lags, No. of Diff)

where: Y Values Range is the range of the time series data to be evaluated,

No. of Lags is the number of higher order lags to test, and

No. of Diff is the number of differences of the original data series Y to test.

Both of these functions can be used as “scalar” or “array” functions. When used as a scalar, the functions return a single value in the cell which is highlighted. The value returned is the correlation coefficient or the partial autocorrelation coefficient. To use these functions in their array form, highlight three cells in a 3x1 or 1x3 pattern, enter the function name and parameters indicated above, and then press the *Control Shift Enter* keys. Three values will be calculated and placed in the

	C	D	E	F	G	H	I	J
26	Demonstrate the AR and ARSCHARZ Functions to Test the Number of Lags for							
27	Alternative Assumptions about Constant and Numbers of Differences.							
28		No. of		Calculated				
29	Constant	Differences		No of Lags		Formula		
30				=ARSCHWARZ()				
31	0	1		1.23	=ARSCHWARZ(\$A\$7:\$A\$91,C31,D31)			
32	0	2		1.45	=ARSCHWARZ(\$A\$7:\$A\$91,C32,D32)			
33	0	3		1.73	=ARSCHWARZ(\$A\$7:\$A\$91,C33,D33)			
34	1	1		1.23	=ARSCHWARZ(\$A\$7:\$A\$91,C34,D34)			
35	1	2		1.45	=ARSCHWARZ(\$A\$7:\$A\$91,C35,D35)			
36	1	3		1.73	=ARSCHWARZ(\$A\$7:\$A\$91,C36,D36)			
37				=ARLAG()				
38	0	1		1	=ARLAG(\$A\$7:\$A\$91,C38,D38)			
39	0	2		2	=ARLAG(\$A\$7:\$A\$91,C39,D39)			
40	0	3		5	=ARLAG(\$A\$7:\$A\$91,C40,D40)			
41	1	1		1	=ARLAG(\$A\$7:\$A\$91,C41,D41)			
42	1	2		2	=ARLAG(\$A\$7:\$A\$91,C42,D42)			
43	1	3		5	=ARLAG(\$A\$7:\$A\$91,C43,D43)			

highlighted array. The first value (top or left most) is the autocorrelation or partial autocorrelation coefficient. The next (middle) value is the Student's -t statistic for the coefficient. The last value is the standard error for the coefficient. In the array form these functions can be used to develop tables showing the autocorrelation coefficients and their levels of statistical significance for alternative numbers of lags and differences.

The example on the right demonstrates using the two functions to estimate sample autocorrelation and partial autocorrelation coefficients. The example comes from the Tests worksheet of the Time Series Demo.xls workbook. Four different lags and first and second differences were tested for the data series. Both autocorrelation functions are demonstrated in array form and the partial autocorrelation coefficient function is demonstrated as a scalar to develop a table of test statistics.

	C	D	E	F	G	H	I	J	K	L
48	Demonstrate the Auto-Correlation and Partial Auto-Correlation Coefficients Functions									
49	Sample and Partial Auto Correlation Coefficient Functions Demonstrated as Array Functions									
50	Test Lags of 1, 2, 3, and 4 periods, assuming First or Second Differenced Data									
51	Sample Auto Corr Student-t Std Error for									
52	No Lags	No. Diff		Coefficient	Statistic	Auto Corr	Coef	Formula for =Autocorr() in Column F		
53	1	1		-0.2885	-2.6444	0.10911		=AUTOCORR(\$A\$7:\$A\$91,C53,D53)		
54	2	1		-0.2054	-1.743	0.11784		=AUTOCORR(\$A\$7:\$A\$91,C54,D54)		
55	3	1		0.17633	1.44494	0.12203		=AUTOCORR(\$A\$7:\$A\$91,C55,D55)		
56	4	1		-0.1216	-0.9723	0.12503		=AUTOCORR(\$A\$7:\$A\$91,C56,D56)		
57	1	2		-0.5313	-4.8405	0.10976		=AUTOCORR(\$A\$7:\$A\$91,C57,D57)		
58	2	2		-0.1176	-0.8566	0.1373		=AUTOCORR(\$A\$7:\$A\$91,C58,D58)		
59	3	2		0.26795	1.93455	0.13851		=AUTOCORR(\$A\$7:\$A\$91,C59,D59)		
60	4	2		-0.1289	-0.8915	0.14462		=AUTOCORR(\$A\$7:\$A\$91,C60,D60)		
61	Partial									
62	Auto Corr Student-t Std Error for Partial									
63	No Lags	No. Diff		Coefficient	Statistic	Auto Corr	Coef	Formula for =PAutocorr() in Column F		
64	1	1		-0.2885	-2.6444	0.10911		=PAUTOCORR(\$A\$7:\$A\$91,C64,D64)		
65	2	1		-0.3149	-2.8858	0.10911		=PAUTOCORR(\$A\$7:\$A\$91,C65,D65)		
66	3	1		0.00915	0.08384	0.10911		=PAUTOCORR(\$A\$7:\$A\$91,C66,D66)		
67	4	1		-0.1411	-1.293	0.10911		=PAUTOCORR(\$A\$7:\$A\$91,C67,D67)		
68	1	2		-0.5313	-4.8405	0.10976		=PAUTOCORR(\$A\$7:\$A\$91,C68,D68)		
69	2	2		-0.5572	-5.0763	0.10976		=PAUTOCORR(\$A\$7:\$A\$91,C69,D69)		
70	3	2		-0.2534	-2.3085	0.10976		=PAUTOCORR(\$A\$7:\$A\$91,C70,D70)		
71	4	2		-0.2035	-1.8535	0.10976		=PAUTOCORR(\$A\$7:\$A\$91,C71,D71)		

15.4 Maximum Likelihood Ratio Test

A maximum likelihood ratio test (LRT) is included as a function in Simetar to facilitate estimation of the number of lags for an unrestricted vector autoregressive (VAR) model. The LRT is estimated for alternative possible lags using the following function:

=LRT (Y Values Range, No. of Lags, Constant, No. of Diff, Error Correction)

where: Y Values Range is the range of the time series data to be evaluated for potential inclusion in a VAR. Two or more data series must be identified.

No. of Lags is the number of lags to test,

Constant is a switch as to whether a constant term (True or 1) is to be included or not (False or 0),


No. of Diff is the number of differences of the original data series to test, and

Error Correction is whether to perform an error correction (True or 1) on the data or not (False or 0).

The =LRT() is demonstrated in the Tests worksheet of the Time Series Demo.xls workbook. Two data series were tested for 7 different lags assuming three differences, a constant, and error correction. The parameters for the =LRT() are displayed in a table below the LRTs so one can easily change a parameter and observe the changes in the LRTs.

	C	D	E	F	G	H	I	J	K
82	Demonstrate the LRT or Likelihood Ratio Test								
83					Lags				
84		1	2	3	4	5	6	7	
85	Wichita	45.25	19.23	4.54	21.71	7.12	7.68	6.65	
86	Salina	-36.21	0.09	7.68	0.98	1.37	1.37	10.13	
87		Constant	Diff	Error Correction					
88		TRUE	3	TRUE	Change the bold values and observe the LRTs.				
89		Formula in cell D85							
90		=LRT(\$A\$7:\$A\$91,D84,\$D\$88,\$E\$88,\$F\$88)							

15.5 Estimating and Forecasting Autoregressive (AR) Models

The Time Series Analysis menu (Figure 49) provides the mechanism to program the information necessary to estimate and forecast an auto-regressive (AR) model. The Time Series Analysis menu is activated by selecting the  icon. If you specify the data to analyze as a single variable (column of data) in the *Data Series* window, the Time Series Analyzer will estimate an AR model. (Specifying two or more columns causes Simetar to estimate a VAR model.) The *Number of Lags* and *Number of Differences* for the original data must be specified for the AR model. In addition, provisions are available in the dialog box to indicate whether or not the *Constant is Zero*. The number of *Forecast Periods* to project using the estimated model is also specified in the dialog box. It is recommended that the Time Series menu be programmed to: (a) calculate the residuals, (b) graph the historical and projected values, and (c) graph the impulse response function (see example below).

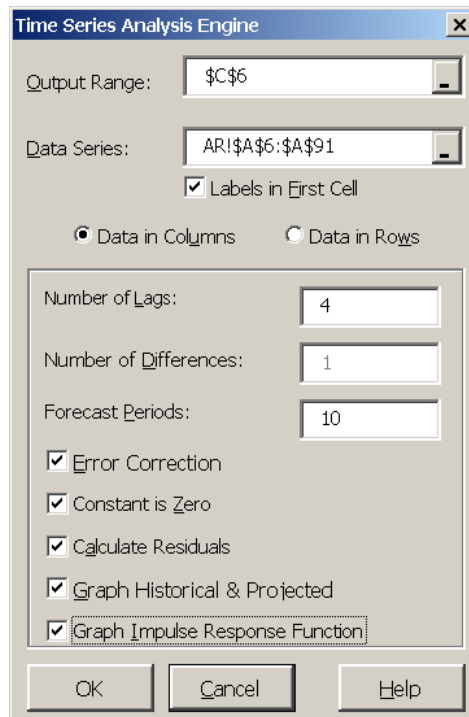
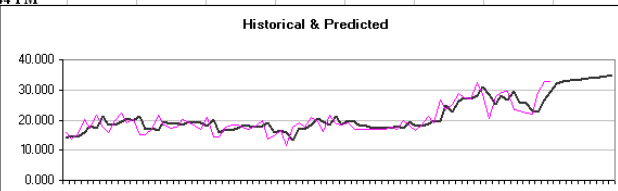
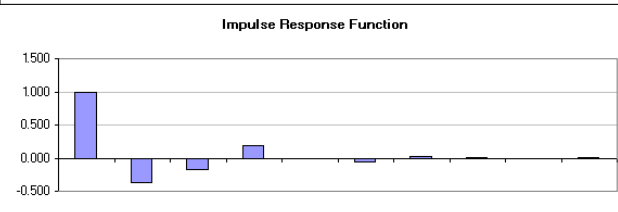


Figure 49. Time Series Analysis Dialog Box.

The results of estimating an AR model with four lags and one difference or an AR (4,1) model is presented below and in the AR worksheet of the Time Series Demo.xls workbook. Several supporting tests are provided along with the coefficients, namely, the Schwarz test, and two Dickey-Fuller tests. The forecast values for the AR model are provided for 10 periods, as programmed in the dialog box, and are labeled “Forecast.” “Impulse Response” values are provided for each forecast period (see the example below). Student-t statistics for the sample and partial autocorrelation coefficients are provided for the 10 periods of forecast output.

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
6	AR Series Analysis Results for 4 Lags & 1 Difference, 12/18/2004 10:09:44 PM															
7	Constant		WichitaL1	WichitaL2	WichitaL3	WichitaL4										
8	Wichita	0.407	-0.383	-0.327	0.000	0.000										
9	S.E. of Coefficients															
10	Wichita	0.333	0.108	0.108	0.000	0.000										
11	Restriction Matrix															
12	Wichita	1	1	1	0	0										
13	Difference															
14	Character	Dickey-Fu	Aug. Dick	Schwarz	S.D. Resid	MAPE	AIC									
15	Wichita	-12.146	-9.921	1.227	2.900186	11.36	2.20									
16	Forecast	Impulse	Auto-	t-Statistic	Partial	t-Statistic										
17	Response	Correlatio	(AutoCorr)	AutoCorr	(Part. Aut)	Period										
18	32.051	1.000	-0.28853	-2.64439	-0.28853	-2.64439	1									
19	32.793	-0.369	-0.2054	-1.74304	-0.31486	-2.88577	2									
20	33.202	-0.179	0.176326	1.444937	0.009148	0.083843	3									
21	33.209	0.183	-0.12157	-0.97233	-0.14108	-1.29303	4									
22	33.479	-0.011	-0.08964	-0.70901	-0.15115	-1.3853	5									
23	33.780	-0.054	0.271286	2.133082	0.157704	1.445381	6									
24	33.983	0.023	-0.18847	-1.4076	-0.09713	-0.89017	7									
25	34.213	0.008	-0.01516	-0.11063	0.00145	0.013291	8									
26	34.465	-0.010	0.039689	0.289632	-0.0951	-0.87162	9									
27	34.700	0.001	-0.16265	-1.18571	-0.17961	-1.64619	10									





The time series output generated by Simetar is dynamic meaning that the beta coefficients in the AR model will update if you change the values in the original data or replace the input data array with another series of data. An added feature is the capability to impose restrictions on the initial AR model by dropping out/re-entering lags in real time. The Restriction Matrix has 1's

beneath each lags' coefficient. When the restriction value of 1 is changed to 0 the model is re-estimated without that particular variable or lag. The example AR model in DemoSimetar-Ar was run with 4 lags so the user can experiment with deleting unnecessary lags using the Restriction Matrix. When the 2nd, 3rd and 4th lags are restricted out the standard deviation for the residuals increases slightly from 2.86 to 3.06. As these higher order lags are removed the MAPE increases only about 1.3 percentage points. The AIC is minimized when lags 3 and 4 are removed.

Note that the initial number of lags and differences specified for the AR model determines the number of observations used to estimate the coefficients. When an AR model of 1st differenced data is estimated with four lags initially but the 3rd and 4th lags are restricted out, the resulting coefficients will not equal those for an AR(1) model estimated with two lags. The reason the coefficients are slightly different is that the latter model uses two more observations to estimate the parameters. It is recommended that the restricted AR model be re-estimated using the exact number of lags once the restricted model is acceptable.

As the restrictions on the lags are imposed on the unrestricted model the following test statistics do not change: Dickey-Fuller Test, Augmented Dickey-Fuller Test, and Schwarz Test (see example above). These statistics do not change because they reflect the number of differences specified for the unrestricted model. For example, the Dickey-Fuller Test statistic for an AR(4,1) model is calculated as =DF(data,,,1) and for an AR(4,2) model it is =DF(data,,,2). The Schwarz Test statistic is based on the number of differences [=ARSCHWARZ(data,,No. of Differences)] and does not change as the number of lags is restricted.


It is possible to interactively analyze the impact of changing the number of differences to the data in the AR model. In the second row of the Restriction Matrix (see the example above) is the word Differences followed by a value, in this case 1. The 1 in the Difference row means the data have been differenced once. To "re-run" the model with second differenced data, type a 2 into the restriction matrix in place of the 1. This change causes Simetar to re-estimate all of the parameters and update the goodness of fit test statistics.

The predicted values over the historical period and their residuals are provided for the AR model. The residuals are also expressed as a fraction of the predicted data. The predicted values and the residuals begin with observation 6, for this example, because the lag/difference structure of an AR(4,1) model uses the first 5 observations.

A graph of the historical and predicted values for the data series is generated by the Time Series function. The thin line represents the original data while the bold line represents the predicted values. Projections beyond the historical data in the graph correspond to the 10 period forecast requested in the dialog box (Figure 43).

A graph of the Impulse Response Function is also included in the forecast. The impulse response values are included in the output, but they are easier to see in the graph. A stationary model will exhibit continuously decreasing impulse responses to a 1 unit change at the outset of the period, as depicted by the graph in the AR Worksheet. The Impulse Response Function graph changes as the lags in the model are restricted out. Not shown in example above are the autocorrelation and partial autocorrelation function graphs for the AR model.

15.6 Estimating and Forecasting Vector Autoregressive (VAR) Models

The Time Series Analysis Engine dialog box (Figure 50) can be used to estimate and forecast VAR models. VAR model analyses begin by selecting the  icon on the Simetar toolbar. To estimate a VAR model, take all the steps used to estimate an AR model with one exception, specify two or more adjacent series in the *Data Series* menu (Figure 50). When two or more data series are specified, Simetar uses the more general estimation procedure for a VAR. The number of lags and differences should be specified based on prior analyses and tests.

The results of estimating and forecasting a two variable unrestricted VAR model are presented in the VAR worksheet of the Time Series Demo.xls Workbook and below. The Time Series function estimated the parameters for the VAR model using 4 lags and 1 difference with a constant, so 18 parameters are presented in the results. Various time series tests statistics for the model are presented below the parameters.

The first and second rows of the Restriction Matrix contain 1's indicating all lags are initially in the model. These restriction values can be changed to 0's to re-fit the VAR in real time by selectively deleting lags for one or both of the variables (see the example below). Changing the 1's to 0's and observing the change in the test statistics will enable the user to instantly experiment with a large number of model specifications. The interaction among the variables and their lags can be tested interactively using this feature in the Simetar VAR. The third row in the Restriction Matrix provides the switch to re-fit the VAR model with alternative numbers of differences, in real time.

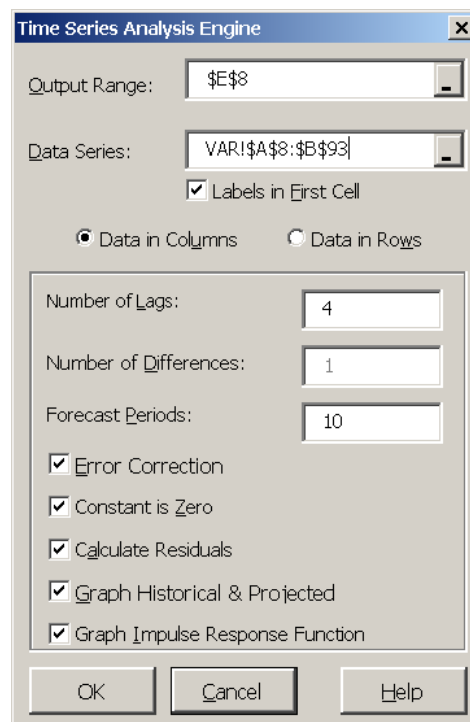
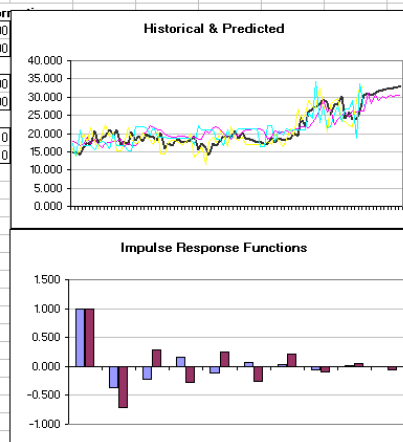


Figure 50. Time Series Analysis Dialog Box for a VAR

	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
8	VAR Series Analysis Results for 4 Lags & 1 Difference with Error Correction, 12/18/2004 10:28:14 PM																	
9		Constant	WichitaL	WichitaL	WichitaL	WichitaL	SalinaL	SalinaL	SalinaL	SalinaL	SalinaL	SalinaL	SalinaL	SalinaL	SalinaL	SalinaL	SalinaL	SalinaL
10	Wichita	0.431	-0.392	-0.392	-0.092	-0.175	0.094	0.004	-0.047	-0.015	0.000							
11	Salina	0.272	0.044	0.121	0.097	0.092	-0.779	-0.334	-0.224	-0.009	0.000							
12	S.E. of Coefficients																	
13	Wichita	0.345	0.118	0.133	0.140	0.125	0.123	0.159	0.157	0.136	0.000							
14	Salina	0.333	0.114	0.128	0.136	0.121	0.119	0.153	0.151	0.131	0.000							
15	Restriction Matrix																	
16	Wichita	1	1	1	1	1	1	1	1	1	1	0						
17	Salina	1	1	1	1	1	1	1	1	1	1	0						
18	Difference	1																
19	Character Dickey-Fu Aug. Dick LRT LRT Critit S.D. Resi MAPE AIC SIC																	
20	Wichita	-12.146	-9.921	18.196	2.733	2.84	10.94	2.3148616	2.5828396									
21	Salina	-17.935	-9.176			2.75	9.68	2.2488177	2.5167957									
22	Forecast Impulse Response																	
23	Wichita	Salina	Wichita	Salina	Period													
24		31.121	30.950	1.000	1.000	1												
25		30.417	28.194	-0.373	-0.714	2												
26		30.980	30.800	-0.225	0.282	3												
27		31.790	28.960	0.158	-0.275	4												
28		31.974	30.248	-0.116	0.257	5												
29		32.120	29.668	0.061	-0.252	6												
30		32.247	30.510	0.039	0.205	7												
31		32.456	30.163	-0.053	-0.099	8												
32		32.688	30.598	0.023	0.056	9												
33		32.918	30.524	-0.003	-0.057	10												




Forecasted values for both of the data series are provided in the output section. Impulse responses for the system of variables are also provided. These impulse response values are also summarized in a graph when requested.

Actual and predicted values over the historical period are presented in the top chart. Numbers behind the predicted values over the historical period are provided, beginning with period 6. The forecast values have a label with the word “Pred” following the variables name. Residuals for the VAR predicted values are also included in the output.

The residuals from the historical data can be used to simulate the unexplained variability or stochastic components of the random variables. Use the residuals to estimate the standard deviation about the forecasted values. Also use the residuals to estimate the correlation matrix for correlating random values about the forecasts.

16.0 Other Statistical and Data Analysis Functions

16.1 Summary Statistics

The dialog box used to calculate summary statistics for a variable (Figure 45) appears when the Summary Statistics menu item or icon  is selected. Select in the *Select Range* box and highlight the range (column or row) to analyze. Next click in the *Output Range* box and click the cell where the results are to be placed. All of the statistics and their names (mean, standard deviation, coefficient of variation, minimum, maximum, lower and upper confidence interval, and sum) will be placed in the worksheet starting with the *Output Range* cell if the *Add Output Labels* button is clicked. The standard deviation can be calculated using either the population or the sample formula. The coefficient of variation, sum, count and autocorrelation coefficient are not calculated unless these statistics are specified by selecting their boxes. Experiment with the dynamic nature of Simetar by changing the values in the original data and observing the updated summary statistics. See Data Analysis Demo.xls for an example.

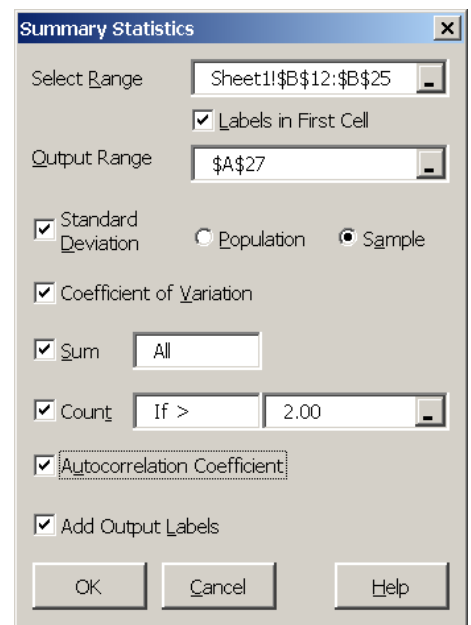


Figure 51. Summary Statistics Dialog Box.

The *Count* and *Sum* options in the Summary Statistics menu are available for conditional counts and sums of the data. Consider the situation where you have 2,500 observations and need to know how many values are less than or equal to 10.0. Perform this calculation by clicking on *Count*, followed by selecting the IF <= box, and then type the target value in the right hand box 10.0. The conditional count will appear with the other statistics.

16.2 Jackknife Estimator

Simetar provides a jackknife function which can be used to estimate parameters for any statistical formula or function in Excel or in Simetar. Given an n-dimensional vector or matrix of data and an associated statistic based on the data, the jackknife procedure sequentially re-estimates the statistic, leaving out the *i*th row at each iteration, where $i = 1, \dots, N$. These *n* statistics are then used to calculate the average statistic, the bias relative to the original statistic, and the jackknife variance of the statistic. The format for the =JACKKNIFE() function is:

```
=JACKKNIFE(DataRange, FormulaRef, RetVariance, Delete_D)
=JACKKNIFE(A2:B20,C2:D3)
```

Where: Data Range is a reference to a range of data that will be resampled to calculate the jackknife estimator. If this range is an $N \times 1$ vector, then the estimator will be calculated based on sequentially removing the *i*th row of the vector, where $i = 1, \dots, N$. Similarly, if this range is an $N \times K$ matrix, the estimator will be calculated based on sequentially removing the *i*th row of the matrix. Thus, multivariate data should be arranged with variables in columns,

FormulaRef is a reference to a range or cell that contains a formula which calculates an estimate based on the given Data Range. The jackknife estimator will be an average of the result of this formula based on the sequentially re-sampled data,

RetVariance is an optional term to include if only the jackknife estimate of the estimator variance is desired. A value of TRUE (or 1) will produce only the variance. A value of FALSE (or 0) will produce the jackknife estimator, bias, and variance. The default value is FALSE, and


Delete_D is an optional term to include if *D* rows are to be deleted at a time instead of one, where *D* is a positive integer less than *n*, the number of rows. The JACKKNIFE Function will then estimate statistics based on removing *D* adjacent rows at a time sequentially. This method is recommended when dealing with nonlinear statistics and should be used in conjunction with random sub-sampling methods. The default is one.

17.0 Function Evaluation

Two Simetar functions are available for evaluating user specified nonlinear functions. The first, =OPT(), finds the minimum or maximum of a function given boundary constraints on the control variables. The =OPT() function can also be used to find the value of *X* when a function equals a target value, as zero. The second function, =RINTEGRAL() integrates a function over a given range. Both functions provide approximate answers using efficient optimal control search and solve algorithms. The level of precision can be increased, but at a slight cost of longer execution times.

17.1 Optimize a Function

The =OPT() function uses the Golden Section method for optimizing a non-linear function specified by the user. The function to optimize (maximize or minimize) can be either typed into the =OPT() function as a literal or as an equation typed into a cell. Optimization Function Demo.xls demonstrates both techniques for optimizing functions.

The easiest method for using the =OPT() function is to type =OPT and then click on Excel's Equation editor, , and fill in the blanks in the OPT equation editor form (Figure 52). The optimization function parameters are:

=OPT (Formula, Constraint Type, Change Variable, Lower Guess, Upper Guess, Max Iterations, Precision)

where: Formula is the function to be optimized, as: $= 100 - 25X + 45X^2$ and must be typed into the referenced cell as a formula,
 Constraint Type must be typed as the word "Min" or "Max" for minimization or maximization, respectively,
 Change Variable is the cell which refers to the X variable in the function and can be any feasible value of X,
 Lower Guess is the minimum X,
 Upper Guess is the maximum X,
 Max Iterations is the maximum number of calculation cycles to use, and
 Precision is the degree of accuracy, such as 0.000001.

The value of X which causes the Y function to be optimized will appear in the cell where =OPT() is typed. Changing the parameters will cause Excel to calculate a new optimal value if the current solution is at a boundary or more precision can be obtained. Changing the function or input values to the function of course changes the =OPT() answer.

17.2 Value of a Function

Given a complex polynomial function that can be programmed in a cell as $Y = f(X)$, Simetar can solve for the value of X where Y equals a target value such as zero. A variation on the =OPT() function can be used to solve this type of optimization problem. The parameters for the function are:

=OPT (Formula, Target Value, Change Variable, Initial Guess, Upper Bound, Max Iterations, Precision)

where: Formula is the cell reference for the function to be optimized,
 Target Value is the value of Y when the function is optimized,
 Change Variable is the cell referring to the X variable in the function and can be any feasible value of X,
 Initial Guess is the lower bound constraint of X,

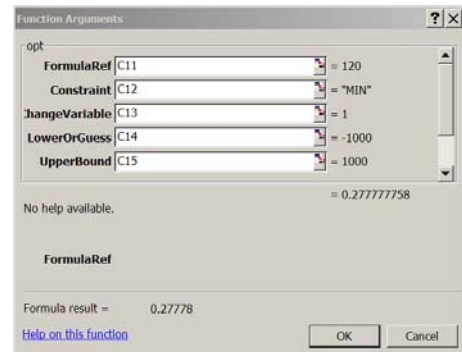


Figure 52. Equation Editor for the Optimization Function =OPT().

Upper Guess is the upper bound constraint of X,
 Max Iterations is the maximum number of calculation cycles, and
 Precision is the degree of accuracy, such as 0.000001.

When the =OPT() function fails to find the target value for Y over the range of the function it returns #VALUE! in the cell where =OPT() is typed. In this case, try another initial guess, the upper bound, the level of precision or the maximum number of iterations. Excel will solve some functions very fast; for example, $Y = X^4$ will find that Y equals 23 at 2.1958 very rapidly. See Optimization Function Demo.xls for this example.

17.3 Integral of a Function

A function can be integrated over a specified range using the =RINTEGRAL() function. This function provides an approximate value for the integral using Riemann Integration.

The level of precision can be increased by increasing the number of partitions. The easiest way to use the function is to develop a table of parameters and then use Excel's Equation editor after typing =RINTEGRAL, as depicted in Figure 53. An example of integrating a function $Y = 100 - 25X + 45X^2$ over the interval of X equal 0 to 100 is provided in Optimization Function Demo.xls. The parameters for the integration function are:

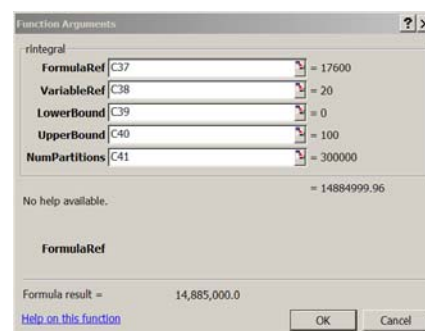



Figure 53. Equation Editor for the Integral Function.

=RINTEGRAL (Formula, Variable Ref, Lower Bound, Upper Bound, Partitions)

where: Formula is the cell reference to the equation to be integrated,
 Variable Ref is the cell reference for the independent variable (X) in the equation,
 Lower Bound is the minimum X for the range of the integration,
 Upper Bound is the maximum X for the range of the integration, and
 Partitions is the number of intervals X range is partitioned into for integration.

The answer will appear in the =RINTEGRAL() function cell. It is recommended that you increase the number of partitions until the change in the integral answer is zero. As you increase the number of partitions, response time will slow. For the example in Optimization Function Demo.xls the true value of 14,885,000 is reached at 300,000 partitions in about 25 seconds.

18.0 Getting Help with Simetar

Simetar Help is provided in two forms: detailed description of the functions and equation editing help. Detailed descriptions are available for all of the Simetar functions by clicking the help icon  on the toolbar. When the help icon is selected the Help Index for Simetar window (Figure 54) appears on the screen. Scroll down to the function of interest and click on the function name. This action results in the requested Simetar Help screen appearing on the screen.

An example of requesting help from the Simetar Help Index for the =NORM() function is

displayed in Figure 55. The help provided in the screen is designed to supplement the material in this manual. You can either print the help screen, return to the Help Index, or close the help screen by clicking on the appropriate button at the bottom of the Help Screen. Additional help on the function is available from Simetar by clicking on the line “Help on the function” as demonstrated for the =NORM() function below.


Excel provides pop-up help menus to assist with writing or editing equations. To access help for equation programming simply type “equal and the function name” in a cell and then click the “=” button or the  icon on the formula bar. An example of how this works for getting help with the =CSND() function is provided in the worksheet example below and Figure 56. In the example the analyst has highlighted three cells (B7:B9) in preparation for using the CSND function as an array. After typing “=CSND” click the “**fx**” button in the formula toolbar at the top of the worksheet and Excel will place a dialog box like Figure 50 on the screen.



Figure 54. Help Index Dialog Box.

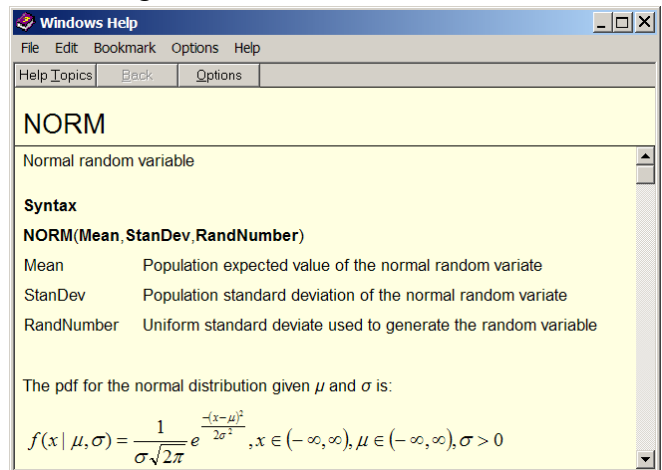


Figure 55. Example of a Simetar Help Screen.

The equation help box in Figure 50 indicates the order of parameters for the =CSND() function and the names of the parameters. You can fill in the worksheet cell locations for the parameters by clicking the miniature grid to the right of each parameter and painting the appropriate cells with the mouse. After filling in values for the parameters select OK.

The equation editor help function can be used to develop new equations and to de-bug existing equations. Select a cell with an existing equation and click the “=” or “**fx**” button on the formula bar to see the equation editing help box. Equation editing help screens are available for all Simetar and Excel functions.

	A	B	C	D
1	Correlation Matrix			ISNDs
2	1	0.25	-0.5	0.550576
3		1	0.3	0.708632
4			1	0.894564
5	Generate 3x1 array of CSNDs			
6		=csnd()		
7				
8				

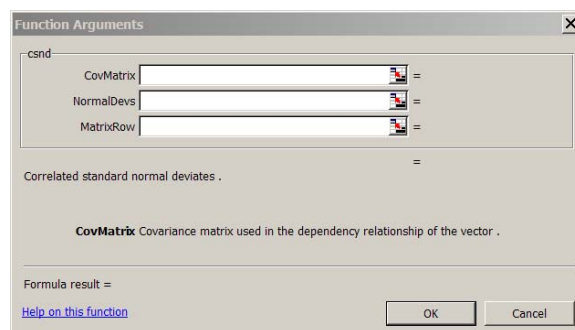


Figure 56. Example of the Equation Help Box.

19.0 Solutions to Problems in Simetar Application

Like all computer programs Simetar 2006 is the result of many enhancements. Each time one function is complete we find two more to add and in the process a better way to do the first function is developed. The program has come a long way given that it began in May 2000. Simetar continues to grow and become more useful.

Most problems are associated with installing Simetar on computers with old operating systems/versions of Excel and operators without administrative privilege. The optimal environment is Windows 2000 operating system with Microsoft Office XP. The Demo programs were developed in this environment. The first time you open one of the Demo workbooks it may warn you of embedded macros – select “Enable Macros” and proceed. Next your Excel may warn you that the Demo has external links – select “No” and proceed. Save the workbook to your hard drive and the next time it is opened you will not have link warnings. The workbook link warnings are caused by your computer storing Simetar in a different location than the developer’s computer. Excel will update the links on its own.

This section documents errors we have observed. Most of the problems occur because Excel’s Calculation is set to Manual or the Operating System burps and sets Calculation to Manual during your Excel session. **Set Calculation to Automatic and leave it there and check it if errors occur.**

19.1 My program was working when I saved it, but now the Simetar functions have #NAME

Sometimes Simetar and Excel gets confused and you need to remind Excel that Simetar is loaded, to do this follow these steps:

- Tools > Add-Ins > Uncheck the box for Simetar
- Then repeat the process
- Tools > Add-Ins > Check the box for Simetar

19.2 File Not Found dialog box with a file name of “PBJ.XLA” listed, appears when I open a workbook

Click the “Cancel” button and Excel will update the links to Simetar and PBJ using the current location of these files on your computer. This error occurs when the workbook was created on a different computer. Save the workbook and the next time it is opened there will be no problem.

19.3 Simetar Functions returns #NAME! instead of values

Sometimes Simetar and Excel gets confused and you need to remind Excel that Simetar is loaded, to do this follow these steps:

- Tools > Add-Ins > Uncheck the box for Simetar
- Then repeat the process
- Tools > Add-Ins > Check the box for Simetar

If your computer is running Excel 97, load Service Pack 2. If your computer is running Excel 95, get a newer version of Excel.

19.4 Scenario names in Stochastic Dominance tables appear as #NUM!

Press Function key F9

Set calculation to automatic by following these steps:

- Tools > Options > Calculation, set the calculation option to Automatic

19.5 Statistics for the first stochastic variable in SimData Worksheet appear as #DIV/0!

Press Function key F9 if the problem goes away, do the following:

- Check Tools > Options > Calculation, set the Calculation option to Automatic

Check if the variable is a constant. If it is then the means will not be zero but the standard deviation and coefficient of variation will be #DIV/0!

19.6 Values for SERF table and chart in SERFTbl1 do not change when you change the ARACs or the utility function

Check Tools > Options > Calculation, set the Calculation option to Automatic

19.7 Results from Testing a Single Variable for Normality returns #VALUE! in place of values

Check Tools > Options > Calculation, set the Calculation option to Automatic

Delete the formats in the cells for the output range that may be left over from previous sessions.

19.8 Results of Compare Two Data Series returns #DIV/0! and #NUM! in place of values

Check Tools > Options > Calculation set the Calculation option to Automatic

19.9 Multiple regression returns #DIV/0! for standard deviation of residuals and/or MAPE is #VALUE!

Check Tools > Options > Calculation, set the Calculation option to Automatic

19.10 Multiple regression does not update the beta hats and goodness of fit statistics when a restriction value is changed

Check Tools > Options > Calculation, set the Calculation option to Automatic

19.11 Multiple regression does not update the beta hats and goodness of fit statistics when one of the X or Y observations is changed

Check Tools > Options > Calculation, set the Calculation option to Automatic

19.12 Multiple regression, time series, and other menu enabled functions return numbers instead of the names for the X and/or Y variables

The dialog boxes allow you to enter “Labels in First Cell,” you did not include the label in the first cell, so Simetar used the first observation as the name of each X variable and/or for Y. Include the variables label when dialog boxes are used to enter data for functions.

19.13 Time series (AR and/or VAR) procedure returns #VALUE! instead of the coefficients

Check Tools > Options > Calculation, set the Calculation option to Automatic

19.14 Stochastic variables (cells) in the worksheet do not change when the Enter or F9 Keys are pressed

Check Tools > Options > Calculation, set the Calculation option to Automatic

Check the Simetar Toolbar to see if worksheet sampling has been set to “Expected Value”, if so click the Expected Value button on the Simetar Toolbar.

19.15 Stochastic variables (cells) in the worksheet are fixed at zero or the mean and do not change when F9 is pressed

The “Expected Value” button in the Expected Value button on the Toolbar is turned on. Turn the option off by clicking on the Expected Value button.

19.16 The CDF or SERF chart has numbers instead of names on the lines and/or the scenario names in the legend are numbers

The “Labels in First Cell” option was turned on so the program used the first observation for each scenario as the scenario names.

Be sure that the label in the first row starts with a letter, not a number, as 1998 or 2000.

19.17 Results and calculations in the simulation output worksheet, SimData, are gone

Simetar writes the iteration results to worksheet SimData after each run. It uses as many columns of the worksheet required for the output variables in the Simulation Engine. If you had tables from a previous simulation run in the columns needed for the current run, they got over written. When you place summary tables, tests, or chart data in SimData, rename the worksheet so it will be protected from the next simulation.

19.18 Simulation used to run fast and now it has slowed down

- Another Excel workbook which contains stochastic variables may be open. When Simetar simulates the stochastic variables in the open workbook, Excel also simulates the workbooks that are minimized.
- The SERF option is dynamic and can slow the simulation down if the model is simulating more than 500 iterations and SERF is tied to the SimData worksheet.
- Simulation can be slowed down if the SimData output is being used to calculate a large number of CDF and PDF charts.
- The number of Key Output Variables that Simetar is collecting for statistical analysis may have been expanded from previous runs.
- The number of Scenarios is greater than previous runs.
- The SimSolver option in the Simulation Engine is turned on.

19.19 A Simetar matrix or array function returns a single value when you expected an array or matrix of answers.

Press F2 to edit the function; if it is typed correctly press three keys: *Control Shift Enter*. Any time an array function is used, you MUST end by pressing these three keys: *Control Shift Enter*.

19.20 Hypothesis Test statistics appear wrong.

- Re-do the test and be careful to indicate no labels in the first row and only include the data.
- Change the variable labels or names so they begin with a letter, as Y1988, not 1988 and re-do the test.
- The t-tests are two tailed tests, so they will not be the values you expect for a one-tailed test.

19.21 After installation, if the Excel Tool Bar does not show “Simetar,” it can be re-loaded to the toolbar using the following steps:

Tools > Select Add-Ins ... > scroll down and click the box for Simetar > OK

20.0 List of All Simetar Functions

Following is a list and short description of all functions in Simetar:

Function Name	Description
· ANOVA	One way analysis of variance
· ARLAG	Recommends the number of lags in an autoregressive model
· ARSCHWARZ	Schwarz criterion associated with recommended number of lags
· AUTOCORR	Autocorrelation function for a univariate time series
· BANDWIDTH	Bandwidth function in kernel density estimation
· BERNOULLI	Bernoulli random variable
· BERNOULLIDIST	Bernoulli distribution function
· BINOMINV	Binomial random variable
· BLOCKIT	Column Vector to a Matrix
· BOOTSTRAPPER	Bootstrap resampling of a univariate or multivariate series
· BOXCOX	Box-Cox transformation of a data series for normalization
· BOXCOXEXP	Estimate of the Box-Cox exponent in a Box-Cox transformation
· BOXM	Box's M statistic for testing multivariate variances
· CAUCHY	Cauchy random variable
· CAUCHYDIST	Cauchy distribution function
· CDFDEV	Indicate goodness of fit between sample data & known distribution data
· CELLSUB	Replace an item or items in a block of data
· CERTEQ	Certainty equivalent of a data series assuming a utility function
· CMOVAVG	Centered moving average
· CONCAT	Concatenate two or more matrices
· COSDIST	Cosine distribution function
· COSINV	Cosine random variable
· CSND	Correlated standard normal deviates
· CUSD	Correlated uniform standard deviates
· DELNUM	Remove the numbers from a string of text and numbers
· DELTEXT	Remove the text from a string of text and numbers
· DEMPIRICAL	Discrete empirical distribution random variable
· DEXPONDIST	Double exponential distribution function
· DEXPONINV	Double exponential random variable
· DF	Dickey-Fuller test statistic
· DIRICHINV	Dirichlet random variable
· EDF	Empirical distribution function
· EMP	Empirical random variable
· EMPCOPULA	Empirical copula function
· EMPIRICAL	Empirical random variable
· EPANDIST	Epanechnikov distribution function
· EWMA	Exponentially weighted moving average
· EXPONINV	Exponential random variable
· EXTVALDIST	Extreme value distribution function
· EXTVALINV	Extreme value random variable
· GEOMDIST	Geometric distribution function
· GEOMINV	Geometric random variable
· GMDIF	Gini's mean difference
· GRK	GRK random variable
· GRKS	GRKS random variable
· GRKSDIST	GRKS distribution function
· GUMBELDIST	Gumbel distribution function
· GUMBELINV	Gumbel random variable
· HOTELLTDIST	Hotelling T-squared distribution function
· HOTELLTINV	Hotelling T-squared random variable
· HYPERGEOMINV	Hypergeometric random variable
· IMPULSE	Impulse response function in a vector autoregression
· INVGAUS	Inverse Gaussian random variable
· INVGAUSDIST	Inverse Gaussian distribution function
· IQR	Inner quartile range of a sample
· ITERATION	Show the iteration number during simulation
· ITERSUM	Sum a value across iterations during a simulation
· JACKKNIFE	Jackknife estimate of statistic, bias, and variance

· KDEINV	Random variable based on a kernel density estimate
· KTAU	Kendall's Tau measure of concordance
· LOGISTICDIST	Logistic distribution function
· LOGISTICINV	Logistic random variable
· LOGIT	Logit binary response regression
· LOGLOGDIST	Log-log distribution function
· LOGLOGINV	Log-log random variable
· LOGLOGISTICDIST	Log-logistic distribution function
· LOGLOGISTICINV	Log-logistic random variable
· LR	Linear regression (OLS)
· LRAIC	Akaike information criterion for a regression
· LRBIG	Linear regression (OLS) for large data sets
· LRDFBETA	Observational diagnostics for a regression
· LRDHATMAT	Diagonal of the hat matrix
· LRDW	Durbin-Watson test statistic in a regression
· LREGLS	Estimated generalized least squares (EGLS)
· LRGLS	Generalized least squares (GLS)
· LRGQ	Goldfeld-Quandt test statistic for a regression
· LROBS	Regression observation count and degrees of freedom
· LRPARTCORR	Partial correlation function in a regression
· LRRESID	Residuals and predicted values in a regression
· LRRHO	Autocorrelation coefficient in the errors of a regression
· LRRIDGE	Ridge regression
· LRSEMICORR	Semi-partial correlation function in a regression
· LRSIC	Schwarz information criterion for a regression
· LRT	Likelihood ratio test in univariate or multivariate autoregression estimation
· LRVIF	Variance inflation factor for a regression
· LRWLS	Weighted least squares (WLS)
· MAE	Mean absolute error
· MAHANGLE	Mahalanobis angle of a data matrix
· MAPE	Mean absolute percent error
· MCENTER	Centering matrix of a specified dimension
· MCHOL	Choleski factorization of an $n \times (n+p)$ matrix, ()
· MCOFACTOR	Cofactor of a square matrix
· MCOR	Correlation matrix
· MCOV	Covariance matrix
· MDAPE	Median absolute percent error
· MDET	Determinant of a square matrix
· MDIAG	Diagonalize a vector or matrix
· MDIST	Squared Mahalanobis distance of two data matrices
· MEDAVG	Median average
· MEQCORR	Equicorrelation matrix of a specified dimension
· MEVAL	Eigenvalues of a square matrix
· MEXP	Exponential power of a matrix
· MGINVERSE	Generalized inverse of a matrix
· MIDEN	Identity matrix
· MINV	Inverse of a square matrix
· MIP	Inner product of two matrices
· MJ	Matrix of 1s
· MKRON	Kronecker multiply two matrices
· MLEBETA	Beta MLE of parameter(s)
· MLEBINOM	Binomial MLE of parameter(s)
· MLEDEXPON	Double Exponential MLE of parameter(s)
· MLEEXPON	Exponential MLE of parameter(s)
· MLEGAMMA	Gamma MLE of parameter(s)
· MLEGEOM	Geometric MLE of parameter(s)
· MLELOGISTIC	Logistic MLE of parameter(s)
· MLELOGLOG	Log-Log MLE of parameter(s)
· MLELOGLOGISTIC	Log-Logistic MLE of parameter(s)
· MLELOGNORM	Lognormal MLE of parameter(s)
· MLENEGBIN	Negative Binomial MLE of parameter(s)
· MLENORM	Normal MLE of parameter(s)
· MLEPARETO	Pareto MLE of parameter(s)
· MLEPOISSON	Poisson MLE of parameter(s)

·	MLEUNIFORM	Uniform MLE of parameter(s)
·	MLEWEIB	Weibull MLE of parameter(s)
·	MNORM	Norm of a matrix
·	MOMBETA	Beta MOM of parameter(s)
·	MOMBINOM	Binomial MOM of parameter(s)
·	MOMDEXPON	Double Exponential MOM of parameter(s)
·	MOMEXPON	Exponential MOM of parameter(s)
·	MOMGAMMA	Gamma MOM of parameter(s)
·	MOMGEOM	Geometric MOM of parameter(s)
·	MOMLOGISTIC	Logistic MOM of parameter(s)
·	MOMLOGLOG	Log-Log MOM of parameter(s)
·	MOMLOGLOGISTIC	Log-Logistic MOM of parameter(s)
·	MOMLOGNORM	Lognormal MOM of parameter(s)
·	MOMNEGBIN	Negative Binomial MOM of parameter(s)
·	MOMNORM	Normal MOM of parameter(s)
·	MOMPARETO	Pareto MOM of parameter(s)
·	MOMPOISSON	Poisson MOM of parameter(s)
·	MOMUNIFORM	Uniform MOM of parameter(s)
·	MOMWEIB	Weibull MOM of parameter(s)
·	MORTH	Orthogonalize a matrix
·	MOVAVG	Moving average
·	MPROD	Multiply two or more conformable matrices
·	MRANK	Rank of a matrix
·	MRECH	Row Echelon Form of a matrix
·	MRRECH	Reduced row echelon form of a matrix
·	MSE	Mean squared error
·	MSQRT	Factor a square, symmetric matrix
·	MSTACK	Stack two or more matrices
·	MSVD	Singular value decomposition of a matrix
·	MSWEEP	Sweep a square matrix on a diagonal element
·	MTOEP	Column vector to a Toeplitz matrix
·	MTPNORM	Modified two-piece normal random variable
·	MTPNORMDIST	Modified two-piece normal distribution function
·	MTRACE	Trace of a square matrix
·	MULTINOMDIST	Multinomial distribution function
·	MULTINOMINV	Multinomial random vector
·	MULTSORT	Sort a matrix by a specified column
·	MVCHT	LRT for complete homogeneity of multiple data matrices
·	MVCV	Multivariate coefficient of variation
·	MVEMP	Multivariate empirical random vector
·	MVEMPIRICAL	Multivariate empirical random vector
·	MVEPANDIST	Multivariate Epanechnikov distribution function
·	MVLOGNORM	Multivariate lognormal random vector
·	MVNORM	Multivariate normal random vector
·	MVNORMDIST	Multivariate normal distribution function
·	MVPDENSITY	Percentile based on a multivariate kernel density estimator
·	MVTINV	Multivariate student's t random variable
·	NEGBINOMINV	Negative binomial random variable
·	NORM	Normal random variable
·	NORMAD	Anderson Darling statistic for test of normality
·	NORMCHI	Chi-squared statistic for a test of normality
·	NORMCVM	Cramer von Mises statistic for test of normality
·	NORMKS	Kolmogorov Smirnov statistic for test of normality
·	NORMSW	Shapiro-Wilks statistic for test of normality
·	OPT	Find an iterative optimum solution
·	PARETO	Pareto random variable
·	PARETODIST	Pareto distribution function
·	PAUTOCORR	Partial autocorrelation function for a univariate time series
·	PDENSITY	Percentile based on a Kernel density estimator
·	PERTDIST	Project evaluation and review technique (PERT) distribution function
·	PERTINV	Project evaluation and review technique (PERT) random variable
·	PNORM	Power normal random variable
·	PNORMDIST	Power normal distribution function
·	POISSONINV	Poisson random variable

·	PROBIT	Probit binary response regression
·	QUANTILE	Find the quantile of an empirical CDF given the probability
·	RANDSORT	Randomly sort a vector
·	RANDWALK	Generate a random walk series
·	RANKCORREL	Rank correlation of two data series
·	REVERSE	Reverse the order of a vector
·	RINTEGRAL	Riemann integral of a bounded function
·	RMSE	Root mean squared error
·	RUSD	Rank correlation matrix
·	SCENARIO	Return a value associated with different scenarios in a simulation
·	SEMICIRCDIST	Semicircle distribution function
·	SEMICIRCINV	Semicircle random variable
·	SEQ	Sequence of numbers
·	SIMETARCR	Returns copyright information for Simetar
·	STRETCHIT	Matrix to a vector
·	TEMPIRICAL	Truncated empirical random variable
·	TGAMMADIST	Truncated gamma distribution function
·	TGAMMAINV	Truncated gamma random variable
·	THEILU2	Theil's U2 statistic for forecasts
·	TNORM	Truncated normal random variable
·	TNORMDIST	Truncated normal distribution function
·	TPNORM	Two-piece normal random variable
·	TPNORMDIST	Two-piece normal distribution function
·	TRANS	Transpose a matrix
·	TRIANGLE	Triangle random variable
·	TRIANGLEDIST	Triangle distribution function
·	TSDECOMP	Time series decomposition
·	TWEIBDIST	Truncated Weibull distribution function
·	TWEIBINV	Truncated Weibull random variable
·	TWOSLS	Two stage least squares (2SLS)
·	UNBOXCOX	Convert a Box-Cox transformed value back to the original level
·	UNIFORM	Uniform random variable
·	UNIFORMDIST	Uniform distribution function
·	USND	Uncorrelated standard normal deviate
·	UUSD	Uncorrelated uniform standard deviate
·	VARAIC	Akaike information criterion in univariate or multivariate autoregression models
·	VAREST	Univariate or multivariate autoregression estimation function
·	VARLRT	Likelihood ratio test in univariate or multivariate autoregression estimation
·	VARRESID	Predictions & residuals in univariate or multivariate autoregression models
·	VFORMULA	View the formula in the referenced cell
·	WAPE	Weighted absolute percent error
·	WBNAME	Return the name of the workbook
·	WEIBDIST	Weibull distribution function
·	WEIBINV	Weibull random variable
·	WILKSLDIST	Approximate cdf of the Wilks' Lambda random variable
·	WILKSLINV	Wilks Lambda random variable
·	WISHDIST	Wishart distribution function
·	WISHINV	Wishart random matrix
·	WSNAME	Return the name of the worksheet

21.0 Cross Reference of Functions and Demonstration Programs

<u>Topic</u>	<u>Demonstration Program Name</u>
ANOVA	Data Analysis Tools Demo.xls
ANOVA test	Hypothesis Tests Demo.xls
AR and VAR models estimated	Time Series Demo.xls
AR model dynamic probabilistic forecast	Time Series Forecasting Demo.xls
AR model estimation	Time Series Functions Demo.xls
AR model estimation	Time Series Analysis Tools Demo.xls
ARLAG function	Time Series Functions Demo.xls
ARSCHWARZ function	Time Series Functions Demo.xls
AUTOCORR function	Time Series Functions Demo.xls
Additive seasonal decomposition forecasting with cycle	Seasonal Decomposition Forecasts Demo.xls
Additive seasonal decomposition forecasting without cycle	Seasonal Decomposition Forecasts Demo.xls
Amortize land debts	Farm Simulator Demo.xls
Amortize loans with monthly payments	Monthly Payments Demo.xls
Augmented Dickie Fuller test	Time Series Functions Demo.xls
Autocorrelation coefficients	Time Series Forecasting Demo.xls
Autocorrelation coefficients	Time Series Analysis Tools Demo.xls
Autocorrelation coefficients	Time Series Demo.xls
Autocorrelation test	Time Series Functions Demo.xls
BERNOULLI function application	Simulate Alternative Distributions Demo.xls
BOXCOX function	Data Analysis Tools Demo.xls
BOXCOXEXP function	Data Analysis Tools Demo.xls
Bad (singular) correlation/covariance matrix	Bad Correlation Matrix Demo.xls
Bernoulli distribution	Conditional Probability Distributions Demo.xls
Bernoulli distribution	Probability Distribution Demo.xls
Bernoulli distribution	Probability Distributions Demo.xls
Bernoulli distribution parameter estimation	Trend Regression to Reduce Risk Demo.xls
Beta distribution	Probability Distribution Demo.xls
Beta distribution	Simulate All Probability Distributions Demo.xls
Beta distribution	Probability Distributions Demo.xls
Bingo	Games of Chance Demo.xls
Binomial distribution	Simulate All Probability Distributions Demo.xls
Binomial distribution	Probability Distributions Demo.xls
Boot strap simulation	Simulate All Probability Distributions Demo.xls
Bootstrap for singular matrix	Bad Correlation Matrix Demo.xls
Bootstrap simulation	Probability Distributions Demo.xls
Bootstrapper distribution	Probability Distribution Demo.xls
Box plot chart of risky alternatives	Analysis of Simulation Results Demo.xls
Box's M test	Data Analysis Tools Demo.xls
Box-Cox transformation	Data Analysis Tools Demo.xls
Business model	Deterministic Demo.xls
Simplified Business model of net returns	Business Model with Risk Demo.xls
CDF chart of random variables	Analysis of Simulation Results Demo.xls
CDFDEV function	Univariate Parameter Estimator Demo.xls
CDFs for 12 distributions	Test Parameters Demo.xls
CV stationarity for Normal distributions	CV Stationarity Normal Demo.xls
CV stationarity for empirical distributions	CV Stationarity Empirical Demo.xls
Capital Investment Analyzer ©	Net Present Value Internal Rate of Return Demo.xls
Cauchy distribution	Probability Distribution Demo.xls
Cauchy distribution	Simulate All Probability Distributions Demo.xls
Cauchy distribution	Probability Distributions Demo.xls
Centering a matrix	Matrix Operation Tools Demo.xls
Centering matrix of size n	Matrices Demo.xls
Chi-Squared distribution	Probability Distribution Demo.xls
Chi-Squared distribution	Probability Distributions Demo.xls
Chi-Squared test	Data Analysis Tools Demo.xls
Choleski decomposition of a covariance matrix	Parameter Estimation Tools Demo.xls
Coin toss	Games of Chance Demo.xls
Column vector to a matrix	Matrix Operation Tools Demo.xls
Compare means and variance for multivariate distributions	Hypothesis Tests Demo.xls

Compare means and variance for univariate distributions	Hypothesis Tests Demo.xls
Compare means for two distributions -- ANOVA	Hypothesis Tests Demo.xls
Compare means for two series	Data Analysis Tools Demo.xls
Compare two data series	Analysis of Simulation Results Demo.xls
Compare two multivariate distributions	Validation Tests Demo.xls
Compare two tests -- t and F tests	Data Analysis Tools Demo.xls
Compare two univariate distributions	Validation Tests Demo.xls
Complete Homogeneity test	Data Analysis Tools Demo.xls
Concatenate data from two locations	Matrix Operation Tools Demo.xls
Concatenate two matrices	Matrices Demo.xls
Conditional distribution for simulating sales bonus	Conditional Probability Distributions Demo.xls
Conditional probability distributions	Conditional Probability Distributions Demo.xls
Confidence interval for seasonal index	Seasonal Analysis Demo.xls
Confidence intervals for Multiple Regression forecasts	Probabilistic OLS Forecasts Demo.xls
Convert a matrix to a vector	Matrices Demo.xls
Convert a vector to a matrix	Matrices Demo.xls
Corporate federal income taxes	Income Tax Demo.xls
Corporate income taxes	Farm Simulator Demo.xls
Correct for CV non-stationarity Normal distribution	CV Stationarity Normal Demo.xls
Correlated standard normal deviates	Probability Distributions Demo.xls
Correlated uniform standard deviates	Probability Distributions Demo.xls
Correlating Normal, Empirical, Uniform in a MV distribution	Multivariate Mixed Probability Distribution Demo.xls
Correlation matrix calculated	Data Analysis Tools Demo.xls
Correlation matrix t test of rho vs. zero	Data Analysis Tools Demo.xls
Correlation matrix test simulated vs. historical	Data Analysis Tools Demo.xls
Correlation matrix validation for MV distributions	Validation Tests Demo.xls
Correlation significance test	Hypothesis Tests Demo.xls
Correlation test of MVE method	Multivariate Empirical Distribution Demo.xls
Cosine distribution	Simulate All Probability Distributions Demo.xls
Cosine distribution	Probability Distributions Demo.xls
Cost of a project with risk	Project Management Demo.xls
Covariance matrix calculated	Data Analysis Tools Demo.xls
Covariance matrix estimation	Parameter Estimation Tools Demo.xls
Covariance matrix estimation	Matrix Operation Tools Demo.xls
Crop Insurance premium estimation	Insurance Premium Demo.xls
Cumulative distributions for ranking risky alternatives	Stochastic Dominance Demo.xls
Cycle length estimation	Probabilistic Cycle Forecasts Demo.xls
Cyclical decomposition of times series data	Exponential Smoothing Demo.xls
Cyclical decomposition of times series data	Moving Average Demo.xls
Cyclical index	Cyclical Analysis Tools Demo.xls
Cyclical index	Exponential Smoothing Demo.xls
Cyclical index	Moving Average Demo.xls
Cyclical index	Seasonal Analysis Demo.xls
DELNUM function	Data Analysis Tools Demo.xls
DELTEXT function	Data Analysis Tools Demo.xls
DEMPIRICAL function application	Simulate Alternative Distributions Demo.xls
DF Betas	Parameter Estimation Tools Demo.xls
DF function	Time Series Functions Demo.xls
DF function	Time Series Analysis Tools Demo.xls
Decomposition forecasting	Seasonal Decomposition Forecasts Demo.xls
Decomposition forecasts	Seasonal Index Forecasts Demo.xls
Decomposition of a time series	Cyclical Analysis Tools Demo.xls
Delivery time and inventory management	Inventory Management Demo.xls
Determinant of a square matrix	Matrices Demo.xls
Determinate of a square matrix	Matrix Operation Tools Demo.xls
Deterministic farm model	Deterministic Demo.xls
Deterministic simulation NPV and IROR	Net Present Value Internal Rate of Return Demo.xls
Deterministic simulation model	Cotton Model Demo.xls
Dice	Games of Chance Demo.xls
Dickie Fuller (DF) test	Time Series Functions Demo.xls
Dickie Fuller test	Time Series Forecasting Demo.xls
Dickie Fuller test	Time Series Analysis Tools Demo.xls
Dickie Fuller test	Time Series Demo.xls
Discrete empirical distribution	Probability Distributions Demo.xls

Discrete uniform distribution	Probability Distribution Demo.xls
Discrete uniform distribution	Simulate Alternative Distributions Demo.xls
Discrete uniform distribution -- numbers and names	Simulate All Probability Distributions Demo.xls
Double exponential distribution	Simulate All Probability Distributions Demo.xls
Dummy variables in Multiple Regression for seasonal analysis	Regression for Seasonal Forecasts Demo.xls
Dynamic forecast of AR model	Time Series Forecasting Demo.xls
E Factors to control heteroskedasticity	Heteroskedasticity Demo.xls
EMP function	Empirical Distribution Demo.xls
EMP function application	Simulate Alternative Distributions Demo.xls
EMP icon for estimating parameters	Multivariate Empirical Distribution Demo.xls
Econometric model for soybeans	Soybean Model Demo.xls
Econometric model for wheat	Wheat Sim Solve Demo.xls
Econometric stochastic model	Soybean Model Demo.xls
Econometric wheat model	Wheat Model Demo.xls
Eigenvalues for a square matrix	Matrix Operation Tools Demo.xls
Eigenvalues for a square matrix	Matrices Demo.xls
Empirical distribution	Probability Distribution Demo.xls
Empirical distribution	Probability Distributions Demo.xls
Empirical distribution -- actual data	Empirical Distribution Demo.xls
Empirical distribution -- actual data w/ CV stationary	CV Stationarity Empirical Demo.xls
Empirical distribution -- deviations from mean	Empirical Distribution Demo.xls
Empirical distribution -- deviations from trend	Empirical Distribution Demo.xls
Empirical distribution -- differences from mean	Empirical Distribution Demo.xls
Empirical distribution -- general and direct	Simulate All Probability Distributions Demo.xls
Empirical distribution -- percent deviates from mean	CV Stationarity Empirical Demo.xls
Empirical distribution parameter estimation	Trend Regression to Reduce Risk Demo.xls
Empirical distribution using interpolation	Empirical Distribution Demo.xls
Empirical distribution using inverse transform method	Inverse Transform Demo.xls
Empirical parameter estimation using actual data	Parameter Estimation Tools Demo.xls
Empirical parameter estimation using deviates from the mean	Parameter Estimation Tools Demo.xls
Empirical parameter estimation using deviates from trend	Parameter Estimation Tools Demo.xls
Empirical parameter estimation using differences from the mean	Parameter Estimation Tools Demo.xls
Equation editor to use Simetar functions	Equation Editor Demo.xls
Equicorrelation matrix	Matrices Demo.xls
Equicorrelation matrix	Matrix Operation Tools Demo.xls
Equilibrium displacement model	Cotton Model Demo.xls
Ethanol feasibility study	Project Feasibility Demo.xls
Excel's equation editor for using Simetar functions	Equation Editor Demo.xls
Exponential distribution	Probability Distribution Demo.xls
Exponential distribution	Simulate All Probability Distributions Demo.xls
Exponential distribution	Probability Distributions Demo.xls
Exponential smoothing Holt method	Exponential Smoothing Forecasts Demo.xls
Exponential smoothing Holt-Winters method	Exponential Smoothing Forecasts Demo.xls
Exponential smoothing for probabilistic forecasts	Exponential Smoothing Demo.xls
Exponential smoothing forecast	Cyclical Analysis Tools Demo.xls
Exponential smoothing forecasts	Exponential Smoothing Forecasts Demo.xls
Exponential smoothing probabilistic forecasts	Exponential Smoothing Forecasts Demo.xls
Exponential smoothing trend only	Exponential Smoothing Forecasts Demo.xls
Extreme value distribution	Simulate All Probability Distributions Demo.xls
Extreme value distribution	Probability Distributions Demo.xls
F distribution	Simulate All Probability Distributions Demo.xls
F distribution	Probability Distributions Demo.xls
F test of variances	Data Analysis Tools Demo.xls
Factor a correlation matrix for a MVE distribution	Parameter Estimation Tools Demo.xls
Factor a correlation matrix for a MVE distribution	Matrix Operation Tools Demo.xls
Factor a square symmetric matrix	Matrices Demo.xls
Fan graph of random variable over time	Analysis of Simulation Results Demo.xls
Farm simulator 3 crops	Farm Simulator Demo.xls
Feasibility of purchasing a business	Investment Management Demo.xls
Feasibility study for new business	Project Feasibility Demo.xls
Federal income taxes	Income Tax Demo.xls
Financial statements	Feedlot Demo.xls
Financial statements	Financial Risk Management Demo.xls
Financial statements for a business	Business Demo.xls

Financial statements for multiple enterprise business	Investment Management Demo.xls
Financial statements multi year business	Deterministic Demo.xls
Financial statements with risk	Farm Simulator Demo.xls
Financial statements with risk	Project Feasibility Demo.xls
First degree stochastic dominance	Stochastic Dominance Demo.xls
Forecasting with AR and VAR models	Time Series Functions Demo.xls
GRK and GRKS distributions	Probability Distributions Demo.xls
GRK distribution	GRK Distribution Demo.xls
GRK distribution	Simulate All Probability Distributions Demo.xls
GRK distributions	Probability Distributions Demo.xls
GRK function application	Simulate Alternative Distributions Demo.xls
GRKS distribution	Probability Distributions Demo.xls
GRKS distribution	GRKS Distribution Demo.xls
GRKS distribution for sparse data	Parameter Estimation Tools Demo.xls
Games of chance	Games of Chance Demo.xls
Gamma distribution	Probability Distribution Demo.xls
Gamma distribution	Simulate All Probability Distributions Demo.xls
Generalized inverse of a square matrix	Matrix Operation Tools Demo.xls
Generalized inverse of a square matrix	Matrices Demo.xls
Generalized stochastic dominance for ranking risky alternatives	Stochastic Dominance Demo.xls
Generate random numbers	Probability Distributions Demo.xls
Geometric distribution	Simulate All Probability Distributions Demo.xls
Geometric distribution	Probability Distributions Demo.xls
Harmonic regression for seasonal analysis	Regression for Seasonal Forecasts Demo.xls
Hedging and options for risk management	Financial Risk Management Demo.xls
Heteroskedasticity correction in simulation	Heteroskedasticity Demo.xls
Heteroskedasticity test	Heteroskedasticity Demo.xls
Histogram of a random variable	Analysis of Simulation Results Demo.xls
Hotelling T-Squared distribution	Simulate All Probability Distributions Demo.xls
Hotelling T-squared distribution	Probability Distributions Demo.xls
Hypergeometric distribution	Probability Distribution Demo.xls
Hypergeometric distribution	Simulate All Probability Distributions Demo.xls
Hypergeometric distribution	Probability Distributions Demo.xls
IROR simulated for a business	Net Present Value Demo.xls
Identity matrix	Matrices Demo.xls
Identity matrix	Matrix Operation Tools Demo.xls
Inflation rates stochastic	Farm Simulator Demo.xls
Inner product of two matrices	Matrix Operation Tools Demo.xls
Inner product of two matrices	Matrices Demo.xls
Insurance premium estimation	Insurance Premium Demo.xls
Integrate a function	Data Analysis Tools Demo.xls
Integrate a function	Optimization Function Demo.xls
Internal rate of return for a risky business	Net Present Value Demo.xls
Interpolate function	Empirical Distribution Demo.xls
Intra- and inter-temporal correlation	Complete Correlation Demo.xls
Inventory management with stochastic demand	Inventory Management Demo.xls
Inverse Gaussian distribution	Simulate All Probability Distributions Demo.xls
Inverse Gaussian distribution	Probability Distributions Demo.xls
Inverse transform method of simulating random variables	Inverse Transform Demo.xls
Invert a nonsingular square matrix	Matrix Operation Tools Demo.xls
Invert a nonsingular square matrix	Matrices Demo.xls
Investment analysis under risk	Project Evaluation Demo.xls
Iteration counter ITERATION function	Simulate All Probability Distributions Demo.xls
Iteration counter function	Probability Distributions Demo.xls
Iteration number comparison	Latin Hypercube vs Monte Carlo Demo.xls
Iteration number comparison	Latin Hypercube Demo.xls
Iteration number comparison	Business Model with Risk Demo.xls
J Factor to correct for non-stationarity of CV	Heteroskedasticity Demo.xls
J-factor for CV stationarity Normal distribution	CV Stationarity Normal Demo.xls
Jack knife a covariance matrix	Jack Knife Demo.xls
Jack knife estimator for statistical functions	Jack Knife Demo.xls
Jack knife summary statistics for distributions	Jack Knife Demo.xls
Kernel density estimator	Probability Distributions Demo.xls
Kernel distribution	Probability Distribution Demo.xls

Kernel distribution	Simulate All Probability Distributions Demo.xls
Kernel distribution for 9 kernels	Probability Distributions Demo.xls
Kernel distribution simulation	Sparse Data Demo.xls
Kronecker multiply two matrices	Matrices Demo.xls
Kronecker product of two matrices	Matrix Operation Tools Demo.xls
Latin hyper cube sampling method	Latin Hypercube vs Monte Carlo Demo.xls
Latin hyper cube sampling method	Latin Hypercube Demo.xls
Latin hyper cube vs. Monte Carlo sampling method	Latin Hypercube vs Monte Carlo Demo.xls
Latin hyper cube vs. Monte Carlo sampling method	Latin Hypercube Demo.xls
Likelihood ration test LRT function	Time Series Analysis Tools Demo.xls
Line graph with labels for points	Analysis of Simulation Results Demo.xls
Loan amortization	Feedlot Demo.xls
Log Normal distribution	Probability Distribution Demo.xls
Log normal distribution	Simulate All Probability Distributions Demo.xls
Log normal distribution	Probability Distributions Demo.xls
Log-log distribution	Simulate All Probability Distributions Demo.xls
Log-log distribution	Probability Distributions Demo.xls
Log-logistic distribution	Simulate All Probability Distributions Demo.xls
Logistic distribution	Simulate All Probability Distributions Demo.xls
Logistic distribution	Probability Distributions Demo.xls
Logit regression	Probit and Logit Demo.xls
Lottery	Games of Chance Demo.xls
MAE	Forecast Errors Demo.xls
MAE -- Mean absolute error	Measuring Forecast Errors Demo.xls
MAPE	Forecast Errors Demo.xls
MAPE -- Mean absolute percent error	Measuring Forecast Errors Demo.xls
MLE and MOM to estimate distribution parameters	Parameter Estimation Demo.xls
MLE for estimating distribution parameters	Parameter Estimation Tools Demo.xls
MLE for estimating distribution parameters	Univariate Parameter Estimator Demo.xls
MOM for estimating distribution parameters	Parameter Estimation Tools Demo.xls
MOM for estimating distribution parameters	Univariate Parameter Estimator Demo.xls
MPCI simulation	Crop Insurance Demo.xls
MSQRT function	Matrix Operation Tools Demo.xls
MSQRT function to factor a square matrix	Matrices Demo.xls
MVE distribution	Complete Correlation Demo.xls
MVE distribution	Multivariate Empirical Distribution Demo.xls
MVE distribution in one step	Multivariate Empirical Distribution Demo.xls
MVE distribution parameter estimation in detail	Multivariate Empirical Distribution Demo.xls
MVE distribution prices and costs	Project Feasibility Demo.xls
MVE in one step	Feedlot Demo.xls
MVE intra- and inter-temporal correlation	Complete Correlation Demo.xls
MVE with exogenous projected means	Farm Simulator Demo.xls
MVE with trend projected means	Farm Simulator Demo.xls
MVN distribution	Multivariate Normal Distribution Demo.xls
MVN distribution in one step	Multivariate Normal Distribution Demo.xls
MVN distribution parameter estimation in detail	Multivariate Normal Distribution Demo.xls
MVN parameter estimation and simulation	Multivariate Normal Distribution Demo.xls
MVN validation test	Multivariate Normal Distribution Demo.xls
Marketing options simulation	Futures and Options Demo.xls
Marketing strategies simulated	Futures and Options Demo.xls
Matrix of 1s	Matrices Demo.xls
Matrix of one's	Matrix Operation Tools Demo.xls
Matrix to a vector	Matrix Operation Tools Demo.xls
Maximum likelihood estimation for parameter estimation	Parameter Estimation Tools Demo.xls
Maximum likelihood estimation for parameter estimation	Univariate Parameter Estimator Demo.xls
Maximum likelihood estimator for parameter estimation	Parameter Estimation Demo.xls
Mean absolute error -- MAE	Measuring Forecast Errors Demo.xls
Mean absolute percent error -- MAPE	Measuring Forecast Errors Demo.xls
Mechanical repair costs/failure simulation	Conditional Probability Distributions Demo.xls
Method of Moments for parameter estimation	Parameter Estimation Demo.xls
Method of moments for parameter estimation	Parameter Estimation Tools Demo.xls
Method of moments for parameter estimation	Univariate Parameter Estimator Demo.xls
Model validation statistical tests	Hypothesis Tests Demo.xls
Modified two piece normal distribution	Simulate All Probability Distributions Demo.xls

Monte Carlo sampling method	Latin Hypercube vs Monte Carlo Demo.xls
Monte Carlo sampling method	Latin Hypercube Demo.xls
Moving average forecast	Cyclical Analysis Tools Demo.xls
Moving average forecast	Moving Average Demo.xls
Moving average forecasts	Moving Average Forecasts Demo.xls
Moving average seasonal index	Seasonal Index Forecasts Demo.xls
Multi peril crop insurance analyzer	Crop Insurance Demo.xls
Multinomial distribution	Simulate All Probability Distributions Demo.xls
Multinomial distribution	Probability Distributions Demo.xls
Multiple Regression forecast stochastic w/ SE of predictions	Probabilistic OLS Forecasts Demo.xls
Multiple Regression forecast stochastic w/ Std Dev	Probabilistic OLS Forecasts Demo.xls
Multiple Regression forecast with stochastic betas	Probabilistic OLS Forecasts Demo.xls
Multiple Regression harmonic and dummy variable regression	Regression for Seasonal Forecasts Demo.xls
Multiple Regression linear trend regression	Trend Forecasts Demo.xls
Multiple Regression multiple regression model	Parameter Estimation Tools Demo.xls
Multiple Regression non-linear trend regression	Trend Forecasts Demo.xls
Multiple Regression probabilistic forecasting	Multiple Regression Forecasts Demo.xls
Multiple Regression regression with restrictions	Parameter Estimation Tools Demo.xls
Multiple Regression to estimate risk for a random variable	Multiple Regression to Reduce Risk Demo.xls
Multiple enterprise business	Business Demo.xls
Multiple enterprise business	Farm Simulator Demo.xls
Multiple enterprise business	Feedlot Demo.xls
Multiple regression	Parameter Estimation Tools Demo.xls
Multiple regression forecasting	Multiple Regression Forecasts Demo.xls
Multiple regression model vs. trend model vs. mean model	Multiple Regression to Reduce Risk Demo.xls
Multiple regression to reduce risk	Trend Regression to Reduce Risk Demo.xls
Multiple regression with probabilistic forecast	Multiple Regression Demo.xls
Multiple year financial statement	Net Present Value Demo.xls
Multiplicative seasonal decomposition forecasting with cycle	Seasonal Decomposition Forecasts Demo.xls
Multiplicative seasonal decomposition forecasting without cycle	Seasonal Decomposition Forecasts Demo.xls
Multiply two matrices	Matrix Operation Tools Demo.xls
Multiply two matrices	Matrices Demo.xls
Multivariate Student's t distribution	Probability Distributions Demo.xls
Multivariate empirical distribution	Multivariate Empirical Distribution Demo.xls
Multivariate empirical distribution	Simulate All Probability Distributions Demo.xls
Multivariate empirical distribution -- 1 and 2 steps	Probability Distributions Demo.xls
Multivariate lognormal distribution	Probability Distributions Demo.xls
Multivariate mixed distribution	Multivariate Mixed Probability Distribution Demo.xls
Multivariate mixed distribution	Simulate All Probability Distributions Demo.xls
Multivariate mixed distribution	Probability Distributions Demo.xls
Multivariate normal distribution	Multivariate Normal Distribution Demo.xls
Multivariate normal distribution	Simulate All Probability Distributions Demo.xls
Multivariate normal distribution -- 1 and 2 steps	Probability Distributions Demo.xls
Multivariate test of two distributions	Data Analysis Tools Demo.xls
NORMAL function application	Simulate Alternative Distributions Demo.xls
NPV	Farm Simulator Demo.xls
NPV	Project Feasibility Demo.xls
NPV - Net Present Value	Investment Management Demo.xls
NPV and IROR simulated for 20 year investment	Net Present Value Internal Rate of Return Demo.xls
NPV for alternative discount rates	Feedlot Demo.xls
NPV optimization for a business	Deterministic Optimal Control Demo.xls
NPV simulated for a business	Net Present Value Demo.xls
Negative binomial distribution	Probability Distribution Demo.xls
Negative binomial distribution	Simulate All Probability Distributions Demo.xls
Negative binomial distribution	Probability Distributions Demo.xls
Negative ending cash reserves	Feedlot Demo.xls
Negative ending cash reserves	Financial Risk Management Demo.xls
Net present value for a risky business	Net Present Value Demo.xls
Net returns for one enterprise	Truncated Normal Distribution Demo.xls
Norm of a square matrix	Matrix Operation Tools Demo.xls
Norm of a square matrix	Matrices Demo.xls
Normal distribution	Probability Distribution Demo.xls
Normal distribution	Probability Distributions Demo.xls
Normal distribution	Test Simetar Demo.xls

Normal distribution -- general and direct	Simulate All Probability Distributions Demo.xls
Normal distribution using inverse transform method	Inverse Transform Demo.xls
Normality tests	Conditional Probability Distributions Demo.xls
Normality tests	Data Analysis Tools Demo.xls
Normality tests	Hypothesis Tests Demo.xls
Normality tests for random variable	Validation Tests Demo.xls
Number of iterations	Latin Hypercube vs Monte Carlo Demo.xls
Number of iterations	Latin Hypercube Demo.xls
Number of iterations test	Business Model with Risk Demo.xls
Observational diagnostics -- DF Betas	Parameter Estimation Tools Demo.xls
Optimal control theory for a deterministic simulation model	Optimal Control Demo.xls
Optimal control theory for crop mix decision	Deterministic Optimal Control Demo.xls
Optimal control theory for simulation model	Deterministic Optimal Control Demo.xls
Optimal control theory to maximize NPV	Deterministic Optimal Control Demo.xls
Optimal control theory to solve of equilibrium prices	Wheat Model Demo.xls
Optimal number of lags ARLAG function	Time Series Functions Demo.xls
Optimize a function OPT	Data Analysis Tools Demo.xls
Optimize a non-linear function	Optimization Function Demo.xls
Options and hedging for risk management	Financial Risk Management Demo.xls
Options contracts simulated for market strategy	Futures and Options Demo.xls
Orthogonalize a matrix	Matrix Operation Tools Demo.xls
Orthogonalize a matrix	Matrices Demo.xls
PDF chart of random variables	Analysis of Simulation Results Demo.xls
PDFs for 12 distributions	Test Parameters Demo.xls
PERT distribution	Probability Distributions Demo.xls
PERT distribution -- general and direct	Simulate All Probability Distributions Demo.xls
Parameter estimation for 16 distributions	Parameter Estimation Tools Demo.xls
Parameter estimation for 16 distributions	Univariate Parameter Estimator Demo.xls
Parameter tests -- t and Chi-Square	Data Analysis Tools Demo.xls
Parametric distribution parameter estimator	Parameter Estimation Demo.xls
Pareto distribution	Probability Distributions Demo.xls
Pareto distribution	Simulate All Probability Distributions Demo.xls
Partial autocorrelation coefficients	Time Series Forecasting Demo.xls
Partial autocorrelation coefficients	Time Series Analysis Tools Demo.xls
Partial autocorrelation coefficients	Time Series Demo.xls
Partial autocorrelation test	Time Series Functions Demo.xls
Percentiles with EDF function	Analysis of Simulation Results Demo.xls
Poisson distribution	Probability Distribution Demo.xls
Poisson distribution	Simulate All Probability Distributions Demo.xls
Poisson distribution	Probability Distributions Demo.xls
Poker	Games of Chance Demo.xls
Portfolio analysis	Portfolio Analysis Demo.xls
Power normal distribution	Simulate All Probability Distributions Demo.xls
Power normal distribution	Probability Distributions Demo.xls
Premium calculation for term life insurance	Life Insurance Demo.xls
Premium calculation for whole life insurance	Life Insurance Demo.xls
Probabilistic forecast of Multiple Regression structural model	Multiple Regression Demo.xls
Probabilistic forecast of monthly data	Seasonal Analysis Demo.xls
Probabilistic forecast of time series model	Time Series Forecasting Demo.xls
Probabilistic forecast of time series model	Time Series Analysis Tools Demo.xls
Probabilistic forecasting of Multiple Regression equations	Multiple Regression Forecasts Demo.xls
Probabilistic forecasting of cycles	Probabilistic Cycle Forecasts Demo.xls
Probabilistic forecasting of harmonic regression	Regression for Seasonal Forecasts Demo.xls
Probabilistic forecasting of seasonal index	Regression for Seasonal Forecasts Demo.xls
Probabilistic forecasting with Multiple Regression	Probabilistic OLS Forecasts Demo.xls
Probabilistic forecasting with moving average	Moving Average Forecasts Demo.xls
Probabilistic forecasts with exponential smoothing	Exponential Smoothing Demo.xls
Probabilistic linear and non-linear trend regression	Trend Forecasts Demo.xls
Probabilistic moving average forecast	Moving Average Demo.xls
Probability annual cash flow deficits	Farm Simulator Demo.xls
Probability annual cash flow deficits	Project Feasibility Demo.xls
Probability losing real net worth	Farm Simulator Demo.xls
Probability losing real net worth	Project Feasibility Demo.xls
Probability of success	Feedlot Demo.xls

Probability-Probability (PP) plot chart	Analysis of Simulation Results Demo.xls
Probit regression	Probit and Logit Demo.xls
Production function with risk	Production Function Demo.xls
Production function with risk	Stochastic Production Function Demo.xls
Production insurance (MPCI)	Financial Risk Management Demo.xls
Project management analysis	Project Management Demo.xls
Project management and evaluation	Project Evaluation Demo.xls
QUANTILE function	Analysis of Simulation Results Demo.xls
Quantile-Quantile (QQ) plot chart	Analysis of Simulation Results Demo.xls
RANDSORT function application	Simulate Alternative Distributions Demo.xls
RMSE	Forecast Errors Demo.xls
RMSE -- Root mean square error	Measuring Forecast Errors Demo.xls
Random sort of objects	Probability Distributions Demo.xls
Random walk distribution	Probability Distributions Demo.xls
Rank insurance strategies	Financial Risk Management Demo.xls
Rank of a matrix	Matrix Operation Tools Demo.xls
Rank of a matrix	Matrices Demo.xls
Rank risky alternatives with SERF	SERF Analysis Demo.xls
Rank risky alternatives with SERF	Simulate Scenarios Demo.xls
Rank risky marketing strategies	Financial Risk Management Demo.xls
Ranking alternative portfolios	Portfolio Analysis Demo.xls
Ranking risky alternatives based on NPV	Net Present Value Demo.xls
Ranking risky alternatives with several methods	SDRF and SERF Ranking Demo.xls
Ranking risky alternatives with several methods	Analysis of Simulation Results Demo.xls
Ranking risky marketing options	Futures and Options Demo.xls
Ranking univariate distributions	Univariate Parameter Estimator Demo.xls
Real rate of return to equity	Investment Management Demo.xls
Regression forecasting	Probabilistic OLS Forecasts Demo.xls
Replacement of machinery compliment by item	Machinery Demo.xls
Residuals from regression to measure risk	Multiple Regression to Reduce Risk Demo.xls
Restricted Multiple Regression estimations	Parameter Estimation Tools Demo.xls
Revenue insurance (CRC)	Financial Risk Management Demo.xls
Reverse the order of a vector	Matrix Operation Tools Demo.xls
Reverse the order of data in a vector	Matrices Demo.xls
Risk premiums for ranking risky alternatives	SDRF and SERF Ranking Demo.xls
Risk premiums for ranking risky alternatives	Analysis of Simulation Results Demo.xls
Risky cost of projects	Project Management Demo.xls
Risky investment analysis	Project Evaluation Demo.xls
Root mean square error -- RMSE	Measuring Forecast Errors Demo.xls
Row echelon for of a matrix	Matrix Operation Tools Demo.xls
Row echelon of a matrix	Matrices Demo.xls
SCENARIO function	Scenario Analysis Demo.xls
SDRF for ranking risky alternatives	Stochastic Dominance Demo.xls
SDRF ranking of risky alternatives	Crop Insurance Demo.xls
SERF and SDRF for ranking risky alternatives	Portfolio Analysis Demo.xls
SERF application	SERF Analysis Demo.xls
SERF ranking of risky alternatives	Crop Insurance Demo.xls
Sampling without replacement	Probability Distributions Demo.xls
Scatter matrix	Matrix Operation Tools Demo.xls
Scenario analysis	Feedlot Demo.xls
Scenario analysis	Analysis of Simulation Results Demo.xls
Scenario analysis	Net Present Value Demo.xls
Scenario analysis of a simple business	Simulate Scenarios Demo.xls
Scenario application to simple profit model	Scenario Analysis Demo.xls
Scenario simulation	Simulate All Probability Distributions Demo.xls
Scenario simulation and ranking	Simulate Scenarios Demo.xls
Schwarz criteria for number of lags	Time Series Forecasting Demo.xls
Schwarz criteria for number of lags	Time Series Analysis Tools Demo.xls
Schwarz criteria for number of lags	Time Series Demo.xls
Schwarz test	Time Series Functions Demo.xls
Seasonal decomposition of monthly & quarterly data	Seasonal Analysis Demo.xls
Seasonal forecast of monthly & quarterly data	Seasonal Analysis Demo.xls
Seasonal index	Seasonal Index Forecasts Demo.xls
Seasonal index	Cyclical Analysis Tools Demo.xls

Seasonal index
 Seasonal index
 Seasonal index
 Second degree stochastic dominance
 Seed for pseudo random number generator
 Semicircle distribution
 Semicircle distribution
 Sensitivity analysis
 Sensitivity analysis
 Sensitivity analysis for an economic model
 Sensitivity elasticities for testing models
 Sequence of numbers
 Sequence of numbers
 SimSolver application
 SimSolver application
 Simple average seasonal index
 Simple regression for multiple variables
 Simple statistics for multiple variables
 Simulate a VAR model
 Simulate net returns model
 Simulate simultaneous equation econometric model
 Simulating risky cost to complete a project
 Simulation engine for Simetar demonstrated
 Simulation engine for Simetar demonstrated
 Simulation example for a simple model
 Simultaneous equation model with stochastic errors
 Simultaneous equation simulation
 Simultaneous equation stochastic model
 Simultaneous equation stochastic model
 Sin Cos in Multiple Regression for cycle estimation
 Singular correlation matrix and MV distributions
 Slot machine
 Sole proprietor federal income taxes
 Solve supply and demand model
 Solver for optimal control
 Solver for simultaneous equations
 Solver to simulate simultaneous equation models
 Solver to solve for equilibrium prices
 Sort a matrix by a column
 Sort a matrix by a column
 Sort a matrix by a row or column
 Sparse data distribution simulation
 Sparse data distributions
 Sparse data distributions using GRKS
 Sparse data kernel distribution
 Stationarity tests
 Stationarity tests
 Stationarity tests
 Statistical tests for model validation
 Stochastic chart
 Stochastic dominance with respect to a function
 Stochastic dominance with respect to a function (SDRF)
 Stochastic dominance with respect to a function (SDRF)
 Stochastic econometric model
 Stochastic efficiency with respect to a function application
 Stochastic efficiency with respect to a function (SERF)
 Stochastic efficiency with respect to a function (SERF)
 Stochastic futures and options prices
 Stochastic production function
 Stochastic production function
 StopLight chart for ranking risky alternatives
 StopLight chart of risky alternatives
 Student t test of means
 Student's t distribution
 Exponential Smoothing Demo.xls
 Moving Average Demo.xls
 Seasonal Analysis Demo.xls
 Stochastic Dominance Demo.xls
 Pseudo Random Number Generator Demo.xls
 Simulate All Probability Distributions Demo.xls
 Probability Distributions Demo.xls
 Simulate All Probability Distributions Demo.xls
 Sensitivity Analysis Demo.xls
 Simulate Sensitivity Elasticity Demo.xls
 Simulate Sensitivity Elasticity Demo.xls
 Matrix Operation Tools Demo.xls
 Matrices Demo.xls
 Wheat Sim Solve Demo.xls
 Demand Supply Model Sim Solve Demo.xls
 Seasonal Index Forecasts Demo.xls
 Parameter Estimation Tools Demo.xls
 Parameter Estimation Tools Demo.xls
 Probabilistic Forecasting a VAR Model Demo.xls
 Analysis of Simulation Results Demo.xls
 Sim Solve Demo.xls
 Project Management Demo.xls
 Simulation Demo.xls
 Test Simetar Demo.xls
 Simulation Demo.xls
 Sim Solve Demo.xls
 Simulate All Probability Distributions Demo.xls
 Wheat Sim Solve Demo.xls
 Demand Supply Model Sim Solve Demo.xls
 Probabilistic Cycle Forecasts Demo.xls
 Bad Correlation Matrix Demo.xls
 Games of Chance Demo.xls
 Income Tax Demo.xls
 Demand Supply Model Sim Solve Demo.xls
 Deterministic Optimal Control Demo.xls
 Simulate All Probability Distributions Demo.xls
 Sim Solve Demo.xls
 Wheat Model Demo.xls
 Matrix Operation Tools Demo.xls
 Matrices Demo.xls
 Data Analysis Tools Demo.xls
 Sparse Data Demo.xls
 GRKS Distribution Demo.xls
 Parameter Estimation Tools Demo.xls
 Probability Distribution Demo.xls
 Time Series Forecasting Demo.xls
 Time Series Analysis Tools Demo.xls
 Time Series Demo.xls
 Hypothesis Tests Demo.xls
 Stochastic Production Function Demo.xls
 Stochastic Dominance Demo.xls
 SDRF and SERF Ranking Demo.xls
 Analysis of Simulation Results Demo.xls
 Soybean Model Demo.xls
 SERF Analysis Demo.xls
 SDRF and SERF Ranking Demo.xls
 Analysis of Simulation Results Demo.xls
 Financial Risk Management Demo.xls
 Production Function Demo.xls
 Stochastic Production Function Demo.xls
 Stochastic Dominance Demo.xls
 Analysis of Simulation Results Demo.xls
 Data Analysis Tools Demo.xls
 Probability Distribution Demo.xls

Student's t distribution	Simulate All Probability Distributions Demo.xls
Student's t distribution	Probability Distributions Demo.xls
Summary statistics	Data Analysis Tools Demo.xls
Summary statistics	Trend Regression to Reduce Risk Demo.xls
Supply and demand model	Demand Supply Model Sim Solve Demo.xls
Supply and utilization model -- cotton	Cotton Model Demo.xls
Sweep a square matrix	Matrix Operation Tools Demo.xls
Sweep a square matrix	Matrices Demo.xls
Symmetric covariance matrix	Matrices Demo.xls
TNORM function	Truncated Normal Distribution Demo.xls
TNORM function	Simulate Alternative Distributions Demo.xls
Test 12 distributions for empirical data	View Distributions Demo.xls
Test alternative distributions for empirical data	View Distributions Demo.xls
Test for presence of a trend	Trend Regression to Reduce Risk Demo.xls
Test mean and standard deviation for a distribution	Hypothesis Tests Demo.xls
Test parameters for simulated variable	Validation Tests Demo.xls
Tests means for two distribution -- ANOVA	Validation Tests Demo.xls
Thiel U2	Forecast Errors Demo.xls
Thiel U2	Measuring Forecast Errors Demo.xls
Time series decomposition	Cyclical Analysis Tools Demo.xls
Time series model VAR	Probabilistic Forecasting a VAR Model Demo.xls
Time to complete a project	Project Management Demo.xls
Toeplitz matrix from an array	Matrix Operation Tools Demo.xls
Toeplitz matrix from an array	Matrices Demo.xls
Trace of a square matrix	Matrix Operation Tools Demo.xls
Trace of a square matrix	Matrices Demo.xls
Transpose a matrix	Matrix Operation Tools Demo.xls
Transpose a matrix or vector of any size	Matrices Demo.xls
Trend regression to reduce risk	Trend Regression to Reduce Risk Demo.xls
Triangle distribution	Probability Distribution Demo.xls
Triangle distribution	Probability Distributions Demo.xls
Triangle distribution -- general and direct	Simulate All Probability Distributions Demo.xls
Truncated Weibull distribution	Probability Distributions Demo.xls
Truncated empirical distribution	Probability Distribution Demo.xls
Truncated empirical distribution	Probability Distributions Demo.xls
Truncated gamma distribution	Probability Distributions Demo.xls
Truncated normal distribution	Probability Distribution Demo.xls
Truncated normal distribution	Simulate Alternative Distributions Demo.xls
Truncated normal distribution	Probability Distributions Demo.xls
Truncated normal distribution -- general and direct	Simulate All Probability Distributions Demo.xls
Truncated normal distribution application	Truncated Normal Distribution Demo.xls
Two Sample Hotelling T-Squared test	Data Analysis Tools Demo.xls
Two piece normal distribution	Probability Distributions Demo.xls
UNBOXCOX function	Data Analysis Tools Demo.xls
UNIFORM function	Uniform Random Number Generator Demo.xls
UNIFORM function application	Simulate Alternative Distributions Demo.xls
UNIFORM vs. Excel's RAND function	Uniform Random Number Generator Demo.xls
Uniform distribution	Probability Distribution Demo.xls
Uniform distribution	Probability Distributions Demo.xls
Uniform distribution	Test Simetar Demo.xls
Uniform distribution -- general and direct	Simulate All Probability Distributions Demo.xls
Uniform distribution to simulate a Normal	Uniform Random Number Generator Demo.xls
Uniform distribution using inverse transform method	Inverse Transform Demo.xls
Univariate distribution parameter estimation	Univariate Parameter Estimator Demo.xls
Univariate distribution parameter estimation	Trend Regression to Reduce Risk Demo.xls
Univariate parameter estimation system	Parameter Estimation Demo.xls
VAR model estimation	Time Series Functions Demo.xls
VAR model estimation	Time Series Analysis Tools Demo.xls
VAR model for two series	Probabilistic Forecasting a VAR Model Demo.xls
Validate correlation of random variables in MV distribution	Hypothesis Tests Demo.xls
Validation for MV distributions correlation matrix	Validation Tests Demo.xls
Validation test of MVE	Multivariate Empirical Distribution Demo.xls
Validation test of MVN	Multivariate Normal Distribution Demo.xls
Validation tests	Hypothesis Tests Demo.xls

Vector to a diagonal matrix
Vector to a diagonal matrix
View distributions as parameter change
WAPE
WAPE -- Weighted absolute percent error
Weibull distribution
Weibull distribution
Weibull distribution
Weighted absolute percent error -- WAPE
Wilk's Lambda distribution
Wilk's lambda distribution
Wishart distribution
Wishart distribution

Matrix Operation Tools Demo.xls
Matrices Demo.xls
Test Parameters Demo.xls
Forecast Errors Demo.xls
Measuring Forecast Errors Demo.xls
Probability Distribution Demo.xls
Simulate All Probability Distributions Demo.xls
Probability Distributions Demo.xls
Measuring Forecast Errors Demo.xls
Probability Distributions Demo.xls
Simulate All Probability Distributions Demo.xls
Simulate All Probability Distributions Demo.xls
Probability Distributions Demo.xls

Appendix A: Getting Started in Microsoft Excel XP

An Excel Workbook consists of one or more Worksheets. The name of the Workbook appears at the top of the screen after the phrase “Microsoft Excel”, see Figure A.1. The tabs at the bottom of the screen are Worksheet names (Figure A.1). The Worksheet names default to Sheet 1, Sheet 2, Sheet 3, ... when a new Workbook is opened. Each Worksheet has 255 columns (A through IV) and 65,536 rows. A Workbook can have more than 250 Worksheets. See Tools > Options > General to set the number of sheets to open for a new Workbook; 3 is plenty as you can add more later. The option for setting the default number of sheets and the default file location for saving Workbooks are indicated later in this Appendix. The primary purpose of Excel is to perform lots of simple mathematical calculations. Each cell can contain an equation, a constant, or text. The syntax for entering these three types of values in a cell are:

- Equations begin with an equal sign, = ,
- Constants begin with any number. Constants can be real numbers or integers.
- Text begins with an alphabetic character, without the need for a quote or other special character.
- A cell that documents a formula should start with a single quote, such as, ‘= to tell Excel the cell is text and not a formula.

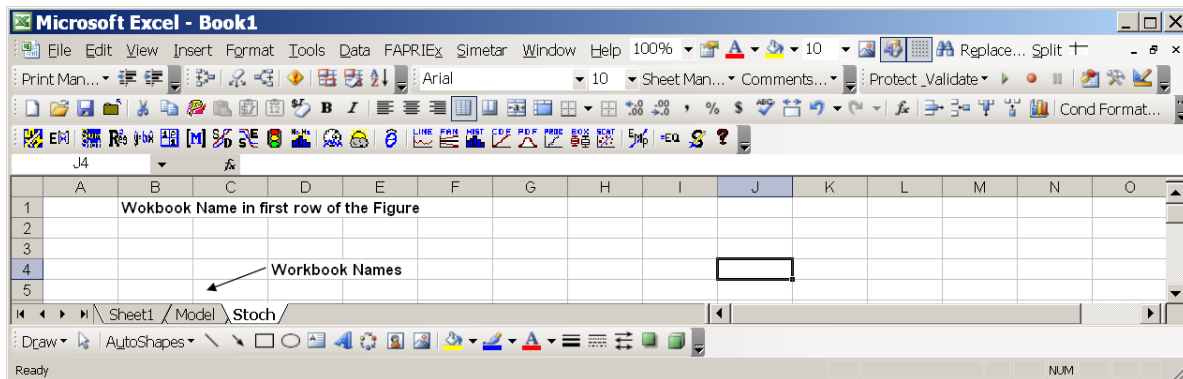


Figure A.1. Workbook Name and WorksHeet Names.

Edit Workbooks and Worksheets

When Excel is opened it begins with a blank Workbook named “Book1.” To organize your work the Workbook should be saved with a unique name. This is done using the Save As Option which is accessed as follows: File > Save As The Save As Menu appears so you can browse to the directory where the file is to be stored, type the new name in the “File name” window and be sure to click on the “Save” button. Keep your Workbook names short, 8 to 12 letters and numbers. Do not use special characters in the name.

To edit the Worksheet name, right click on the Worksheet name and select the Rename option. Next type the new name for the Worksheet. Be sure to use short names for Worksheets, 4 to 8 characters and numbers. Do not use special characters in Worksheet names.

Other functions in the menu for editing Worksheets allow you to copy, delete, move, or insert an entire Worksheet. The copy Worksheet function is particularly useful, as you can copy

a SimData Worksheet output sheet for Simetar that has charts and tables and then use the original SimData Worksheet for the next simulated output and make use of the pre-programmed charts and tables. Moving a Worksheet name within a Workbook can be done by clicking on the Worksheet name, holding down the left mouse button and dragging and dropping the name to its new location.

It is not recommended that you copy a Worksheet from one Workbook to another. This action will create links that are hard to break.

Insert and Delete Rows and Columns

Excel's Worksheets are flexible so the user can add and delete rows and columns. This feature is useful for model development when a new row must be inserted into the existing model or a new column needs to be inserted in a table.

To insert one column click on the main toolbar for Insert > Columns. The new column appears to the left of the cursor. To insert two or more columns, highlight the number of columns to insert and click on Insert > Columns.

To insert one row above the cursor, click on Insert > Rows. To insert multiple rows, highlight the desired number of new rows and click on Insert > Rows.

Removing a row or rows is accomplished by highlighting the row or rows and clicking on Edit > Delete. Follow the same steps to remove a column or columns.

Getting Help

Select the word Help on the main Toolbar (Figure A.1) and then select "Microsoft Excel Help" in the subsequent drop down menu. Help can also be accessed by pressing the F1 function key. Help is provided in several different forms.

- **Table of Contents** -- tab lists the major areas where you can find detailed help. Under each major topic area you will find numerous sub-topics with sub-levels of help under each topic. For example, click on Printing to see the range of topics.
- **Search for:** -- window provides a keyword field so you can type in a word or phrase to gain help for a topic (Figure A.2). Typing in a phrase such as "if" results in a list of different IF statements and how they are programmed. Additional assistance is available by connecting to Microsoft Office Online (Figure A.2). Try using the help by typing "if" in the first line and then follow through with help on the topic of "Check if a number is greater than or less than another number."

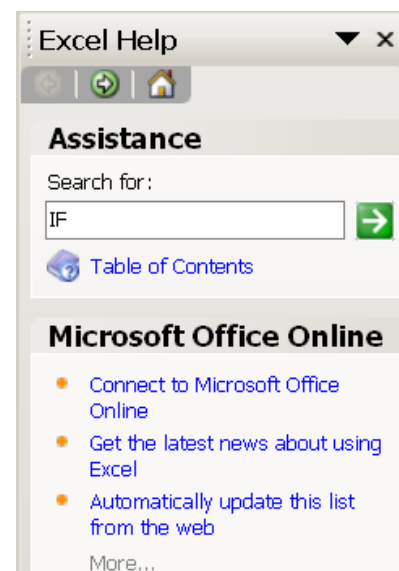


Figure A.2. Excel Help Screen.

Help with Functions

Select the “*fx*” icon or ‘=’ on the bottom toolbar to get a list of all functions in Excel and to get help in using them (Figure A.1). The resulting “Insert Function” menu (Figure A.3) is split in three parts. The top line provides space for you to type a description of what you want to calculate. The second input line is a drop down list of categories of functions, such as mathematical, statistical, etc. The Select a Function window is a scrollable list of functions in the category selected. The Simetar functions are under Simetar in the list of Categories.

To test how the Insert Function works, click on a cell, then click on *fx*, select Statistical in the Select Category box. In the Function screen, scroll down to CORREL and double click it. The format for the function and an explanation appears in a dialog box (Figure A.4). Tips on what to enter in each data entry box or field are provided as you move from one field to another. Follow the instructions and Select OK. The correlation coefficient appears in the cell you selected prior to hitting *fx*.

The *fx* help/dialog box can be accessed to provide help with all types of functions. It is useful for learning the syntax for new functions and to verify correct syntax for functions that are reporting an error. To test this

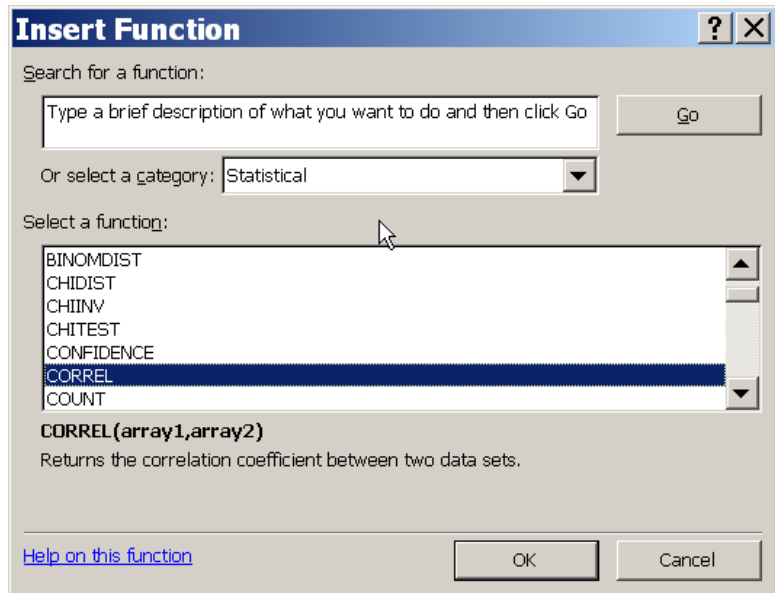


Figure A.3. Help Screen for Inserting a Function.

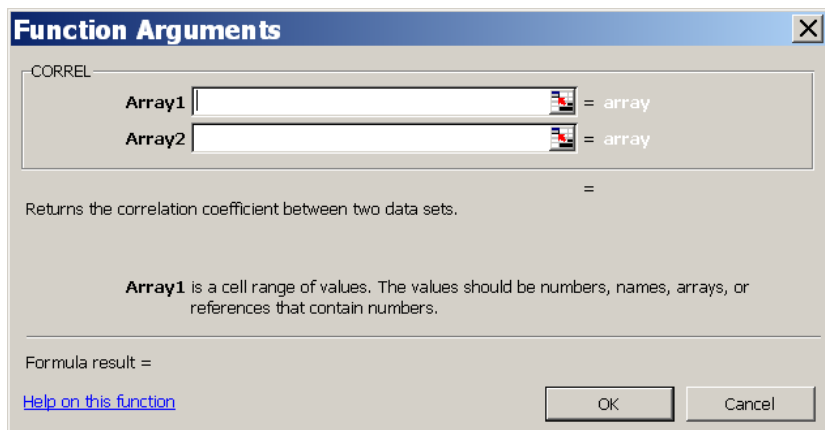


Figure A.4. Example of a Function Argument Help Screen.

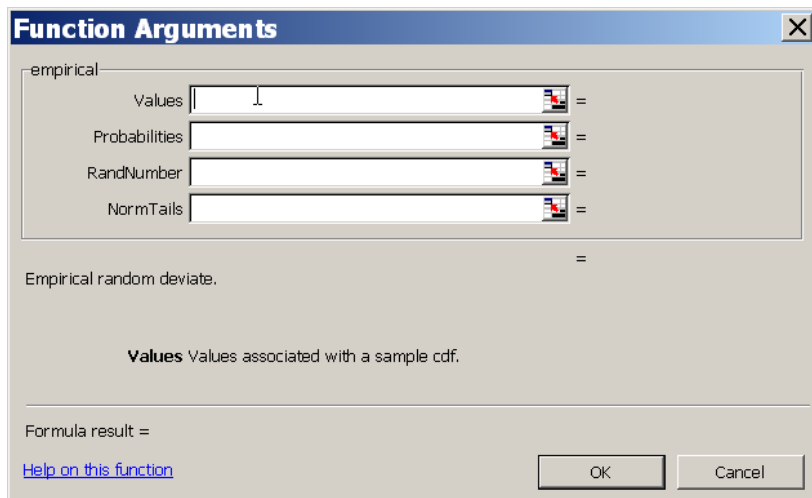


Figure A.5. Example of a Function Editing Screen for EMPIRICAL.

Excel capability, select a cell and type a function name with the left parenthesis, e.g., =EMPIRICAL(and hit the *fx* icon on the toolbar. Excel responds by providing the help menu in Figure A.5, which indicates the parameters for the EMPIRICAL function. Cell locations or numbers can be typed in the boxes provided in the help menu in Figure A.5. Repeat this example for another function, such as: =AVERAGE() or =STDEVP().

Printing Reports

Excel provides sufficient flexibility to both produce business reports and to completely frustrate the user. The good news is that Excel saves each Worksheet's print format information so you only have to go through the print formatting steps once for each Worksheet. The steps to follow for printing a file are:

- First highlight the section of the spreadsheet to print using the left mouse button and dragging the mouse. Always specify what to print, if you ignore this step, Excel will dump the entire Worksheet to the printer.
- Set the Print Area by selecting: File > Print Area > Set Print Area
This action places a colored box with a dashed line around the cells to print.
- Next format your print region by selecting: File > Print Preview
This action opens a window with your selected region placed on a simulated sheet of paper. This is how the report will look if you select the Print option in the menu at the top of the screen (Figure A.6). To make the page look the way you want, use the menu options in Figure A.6.

The "Setup" button in Figure A.6 opens a four level dialogue/menu box for formatting your

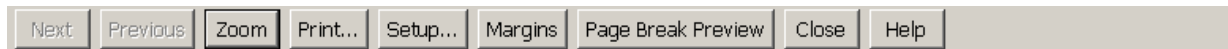


Figure A.6. Menu for Page Preview

report (see Figure A.7).

The four tabs (Page, Margins, Header/Footer and Sheet) are described briefly:

- Page -- tab allows you to set the orientation (Portrait or Landscape). I prefer the Fit to "1" page wide by "N" pages tall option.
- Margins -- tab allows you to change the margins on the page. The margin settings are in effect for all pages selected in the print region. Margins can be set easier from the Main toolbar in Figure A.6, see below.
- Header/Footer -- tab provides text boxes for you to enter your own or select from numerous headers and footers. Drop down menu boxes are provided with several pre-

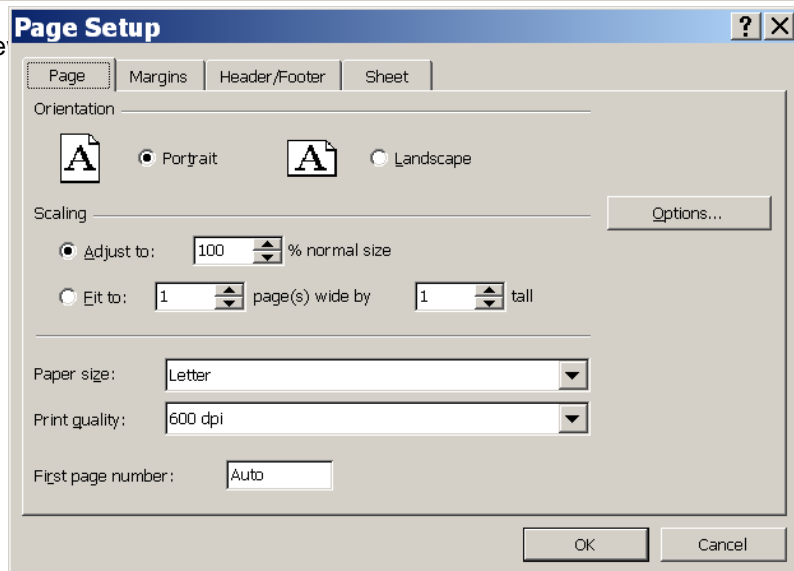


Figure A.7. Page Setup Dialog Box for Printing.

Margins can be set easier from the Main toolbar in Figure A.6, see below.

- Header/Footer -- tab provides text boxes for you to enter your own or select from numerous headers and footers. Drop down menu boxes are provided with several pre-

set titles. The titles include the page number, your name, the name of the file, and the name of your computer.

- Sheet -- tab is used to specify whether you want to print Grid Lines and/or Row and Column Headings.

The “Margins” button on the menu at the top of the screen (Figure A.6) causes Excel to show the margins on the print preview page. To change a margin click on the dotted margin line and drag where you want.

The “Page Break Preview” button on the menu at the top of the screen (Figure A.6) changes the screen back to the Worksheet where you can change the page breaks manually. To change a page break, click on the dashed line and drag it where you want the page break. Once you are satisfied with the page break go back to the print preview screen by selecting File > Print Preview or clicking View > Normal from the Main Toolbar.

The “Print” button on the menu at the top of the screen (Figure A.6) will print what you see in the print preview.

- The “Next” button on the Print Preview toolbar (Figure A.6) advances the screen preview through the remaining sheets in the print area.

Format each Worksheet you want to print by repeating the steps described above. To print the Set Print Areas in multiple Worksheets, select the Worksheets names to print (click on a name, hold the Ctrl key and click on other Worksheet names) and click File > Print.

Formatting Cells in a Worksheet

The format option on the Excel main Toolbar is used to format the cells in a Worksheet. Highlight the cell or cells to format and then select the Format > Cells ... option.

Selecting Format > Cells ... opens a dialog box (Figure A.8) with six tabs to provide a great deal of flexibility in formatting your Worksheet. The function of each tab is described as follows:

- Number -- tab in Figure A.8 provides a list of categories that you can pick from, in a scrollable menu. Use the arrow keys to highlight the preferred category. Sample settings available for each category appear in the right side of the box.

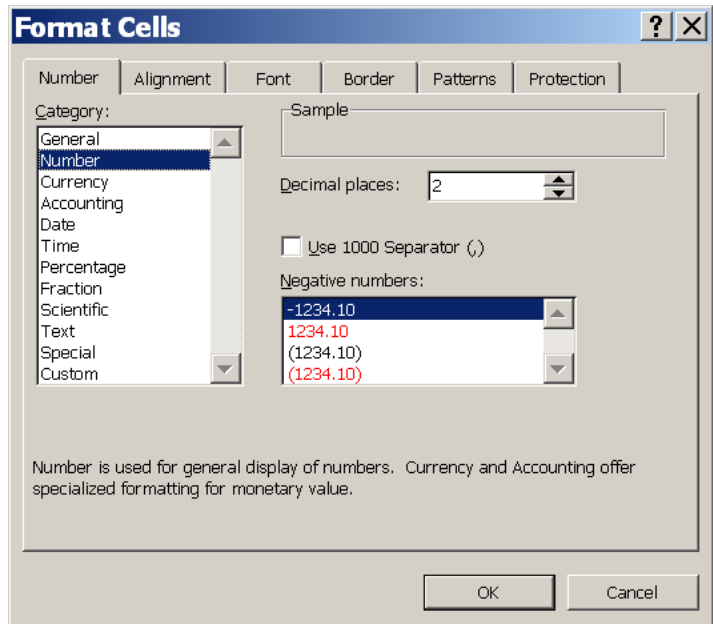


Figure A.8. Dialog Box for Formatting Cells.

- Number: specify number of decimal points, use of a comma separator, and how to

- display negative values.
 - Currency: specify decimals, use of \$, and options for displaying negative values.
 - Percentage: specify number of decimal places.
- Alignment -- tab in Figure A.8 provides options to position numbers and text in individual cells, different from the default settings. The primary option in this tab is Horizontal alignment of text and numbers, which allows you to center, left or right justify the contents of a cell.
 - Font -- tab in Figure A.8 provides several small windows for setting the font type, font style, font size, underlining, and color of the cell contents.
 - Border -- tab in Figure A.8 provides options for specifying line borders around individual or blocks of cells. The size and type of lines used for borders can be specified.
 - Patterns -- tab provides tools for setting a color for individual cells or blocks of cells. A palette of colors is provided to select from.
 - Protection -- tab in Figure A.8 provides a way to lock or hide a cell so it cannot be changed by the user. Protection of an Excel Worksheet involves two steps that must be done in this order:
 - Highlight the cell(s) that the user can change; click Format > Cells > Protection > unclick the box for Locked.
 - Click on Tools > Protection > Protect sheet > click on “Select locked cells” and click on “Select unlocked cells.” You must enter a password to protect the Worksheet. Caution is recommended because if you forget the password you can never edit the Worksheet.
 - Users will be able to change values in the cells that are selected in the first step and cannot change values in any other cells in the Worksheet.

Writing Equations in Excel

The arithmetic operators are:

- Addition +
- Subtraction -
- Division /
- Exponentiation ^
- Multiplication *

An equation of constants could be typed into a cell as:

$$= 10 + (20.5 * 2.01) + (1.04 ^ 2) / 3.0$$

The power of Excel comes from using cell addresses in equations, rather than constants, as:

$$= C11 + (C12 * C13) + (A12 ^ 2) / A3$$

The values in cells A3, C11, C12, and C13, are inserted by Excel into the equation for calculation. Calculation can be done manually by pressing F9 or automatically by selecting

Tools > Options > Calculations Automatic. For most all forms of modeling in Excel it is best to have Calculations updated automatically.

Most users find it easiest to write a formula using a combination of typing and the mouse. Try the following to enter the equation of = A1 + A2 or = 4 + 3.

- Type 4.0 in A1 and 3.0 in A2.
- In cell A3 type = and left (mouse button) click on cell A1 then press the “+” key and left click cell A2.
- Press Enter and you will see the result of the calculation.
- Type a new value, say 10, in A1 and watch the change in A3. (If A3 did not change press F9 or set Calculation to automatic by selecting Tools > Options > Calculations > Automatic.)

This feature of Excel is particularly useful in simulation because all input and control values can be easily changed for alternative scenarios if they are entered in cells rather than being hard coded in equations.

Edit Equations

To edit the equation in A3, highlight the cell and press the F2 function key. The formula will appear in the A3 cell and in the *fx* formula bar. You can edit the formula in either place, using the mouse and the arrow key to position the cursor. Backspace deletes characters to the left and Delete removes characters in the formula to the right of the cursor.

Equations can be edited more efficiently by using the F2 key and the mouse. For example, enter the number 10.0 in cell B1, click on the formula in A3 and press the F2 key. Now change the A3 equation to = A1 + B1 by positioning the cursor over cell A2 until the cursor becomes a cross with four arrows, hold down the left mouse button and drag the box to cell B1, and press Enter, using F4 to Lock Cells.

Excel allows the user to copy equations to speed up the process of developing a model. Because Excel is a “relational” program you must “lock” or “fix” the cell addresses prior to copying, if a cell address is to be constant in all locations. Lock the column reference for a cell with a “\$” sign before the column letter. Lock the row reference for a cell with a “\$” sign before the row number. Lock both row and column for a cell by following these steps.

- Highlight the cell with the equation to be edited by placing the cursor on the cell,
- Press F2 to edit the cell,
- With the mouse or the arrow keys position the cursor on the cell name (e.g. A1) in the formula to be locked, and
- Press F4 to lock the cell reference. Multiple clicks of F4 will cell reference only the column or only the row.

	A	B	C	D	E
8	Centigrade to Fahrenheit Calculator				
9	Cent	Fahren			
10	Freeze		32		
11	Diff	100	292		
12	Difference	100	180	=C10-C10	
13	Ratio		1.8	=C12/B12	
14	Temp	Fahren			
15	1	0	32	=C\$10+C\$13*B15	
16	2	5	41	=C\$10+C\$13*B16	
17	3	10	50	=C\$10+C\$13*B17	
18	4	15	59	=C\$10+C\$13*B18	
19	5	20	68	=C\$10+C\$13*B19	
20	6	25	77	=C\$10+C\$13*B20	
21	7	30	86	=C\$10+C\$13*B21	
22	8	35	95	=C\$10+C\$13*B22	
23	9	40	104	=C\$10+C\$13*B23	

An example of locking (cell referencing) the cells in an equation is provided in Figure A.9. The example is a simple Centigrade to Fahrenheit calculator where the formula is $F = 32 + 1.8 * (C)$. The slope for the line is calculated in C13 and the intercept is in C10. The intercept

Figure A.9. Example of Cell Locking with F4 in an Equation.

and slope constants are in the formulas in C12:C20 to convert nine different Centigrade temperatures. The formula in C12 was entered as $\$B\$10 + \$C\$10 * B12$ and then the formula was copied to cells C13:C20.

Copy and Paste Formulas

Highlight the cell or block of cells to copy using the mouse and press the Copy icon on the toolbar. (The block will have moving dashed lines around it after it is copied into active memory.) Move the cursor to the new location using the arrow keys or the mouse. Press the Paste icon. This can also be done using Ctrl C to copy and Ctrl V to paste.

	A	B	C	D	E	F
22	Copy and Paste Formulas					
23	Interest rate	0.075		0.085		0.095
24	Principal	250,000				
25	Interest Due	18,750		21,250		23,750
26		=B23*\$B\$24		=D23*\$B\$24		=F23*\$B\$24

Figure A.10. Example of Copy and Paste an Equation with Cells Locked Using F4.

An example of copy and paste equations is provided in Figure A.10. The interest payment for a one year loan is calculated for three interest rates and a constant principal borrowed. Interest rates are entered in cells B23, D23, and F23. The formula to calculate interest is entered in B25. Because the principal in the formula in B25 is to be used for all three interest rates, the formula is edited to lock $\$B\24 using the F4 key. The formula in B25 was copied and pasted in cells D25 and F25. The calculated values in D25 and F25 use the interest rate in the same relative position on the Worksheet (2 cells above the formula) as the original formula in A25. As a result use caution when pasting formulas.

Drag Formulas

Programming formulas in a Worksheet can be more efficient if you drag the original formula down (or across) from the original cell. In Figure A.9 the original formula is in cell C12 and it was dragged down for cells C13:C20 faster than the formula can be re-typed once. To drag a formula to an adjacent cell follow these steps.

- Highlight a cell (or range of cells) with a formula.
- Place the cursor on the black square in the lower right hand corner so the cursor becomes a small black cross.
- Next hold down the left mouse button and drag the formula to the next cell or cells.

The formula will now be in both places, the original and the new cell. The cell references in the formula are adjusted to reflect the new location, unless the cell references were row and column locked. Excel is a relational program, so if you filled a cell to the right all cell references in the formula have their original row numbers but all column references are advanced by one letter. Likewise, if the filled cell was below the original, all row numbers would be incremented by one while the column references are unchanged. This feature in Excel is particularly useful for developing equations in a model. An example of dragging a formula to the right is provided in Figure A.11 where we calculate interest payments for five interest rates. The original formula in B32 was dragged to the right to cells C32:F32. The principal amount borrowed is constant for all formulas because of the F4 cell locking on cell B31, but the interest rates change from one

cell to the next.

	A	B	C	D	E	F	
29	Drag Formulas						
30	Interest rates	0.075	0.085	0.095	0.105	0.115	
31	Principal	250,000					
32	Interest Due	18,750	21,250	23,750	26,250	28,750	
33		=B30*\$B\$31		=D30*\$B\$31		=F30*\$B\$31	

Figure A.11. Example of Dragging a Formula to Populate the Equations in a Table.

Moving a Formula or Block of Formulas

Formulas can be moved from one place to another on the Worksheet. This feature is useful in that we can program a formula in one place and move it to where it goes later.

- First highlight the cell or cells containing the formulas to move. This is done by placing the cursor on the first cell and then holding the left mouse button down as you drag the mouse to highlight the cells to move.
- Next move the cursor to an edge of the block, at this point the cursor turns into a white arrow.
- Click and hold down the left mouse button as you move the mouse, and drag the block to its new location.

Using this procedure to move cells with equations automatically updates the cell addresses in the equations as well as all other equations that use the cells being moved. This is not the case if you Copy and Paste equations.

Fill a Series

A column (or row) of numbers which contains a series such as 1, 2, 3, ..., N can be generated by Excel. This is done by typing the first two numbers in the series and having Excel complete the series by following these steps:

- Type values 1.0 and 2.0 into cells A12, A13, respectively (see Figure A.9),
- Use the mouse to highlight the two cells A12:A13 and move the cursor to the bottom right corner where there is a small block box,
- When the cursor touches the small block box it changes to a black cross, hold the left mouse button down and drag the cursor to the end cell, say, A20. (As you drag a series down, a small box to the right of the cursor informs you of the current number in the series.) Release the left mouse button when you are finished.

Frequently Used Excel Functions

– Average

The average function calculates the values in a specified block of row(s) or column(s). The format is:

= average (first cell: last cell) or = average (A6:A18)

All Excel functions allow you to declare the cell addresses in the formulas by using the mouse. After typing the left parentheses, click and hold down the left mouse button on the first cell (A6) and drag to the last cell (A18) and then release the button. Finish the formula by typing the right parentheses and pressing Enter.

– Standard Deviation for a Population

= stdevp (first cell: last cell)

– Minimum Value

= min (first cell: last cell)

– Maximum Value

= max (first cell: last cell)

– Sum of a Series

= sum (first cell: last cell)

– IF Statements

Excel provides for conditional equations using a simple “IF, Then, Else” format which is essential for simulation models. An IF statement follows a set format:

= IF (condition to be tested, value if condition is true (then), value if condition is false (else))

The condition can be a comparison of one cell to a constant or one calculated value to another. The “then” and “else” values can be constants or calculations or cell addresses, or text. Examples of several common IF statements are demonstrated that compare two cells (A4 and A3) and make the “then” value 10 and the “else” equal to a 20. An example is also provided in Figure A.12.

IF Greater than

= IF (A4 > A3, 10, 20)

IF Less than

= IF (A4 < A3, 10, 20)

IF Equal to

= IF (A4 = A3, 10, 20)

IF Less than or Equal to

= IF (A4 <= A3, 10, 20)

IF Greater than or Equal to

= IF (A4 >= A3, 10, 20)

Not Equal to

= IF (A4 <> A3, 10, 20)

	A	B	C	D	E	F
30	Interest rates	0.075	0.085	0.095	0.105	0.115
31	Principal	250,000				
32	Interest Due	18,750	21,250	23,750	26,250	28,750
33		=B30*\$B\$31		=D30*\$B\$31		=F30*\$B\$31
34						
35	IF Statements					
36	If Greater Than		21250	=IF(C32>B32,C32,0)		
37	If Less Than		18750	=IF(B32<C32,B32,0)		
38	If Equal		Not Equal	=IF(C30=D30,"Equal","Not Equal")		
39	If Equal		0	=IF(C30=D30,C30,0)		
40	If Not Equal		1	=IF(B30<>F30,1,0)		
41	If And		1	=IF(AND(B30<C30,D30<E30),1,0)		
42	If Or		1	=IF(OR(C30>B30,E30>D30),1,0)		

Figure A.12. Example of Basic IF Functions in Excel.

Compound IF statements are available, for problems that involve two conditions, as:

IF And

For example, if A4 is less than A3 and A4 is positive:

= IF (and (A4 < A3, A4 > 0), 10, 20)

IF Or

For example, if A4 is less than A3 or A4 greater than two:

= IF (or (A4 < A3, A4 > 2), 10, 20)

You can of course include IF statements inside other IF statements. Use caution, and test thoroughly before using complex IF statements in your models.

VLOOKUP and HLOOKUP Functions

Excel provides two functions for extracting information from a table. The VLOOKUP function extracts information from tables arranged vertically and HLOOKUP extracts information from tables arranged horizontally.

– **HLOOKUP Function**

The HLOOKUP operates on horizontal tables such as the one displayed in Figure A.13. For this example, assume a horizontal table is defined by rows 46-49 in columns B-E, as depicted in Figure A.13. The general format for the HLOOKUP function is:

=HLOOKUP (Target Value in Row 1, Table Location, Row to Look in)

To extract values from the third column of the table, program HLOOKUP as follows:

- Information in row 2 is obtained by
=HLOOKUP (3, A2: G6,2) which returns 22.1
- Information in row 3 is obtained by
=HLOOKUP (3, A2: G6, 3) which returns 4.5
- Information in row 4 is obtained by
=HLOOKUP (3, A2: G6, 4) which returns 128

The way this function operates is that it looks at the values in row 1 and compares them to the Target Value in the HLOOKUP function to locate the column; the row to extract from is specified by the last parameter in HLOOKUP. The function is quite general in that the Target Value in row 1 and the Rows can be specified by the value in a cell (Figure A.13).

A word of warning is that the columns must be arranged so the values in row 1 are unique and increase in value from left to right. If the values in row 1 are not unique, Excel will extract information from the first column where it finds a match. For example, if a 2 appeared in row 1 for both columns 2 and 4, then Excel only pulls in values for column 2 when a 2 is used as the first parameter.

	A	B	C	D	E	F	G
45	VLOOKUP of Quartely Data for Years 1-10						
46		1	2	3	4		
47	Premium	14.5	18.5	22.1	32.5		
48	Price Election	4.5	4.5	4.5	4.5		
49	Yield Guarantee	105	120	128	145		
50	Column to Extract	4					
51		Values Extracted					
52	Premium	2	32.5	=HLOOKUP(\$C\$50,\$B\$46:\$E\$49,B52)			
53	Price Election	3	4.5	=HLOOKUP(\$C\$50,\$B\$46:\$E\$49,B53)			
54	Yield Guarantee	4	145	=HLOOKUP(\$C\$50,\$B\$46:\$E\$49,B54)			

Figure A.13. HLOOKUP Function to Extract Information from a Column of a Table.

– VLOOKUP Function

The VLOOKUP function operates on a vertical table such as the one displayed in Figure A.14. The VLOOKUP function pulls data from a row of a table. The row to use is based on matching the Target Value to the values in the first column of the table. The general format for the VLOOKUP function is:

= VLOOKUP (Target Value in Column 1, Table Location, Column to Look in)

To extract information from the income tax schedule in Figure A.14, one would start with the taxable income, say 76000, in cell C69 and use the following specifications:

- Information in column 2 is obtained by
VLOOKUP (B17, A60:D67, 1) which returns 75000.0

- Information in column 3 is obtained by
=VLOOKUP (B17, A60:D67, 3) which returns 13.750.0
- Information in column 4 is obtained by
=VLOOKUP (B17, A60:D67, 4) which returns 0.34.

The VLOOKUP function assumes the values in the first column are sorted from the lowest to the highest, and each value is unique. If Excel does not find a perfect match for the values in column 1, it uses the row that has the largest value which is less than the target value. In other words if the target value or taxable income is \$51,000 then Excel uses the values found in row 2 of the A60:D67 table (Figure A.14). If taxable income equals a value between \$100,000 and \$334,999.99 the function will use values in row 4 of A60:D67, as specified by the column specifier.

	A	B	C	D	E	F	G
57	Use VLOOKUP to find values in the tax schedule						
58	Taxable income C	But Not	Base Tax	Marginal			
59	of--	Over--	is:	Tax Rate			
60	0	50,000	0	0.15			
61	50,000	75,000	7,500	0.25			
62	75,000	100,000	13,750	0.34			
63	100,000	335,000	22,250	0.39			
64	335,000	10,000,000	113,900	0.34			
65	10,000,000	15,000,000	3,400,000	0.35			
66	15,000,000	18,333,333	5,150,000	0.38			
67	18,333,333	#####	0	0.35			
68		Year 1	Year 2	Year 3			
69	Taxable Income	49900	78000	101000			
70	Tax Base	0	75,000	100,000	=VLOOKUP(D\$69,\$A\$60:\$D\$67,1)		
71	Minimum Inc.	0	13,750	22,250	=VLOOKUP(D\$69,\$A\$60:\$D\$67,3)		
72	Tax Rate	0.15	0.34	0.39	=VLOOKUP(D\$69,\$A\$60:\$D\$67,4)		

Figure A.14. VLOOKUP Function to Extract Information from a Row of a Table.

Array Function

Matrix operations in Excel are handled by using array functions. The product of multiplying two matrices, $Y = X * Z$, is an array, not a cell. To insure that the product, Y, is placed into an array rather than a cell one must highlight the correct size of an array to store the results, type the function correctly, and most importantly end the command by pressing “Control Shift Enter” all at once. To multiply two matrices, X in A1:D4 and Z in E1:H4 we highlight a 4x4 location for the product in say A5:D8 and type the command =MMULT(A1:D4, E1:H4), and then press Control Shift Enter. The Y matrix is now treated by Excel as an “array” so you cannot edit or delete an individual cell in the array.

	A	B	C	D	E	F	G	H	I
99	Multiply X'Y								
100	10384	=MPROD(\$A\$93:\$O\$96,\$B\$76:\$B\$90)							
101	417140.9								
102	655472								
103	1422831.2								
104									
105	Multiply X'X				Inverse of (X'X)				
106	15	600.7	944.2	2031.8	15.12683	-0.26373	-0.05804	-0.00624	
107	600.7	24181.13	37894.8	81727.8	-0.26373	0.012266	-0.00085	-0.00129	
108	944.2	37894.8	60105.42	128243	-0.05804	-0.00085	0.001637	-8.2E-05	
109	2031.8	81727.84	128242.9	278427	-0.00624	-0.00129	-8.2E-05	0.000465	
110	=MPROD(\$A\$93:\$O\$96,\$C\$76:\$F\$90)					=MINV(\$A\$106:\$D\$109)			
111									
112	Product of Inverse[(X'X)] and (X'Y) or the Beta Coefficients for a Multiple Regression								
113	142.343413	=MPROD(\$F\$106:\$I\$109,\$A\$100:\$A\$103)							
114	-6.600445639								
115	0.566832915								
116	5.747875215								

Figure A.15. Example of Using Array Functions in Excel.

The array function capability in Excel is used to speedup calculations and to insure that functions are solved as an array. Examples of functions that benefit from the array function mode are: matrix inversion, factoring a matrix, correlating random variables, and matrix

multiplication, to name a few. For an example of how array functions can be used in Excel to solve a practical problem see Figure A.15. Array functions are demonstrated to calculate betas for a multiple regression using the formula:

$$\beta = (X'X)^{-1} X'Y.$$

Split Screen

To see two or even four different sections of a Worksheet at once you can split the screen with either horizontal or vertical segments. Splitting the screen into two horizontal segments, left click on the small bar segment at the top of the scroll bar on the right side of the screen (see Figure A.16). While holding down the left mouse button drag the cursor halfway down the screen. Now you have two screens and the rows are independent in each, in other words the top screen can show lines 1-15 and the bottom segment can show 105-120, for the same columns.

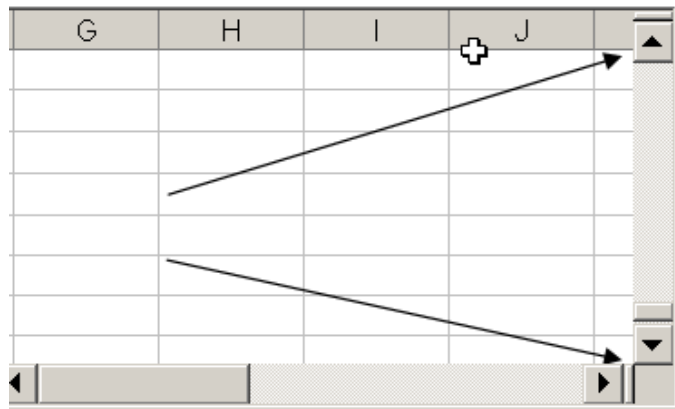


Figure A.16. Location of Short Bars to Split Screen Horizontally and Vertically.

The screen can be split into two vertical segments by left clicking on the small vertical bar on the right hand side of the horizontal scroll bar (see Figure A.16) and dragging the cursor to the left. The rows are constant but you can view different columns in each segment of the screen.

To see four parts of the Worksheet at once split the screen vertically and horizontally using the steps described above. Another way to split the screen in four parts is to click on a cell in the center of the Worksheet and click **Window > Split**. The horizontal and vertical lines can be moved so they are more convenient by left clicking on the thick line and dragging it to the desired place on the screen. Remove a line by double clicking it.

View Two Worksheets

Viewing two Worksheets in the same Workbook at once can be very useful. To accomplish this follow these steps:

- Only have one Workbook open.
- **Windows > New Window**
- **Windows > Compare Side by Side with ...** the name of your open Worksheet, such as Model. This gives you two active windows of the initial open Worksheet (Model).
- Click the Worksheet name in the Worksheet tabs list that you want to compare as Stoch. Now you have two screens, one has Worksheet Model and the other has Stoch.
- Closing one of the Worksheets returns the screen to displaying only one Worksheet.

View Two Workbooks

To view two Workbooks simultaneously in two screens do the following:

- Open two Excel Workbooks, say Test A and Book 2.
- Assuming Test A is open then click on Window > Compare Side by Side with Book 2.

Both Workbooks will have their own scroll bars, grid names and you can navigate independently in the Workbooks.

Preparing Charts

Excel provides the Chart Wizard menu for developing and refining charts. The line, fan, and CDF chart functions in Simetar are provided to bypass the steps described here. However, you should be aware of Excel's capabilities so you can customize charts created with Simetar. The following section describes how you can develop a CDF chart with Excel.

- Put the X axis values in the left hand column, and put the Y values in the right hand column. The columns must be adjacent, as in Columns B and C. For a CDF chart the Y axis has values from 0 to 1 in ascending order and the values in X are the sorted values for the random variable matching the probabilities in the Y data.
- Highlight the range of values in the two columns to be included in the graph and select: Insert > Chart ... > XY Scatter (see Figure A.17).
- XY Scatter offers five types of graphs: just points, smooth lines and points, smooth lines, straight lines with points, and straight lines. Select the graph type with smooth lines and no point markers, then select Next.
- Excel then shifts to Step 2 of the Chart Options, showing you basically what you will get. If necessary select the < Back button to make a different selection at this point.

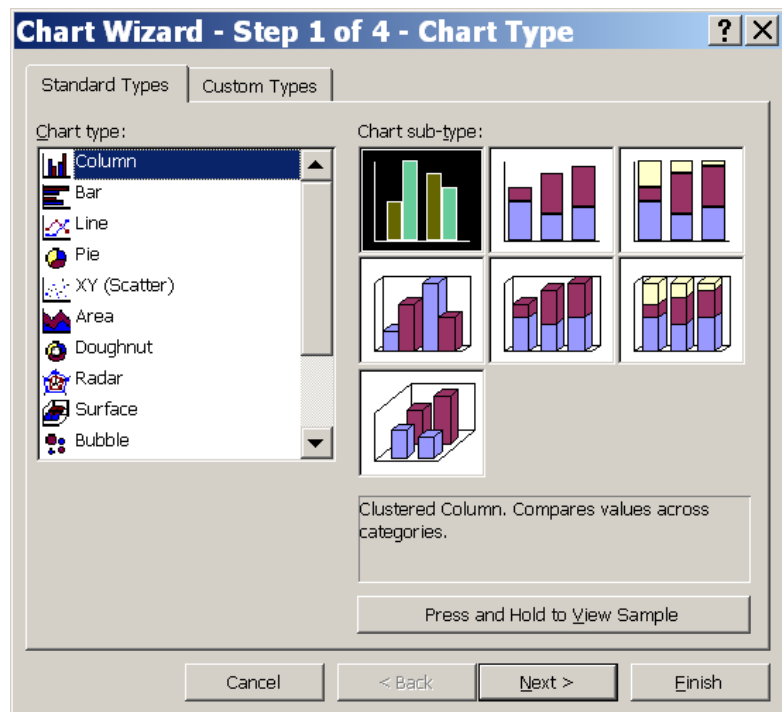


Figure A.17. Example of Using Chart Wizard to Develop a Chart.

- Select the Series tab at the top of the Chart Wizard menu. This allows you to type in the name of the variable in the field to the right of Name.
- You may add a second or third line to the graph by clicking Add or you can remove

- a line by clicking Remove.
 - Select Next > when you are through with the Series options.
- Excel then moves to Step 3 of the Chart Options with a new menu.
- Titles tab provides a mechanism for typing in a name for the chart and for typing names for the X and Y axes.
 - Axes tab allows you to specify whether the X and Y axes will have values printed.
 - Grid Lines tab allows you to specify whether X and Y grid lines will be included or left out.
 - Legend tab allows you to hide the legend or to position it around the graph (right, left, corner, top, bottom).
 - Data Labels tab allows you to specify whether the actual values for the line appear on the line. Show Value places the Y values on the line and Show Label places the X values on the line. None hides all numbers.
 - Select Next > when you are through with these selections.
- Excel then moves to Step 4 of the Chart Wizard. Here you have two options, to either make the chart into a new sheet called Chart 1 or to add it as an object to your current Worksheet.
- As a new sheet the chart will appear on a full screen in a new Worksheet named Chart 1. This is a big picture of the chart that can be viewed easily and edited further.
 - As an object in the current sheet, the chart is smaller (but can be resized) and sits on the current Worksheet. It can be positioned any place on the Worksheet and printed along with the data. The chart can be sent to a new Worksheet by right clicking the chart and clicking on the option to change location.

Editing a chart gets complicated so I will only describe the basic things you can do to edit a chart.

- Change the scale on the Y or X axes.
- Right click the Y or X axis and select Format Axis.
 - Scale tab -- shows the minimum and maximum values (Figure A.18). Type in the new maximum and minimums you want on the chart. You can also set the increments printed on the axes.
 - Number tab -- (in Figure A.18) allows you the opportunity to fix the format of the numbers on the axis, including the number of decimal places.
 - Font size and type are changeable using the Font tab in Figure A.18.

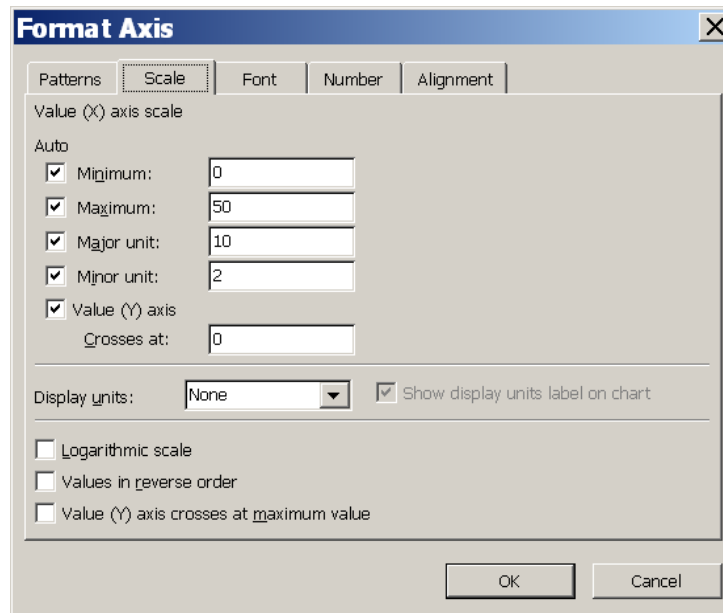


Figure A.18. Example of Setting Scale for the Chart Axis.

- Set the color for the chart.
 - Right click in the interior of the chart without the arrow touching a grid line or the border. Select Format Plot Area ...
 - Click the new color for the chart and OK. Fill effects can be added.
- Change the chart title.
 - Right click the chart title and select the Font, Pattern or Alignment tabs.
 - To change the words in the title, left click on the title, place the cursor where the insert is to be made and type or delete the title. When you are finished click the mouse on any cell outside the chart.
- Change the names of the labels on the lines.
 - Right click in the outer area of the Chart. Select the Source Data ... option. Select the Series tab in the resulting Source Data dialog box.
 - Click on a series name, as Series 1 and then type a new name in the Name box on the right side of the dialog box. Repeat this for every line on the chart. Rather than typing in the name, a cell reference of the variable's name can be assigned via the name cell.
 - Select OK when you are finished.
- Add a line in a chart.
 - Right click in the outer area of the chart. Select the Source Data ... option.
 - Click the Add button in the Source Data dialog box.

- Specify the X values to add by clicking on the miniature spreadsheet icon to the right of X values.
 - Use the mouse to highlight the new column of X values to add to the graph and then press Enter.
 - Repeat the last two steps to specify the Y values to add to the chart.
 - Add a name for the new line by typing a name into the Name text window.
 - Select OK when you are finished.
- Delete a line from a chart.
- Left click on a line in the chart and press the Delete key.
- Change Color, Size, and Markers for a line
- Right click on the line you want to change. Select the Format Data Series ... option. This results in the Format Data Series dialog box appearing on the screen (Figure A.19).
 - Patterns tab allows you to customize the color, the size and the style of the line. Also you can specify a marker, its size and color to add to the line.
 - Series Order tab in Figure A.19 gives you a method for moving the line's name up or down in the legend.
 - Select OK when you are finished.

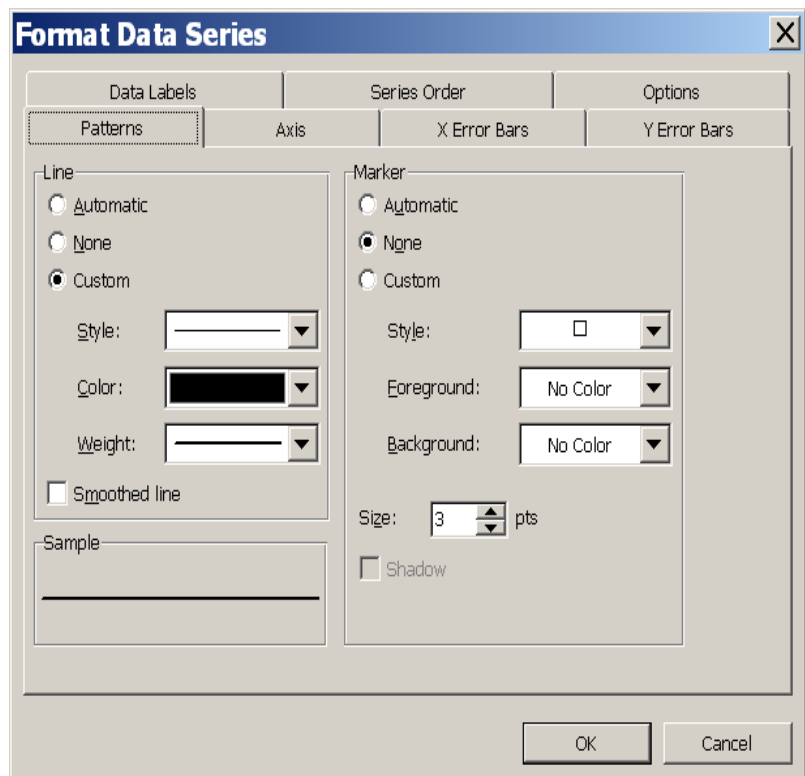


Figure A.19. Dialog Box for Formatting Lines in Charts.

Drawing Toolbar

Placing drawn objects on a Worksheet is easy with Excel's Drawing Toolbar (Figure A.20). A flowchart or a project management chart can be added to a Worksheet using the "Auto Shapes" icon on the Drawing Toolbar. A list of shapes and connectors that Excel can put on the Worksheet is accessed by:

Auto Shapes > Connectors

Auto Shapes > Basic Shapes



Figure A.20. Excel's Drawing Toolbar.

To construct a flowchart, select a shape using Auto Shapes > Basic Shapes > and then click one of the shapes desired, say a rectangle. After clicking the desired shape, the menus disappear and the pointer (cursor) turns into a small black cross. Place the black cross on the Worksheet, press the left mouse button and drag the mouse down and to the right to make the object the size you want. After releasing the left mouse button, you may type text and it will appear in the object. Clicking the object allows you to resize, edit the text and move the object. If a particular object, say a rectangle is to be used 10 times, a shortcut is to draw the object, copy it with the copy icon on the main toolbar, and press the paste button nine times.

Drawing lines and arrows to connect the objects on the Worksheet, is demonstrated in Figure A.21, is done by first selecting a connector. Select a connector by selecting Auto Shapes > Connectors > and then click on the desired connector. The menu will disappear and a black cross appears on the screen. Move the black cross to the object

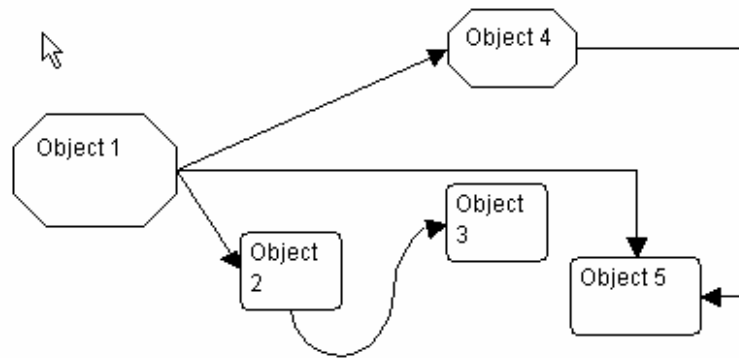


Figure A.21. Example of a Drawn Flowchart.

that the connector will go from and blue cubes will appear on each side of the object. Place the black cross on the side you want the connector to appear on and press the left mouse button; while holding the left mouse button down drag the cursor to the receiving object. In Figure A.21 this would involve dragging the connector from object 1 to object 2. Repeat the Steps to select the appropriate connector between objects 2 and 3. As demonstrated in Figure A.21 an object can have multiple connectors coming into or out of an object.

Excel drawn objects actually are on the surface of the Worksheet and can be moved all around. Select an object to move by placing the pointer on the object and pressing the left mouse button. While holding the left mouse button down, drag the object where it should be placed. Notice that when this is done, the connectors automatically adjust. The connectors are also redrawn automatically when an object is re-sized.

If you decide to change the shape of a drawn object, say from a rectangle to a circle, select (click on it) the object and select: Draw > Change Auto Shape > Basic Shapes > then select the new object. This series of steps will change the shapes of the objects after the chart has been drawn and you know what you want. Changing a drawn object's shape automatically changes the objects connectors.

Text can be added to an object by right clicking on the object and selecting Add Text.

Once the chart is completed, it needs to be grouped as a single object. This is done by selecting the arrow icon on the Draw Toolbar (Figure A.21), use the left mouse button and click on the first object, then hold down the Shift key and click on each of the remaining objects, release the Shift key and press the right mouse button, select “Grouping”, and finally select “Group” in the menu. Once a chart has been grouped as a single object, you can move and resize the whole chart as if it were one object.

I recommend that the completed chart or object be formatted so it does not move or size with the cells. This is done by selecting the object, clicking the right mouse button, selecting “Format Object”, selecting “Properties”, and then selecting “Don’t move or size with cells.”

Key Strokes to Save Time

- Copy highlighted text, while holding down the Ctrl key press C. This process holds for several different key strokes Ctrl C
- Paste Ctrl V
- Bold highlighted text Ctrl B
- Undo the last keystroke Ctrl Z
- Save the Workbook Ctrl S
- Cut highlighted text Ctrl X
- Format highlighted cells Ctrl 1
- Spell checker opened by pressing F7
- Help opened by pressing F1

Appendix B: Setting Excel's Options and Customizing the Tool Bar

Setting Options for Excel

The options are set by selecting Tools > Options. This brings up a menu with eight tabs, five of which are described in this section (Figure B.1).

– View Tab

By clicking on the buttons you should turn on the following options:

- Formula bar
- Status bar
- Windows in taskbar
- Comment indicator only
- Show all objects
- Page breaks off
- Gridlines on
- Row and column headers on
- Outline symbols on
- Zero values on
- Horizontal scroll bar on
- Vertical scroll bar on
- Sheet tabs on
- Color: Automatic

– Calculation Tab

Under the Calculation tab turn on the following options by clicking on the appropriate buttons.

- Automatic calculation
- Update remote references
- Save external link values

It is strongly recommended that calculation be set to automatic if you are using Simetar. This setting prevents issues with certain functions and calculating statistics and names for simulation results in SimData.

– Edit Tab

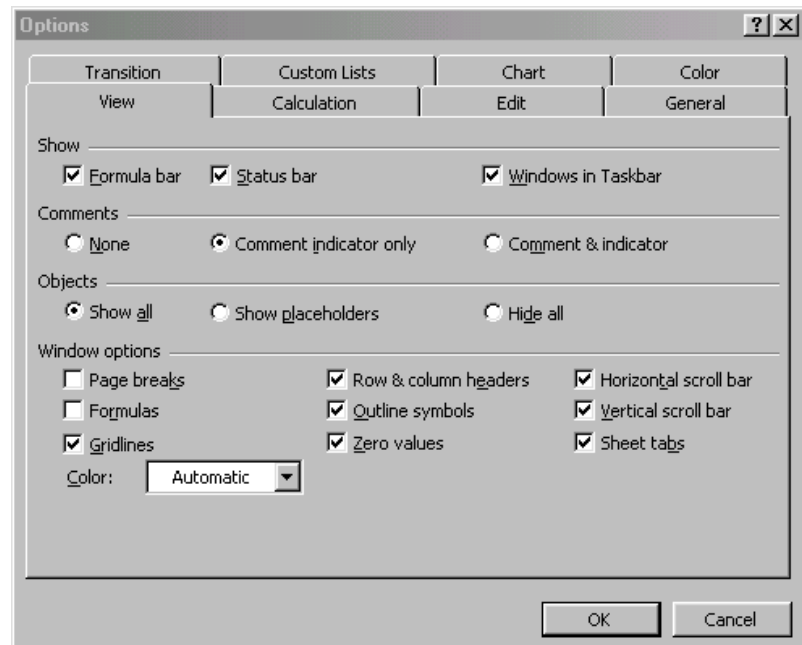


Figure B.1. Excel's Options Setup Dialog Box.

In the Edit tab turn on all of the options except:

- Fixed decimal places
- Move selection after enter

– General Tab

The folder you intend to use for your Excel programs can be indicated here so saving and retrieving Excel workbooks will be made easier.

- Set the recently used file list to “9” entries.
- Set the sheets in a new workbook to “3”.
- Type in the default folder name you want to use for all of your Excel programs.

– Transition Tab

- Save Excel files as: “Microsoft Excel Workbook.”
- Set Microsoft Excel Menus to on.

Customizing the Excel Tool Bar

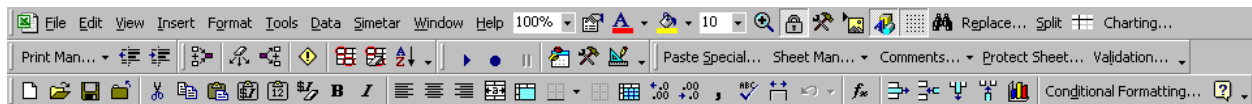


Figure B.2. Example of a Customized Excel Toolbar for Simulation.

You can create your own customized toolbar by following the steps described in this section. To change the toolbar click Tools > Customize > Commands, which generates Figure B.3. Then select the Category you want to select from. Next click on a command icon in the Commands menu and drop it to the Toolbar where you want it. Various command icons useful for simulation that are described below appear on the customized Toolbar in Figure B.2 and all were added to the basic Excel Toolbar using the steps described below.

– Add Auditing Icons to Toolbar

The audit icons are under Tools (see Figure B.3) assist you in tracing which cells are used to calculate the value in a target cell, and where a particular cell's

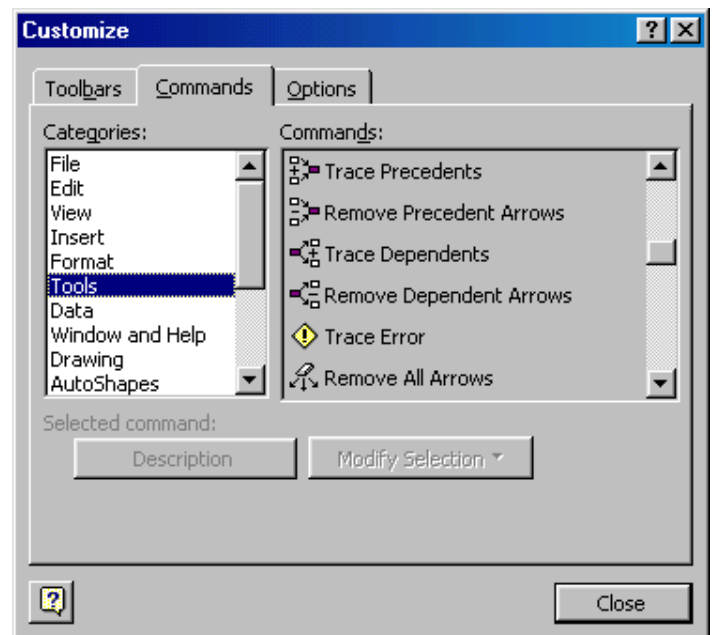


Figure B.3. Example of Excel's Dialog Box for Customizing the Toolbar.

values are used in subsequent calculations. Add the audit icons to the Toolbar as follows:

Tools > Customize ...

Scroll down in the Categories menu and click on Tools (see Figure B.3). Scroll down the Commands list until the icons in Figure B.3 appear. Click on “Trace Precedents” and drag it to the Toolbar, next click on the eraser “Remove All Arrows” and drop it on the Toolbar, and finally drag the “Trace Dependents” to the Toolbar.

After you close the Customize screen the three new icons remain on your Toolbar. To remove an icon reverse the steps.

– **Add Cell Comment Icons to Toolbar**

Icons to insert comments, show all comments and hide all comments can be added to the Toolbar by following these steps:

Tools > Customize ... > Toolbars > Reviewing

This action will place several icons on your Toolbar. The new icons maybe more than you need, so you can shorten it by clicking on an unneeded icon and dropping it on the spreadsheet.

– **Print Icons**

Printing in Excel is a multiple step process so creating a special menu for printing can save some time. You can also create other special menus that can contain multiple icons/commands. Develop and populate a special print menu as follows:

Tools > Customize ... > Categories > New Menu

Scroll the bottom of Categories and click on “New Menu.” Drag the “New Menu” in the Commands menu to the Toolbar where you want it to stay. Right click on the “New Menu” icon on your Toolbar and rename it by typing “Print Options” in the “Name:” box.

Left click on the “Print Options” menu and a small empty box appears. File icons into this empty box as follows:

Tools > Customize ... > File > Set Print Area

Drag the Set Print Area icon to the empty box under Print Options. This is now the first command in your special menu. Next drag File > Print Preview to the Print Options menu, followed by Print, Page Break Preview, Page Setup ..., and Header and Footer.

– **Icons to Add and Delete Rows and Columns**

Adding/deleting a row or a column in a spreadsheet should be a one click operation. This can be done by adding four icons to the Toolbar.

Tools > Customize ... > Commands > Insert

Drag the Rows icon to the Toolbar and drop it. Repeat this action for the Columns icon.

Tools > Customize ... > Commands > Edit

Scroll down until you find the Delete Rows and Delete Columns icons. Drag these two icons to the Toolbar and drop them just before or just after their respective column icons.

– Add Paste Special Icons

Pasting a cell to a new location can be done as a formula (the default) or as a value. The Paste Special Values icon is useful to have on the Toolbar.

Tools > Customize ... > Edit

Scroll down to Paste Values and drag it to the Toolbar. Add another edit icon “Clear Formatting” to the Toolbar. This icon is the fastest way to remove lines and special formats for a single cell or a block of cells.

– Add Validation Icon

Tools > Customize ... > Commands > Data

Scroll down to the word/icon Validation and drag it to the Toolbar.

From the Toolbar, clicking the Validation icon opens a menu screen that allows you to specify the type of input a cell can contain. For example, the decimal > Between option allows you to restrict the data typed into the cell to be within a range, say 0.0 to 1.0. As you see this feature can be added to a Worksheet cell to help users enter the input data in the correct format.

– Take Junk Off the Toolbar

The basic Excel Toolbar comes with some icons that could be removed without hurting your productivity. To remove these icons,

Tools > Customize ... > Commands

Next click on the offending icons and drag and drop them on the worksheet. If you need them later they can be restored to the Toolbar. Icons that are good candidates for elimination are:

- Print (a little printer icon fourth from the left).
- Format Painter (the paint brush)
- Redo (a right hand arrow)
- Insert Hyperlink (globe with a chain link)
- Web Toolbar (globe with two arrows)
- Map (globe on a stand)
- Drawing

– Save Your Custom Toolbar for Windows 2000 and XP

After a custom Excel toolbar is created, Windows 2000 and XP saves it into a personal directory that loads up when you log on. All custom settings are stored in this directory so if you log in as the Administrator your toolbar will not be in place. The personalized directory for settings and toolbars is:

C:\Documents and Settings*your name*\Application Data\Microsoft\Excel

Toolbar changes are stored to this file and named Excel.xlb. It is recommended that you keep a backup of your xlb file.

References

Leonhard, W., L. Hudspeth, and T.J. Lee. Excel 97 Annoyances. Sebastopol, CA: O'Reilly & Associates, Inc., 1997.