Agribusiness Analysis and Forecasting

Simulation Basics

Aleks Maisashvili

Texas A&M University

Stochastic Simulation

In economics we use simulation because we can not experiment on live subjects, a business, or the economy without injury.

In other fields they can create an experiment

- Health sciences they feed (or treat) lots of lab rats on different chemicals to see the results.
- Animal science researchers feed multiple pens of steers, chickens, cows, etc. on different rations.
- Engineers run a motor under different controlled situations (temp, RPMs, lubricants, fuel mixes).
- Vets treat different pens of animals with different meds.
- Agronomists set up randomized block treatments for a particular seed variety with different fertilizer levels.

Probability Distributions

Parametric and Non-Parametric Distributions

- Parametric Dist. have known and well defined parameters that force their shapes to known patterns.
  - Normal Distribution - Mean and Standard Deviation.
  - Uniform - Minimum and Maximum
  - Bernoulli - Probability of true
  - Beta - Alpha, Beta, Minimum, Maximum
- Non-Parametric Distributions do not have pre-set shapes based on known parameters.
  - The parameters are estimated each time to make the shape of the distribution fit the data.
  - Empirical – Actual Observations and their Probabilities.

Typical Problem for Risk Analysis

- We have a stochastic variable that needs to be included in a business model. For example:
  - Price forecast has residuals we could not explain and they are the stochastic component we need to simulate.
  - Crop yield is forecasted by trend but it has residuals that are stochastic; risk caused by weather.
- Model will be solved (sampled) many times using alternative draws of random values for prices and yields.
- We have the data and a forecast model, next we need to estimate parameters to define the stochastic variables.
  - NOTE: Parameters is the generic name for values that determine the location and shape of the distribution.

Steps for Simulating Random Variables

- First step: be certain that the variable that you will directly scholastically draw is suitable
- Every stochastic draw you will make for a variable will be independent of every other draw, even for the same variable in different time periods.
- The properties of the variable must be consistent with this simulation process.
- In short, we need all draws for an individual variable to be *independently, identically distributed* (i.i.d.).
- We must therefore be certain that the variables we directly simulate have
  - Constant mean
  - Constant, finite variance
  - No autocorrelation

Steps for Simulating Random Variables

- For parametric distributions, we must make an assumption on a probability distribution for the random variables (e.g., Normal or Beta or Uniform...).
- Estimate/fit the parameters values to define the assumed distribution.
- Parameters for parametric distributions we will be using are:
  - Normal (Mean, Std Deviation)
  - Beta (Alpha, Beta, Min, Max)
  - Uniform (Min, Max)
  - Bernoulli (probability of true)

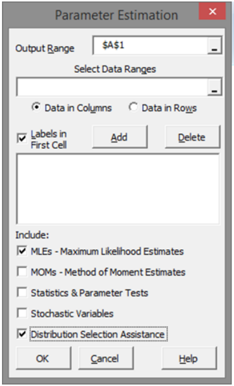
Steps for Parameter Estimation

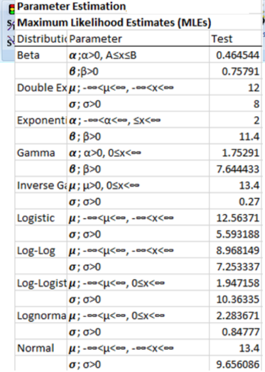
1. Again, be sure that you have removed any trend, cycle or structural pattern. Be sure that you have a constant mean and variance. i.i.d.!
2. Estimate parameters for several assumed distributions using historical data.
3. Simulate the data under different distributions.
4. Pick the best distribution based on.
   - Mean, Standard Deviation - use validation tests.
   - Minimum and Maximum.
   - Shape of the CDF vs. historical series.
   - Penalty function =CDFDEV() to quantify differences.

Parameter Estimator in Simetar

Use Theta Icon in Simetar

- Estimate parameters for 17 parametric distributions.
- Select MLE or MOM for parameter estimation.
- The tool provides ready-made cells simulating your variable under the various distributions.

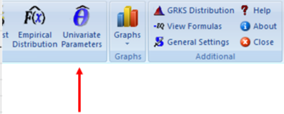




Parameter Estimation

Maximum Likelihood Estimates (MLEs)

Distributi	Parameter	Test
Beta	α ;α>0, A≤x≤B	0.464544
	β ;β>0	0.75791
Double Ex	μ ; -∞<μ<∞, -∞<x<∞	12
	σ ; σ>0	8
Exponenti	α ; -∞<α<∞, ≤x<∞	2
	β ; β>0	11.4
Gamma	α ; α>0, 0≤x<∞	1.75291
	β ; β>0	7.644433
Inverse G	μ ; μ>0, 0≤x<∞	13.4
	σ ; σ>0	0.27
Logistic	μ ; -∞<μ<∞, -∞<x<∞	12.56371
	σ ; σ>0	5.593188
Log-Log	μ ; -∞<μ<∞, -∞<x<∞	8.968149
	σ ; σ>0	7.253337
Log-Logist	μ ; -∞<μ<∞, 0≤x<∞	1.947158
	σ ; σ>0	10.36335
Lognorma	μ ; -∞<μ<∞, 0≤x<∞	2.283671
	σ ; σ>0	0.84777
Normal	μ ; -∞<μ<∞, -∞<x<∞	13.4
	σ ; σ>0	9.656086



Uniform Distribution

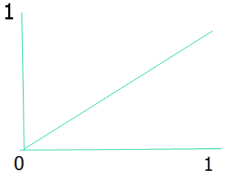
- Random variable where every interval has an equal probability of being observed (drawn).

  if X is Uniform(0, 1) then $P(0.1 < x < 0.2) = P(0.5 < x < 0.6)$

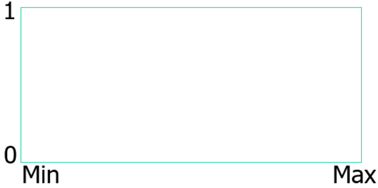
- Simulating Uniform in Simetar enter parameters as:
  - =UNIFORM(Minimum , Maximum)
  - =UNIFORM(0,1) which is the same as =UNIFORM() (this is **standard** uniform, or uniform standard deviate (USD))
  - =UNIFORM(10,25), etc.
- A standard uniform RV is used to simulate all distributions. For example a normal distribution:
  - =norm(mean, standard deviation, U), where U is distributed standard uniform.

Standard Uniform Distribution

- CDF of the Standard Uniform Distribution.



- PDF of Standard Uniform Distribution.

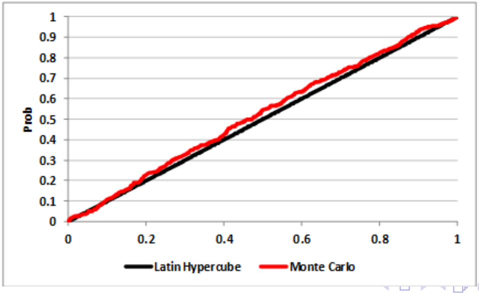


Basic Simulation Definitions

- Stochastic Simulation Model - means the model has at least one random variable.
- Monte Carlo simulation model - same as a stochastic simulation model.
- Two ways to sample or simulate random values:
  1. **Monte Carlo** sampling - draw random values for the variables purely at random.
  2. **Latin Hyper Cube** sampling - draw random values using a systematic approach so we are certain that we sample ALL regions of the probability distribution.
- Monte Carlo sampling requires larger number of iterations to insure that model samples all regions of the probability distribution.

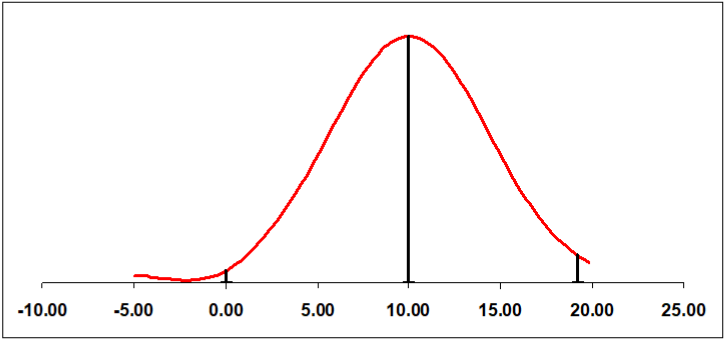
MC vs. LHC Sampling

- For a standard uniform random variable (uniform over the unit interval), the CDF is a 45-degree straight line.
- MC empirical CDF deviates from the 45-degree line.
- LHC empirical CDF is right on top of the population CDF.
- This is with 500 iterations.
- Simetar default is LHC.



When to Use the Normal Distribution

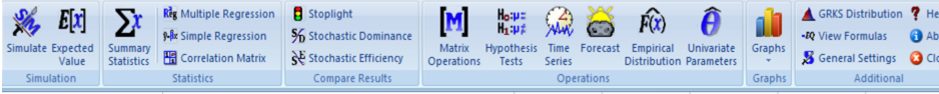
- Use the Normal distribution if you have lots of observations and have tested for normality.
- BUT watch for infeasible values from a *Normal* distribution (negative yields and prices).



How to Test for Normality

Simetar provides an easy to use procedure for testing Normality that includes:

- S-W (Shapiro-Wilk)
- A-D (Anderson-Darling)
- CvM (Cramer-Von Mises)
- K-S (Kolmogorov-Smirnov)
- Chi-Squared

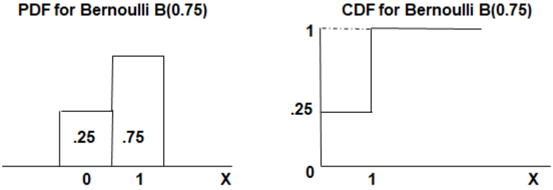


Test for Normality of Distribution for SOYBEAN			
Confidence Level		95.00%	
Procedure	Test Value	p-Value	
S-W	0.932020794	0.151105044	*Fail to Reject the Ho that the Distribution is Normally Distributed**
A-D	0.572620313	0.119703423	*Fail to Reject the Ho that the Distribution is Normally Distributed**
CvM	0.098783019	0.108520615	*Fail to Reject the Ho that the Distribution is Normally Distributed**
K-S	0.173747255	NA	*Consult Critical Value Table*
Chi-Sqared	12.52380952	0.862032304	*Fail to Reject the Ho that the Distribution is Normally Distributed**
			**Based on approximate p-values*

Truncated Normal

- General formula for the Truncated Normal
  =TNORM(Mean, Std Dev, [Min], [Max],[USD])
- Truncated Downside only
  =TNORM(10, 3, 5)
- Truncated Upside only
  =TNORM(10, 3, , 15)
- Truncated Both ends
  =TNORM(10, 3, 5, 15)
- Truncated both ends with a USD in general form
  =TNORM(10, 3, 5, 15, [USD])

Bernoulli Distribution



PDF and CDF for a Bernoulli Distribution.

- Parameter is p or the probability that the random variable is 1.0 or TRUE.
- Simulate Bernoulli as:
  =Bernoulli(p)
  =Bernoulli(0.25)

Bernoulli Distribution Application

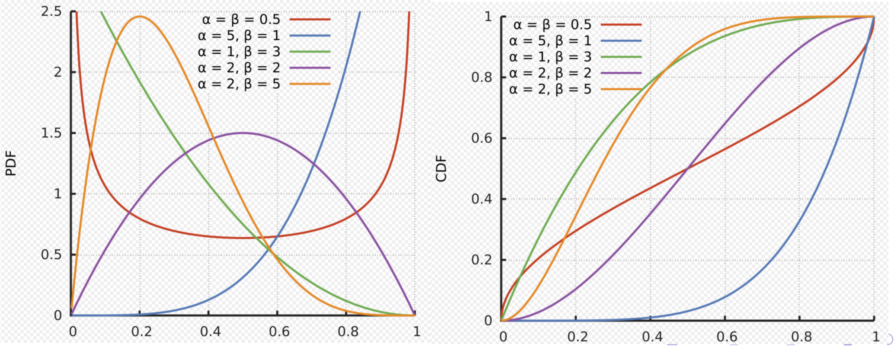
	A	B	C	D	E
13	Conditional Probability Distribution Example of Rain				
14	P(rain) in June	0.2			
15	Quantity of Rain IF it rains				
16	Min	2			
17	Max	5			
18	Use a Uniform distribution to simulate the amount of the rainfall				
19	Rainfall If it rained	3.728058	=UNIFORM(B16,B17)		
20					
21	Did it Rain?	1	=BERNOULLI(B14)		
22	This is the value we want to simualt				
23	If It Rained the Amount	3.728058	=B21*B19		
24	How we could use the stochastic rainfall value in a simulation model				
25	Assume a yield function for cotton that was Y = 400 + 15*Rainfall in June				
26					
27	Simulated Yield is	455.9209	=400+15*B23		
28	Press F9 several times to see the impact of random rainfall on yield				

Bernoulli Distribution Application

	A	B	C	D	E	F	G
32	Simulate Machinery Repair Costs						
33	Assume a 5% chance of a repair						
34	Repairs are $10,000, $20,000 or $30,000						
35	Bernoulli parameter	0.05					
36	Repairs costs range are:	10000	20000	30000			
37	If Repair is needed what is the stochastic repiar cost?				30000	=DEMPIRICAL(B36:D36)	
38	Repair?	1	=BERNOULLI(B35)				
39							
40	Simualted Repair Cost	30000	=B38*E37				
41	You must hit F9 about 22 times to get a vlue for simulated repair greater than zero.						
42	Think about it there is only a 5% chance of a reapir or 1 in 20 chance.						

Beta Distribution

- Beta is a continuous probability distribution.
- It is parametrized by two positive **shape parameters**, denoted by α and β.
- These two parameters define the shape of the distribution.
- Simulate *Beta* distribution using the function: =beta.inv(USD, alpha, beta, minimum, maximum)



Recap: Parametric vs. Non-Parametric Distributions

- Parametric Distributions
  - Fixed form, shape dependent on parameters.
  - Uniform, Normal, Beta, and Bernoulli.

- Non-Parametric Distributions
  - Not a fixed form that is parameter dependent.
  - Discrete Empirical, Empirical.

(Continuous) Empirical Distribution

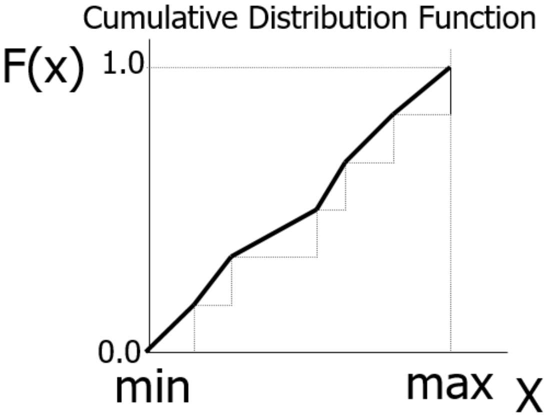
An empirical distribution is defined totally by the observed data for the variable. There is no assumed distributional shape.

Steps to simulate an empirical distribution.

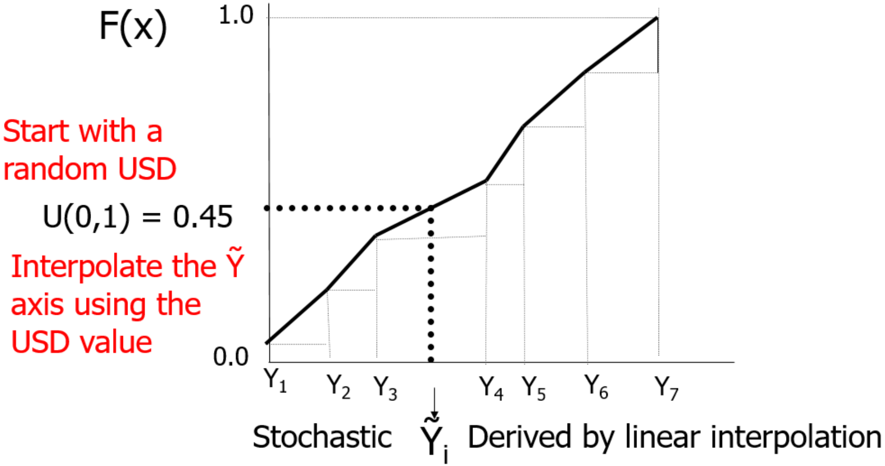
1. Sort the historical values from lowest to highest.
2. Assign a cumulative probability to the sorted deviates (usually assume equal probability for each value). Cumulative probabilities go from 0.0 to 1.0.
3. Assume the distribution is continuous, so interpolate between the observed points.
4. Use the Inverse Transform formula to simulate the distribution. This requires simulation of a standard uniform RV to use in the interpolation.

In Simetar: $=$EMPIRICAL($x_1, x_2, x_3, ...$)

CDF for an Empirical Distribution



Inverse Transform for Simulating an Empirical Distribution



Using the Empirical Distribution

- Empirical distribution should be used if
  - Random variable is continuous over its range.
  - You have fewer than 15 observations for the variable, and/or.
  - You cannot easily estimate parameters for a parametric dist.
- Suppose we have only 10 observed yields:
  - Yield can be any positive value, not discrete values.
  - We don't have enough observations to test for normality or other parametric distributions.
  - We know the 10 random values were observed with a probability of 1/10, or one observation each year.
  - So $F(x)$ goes from 0.0 to 1.0 in equal increments.

EMP Distribution

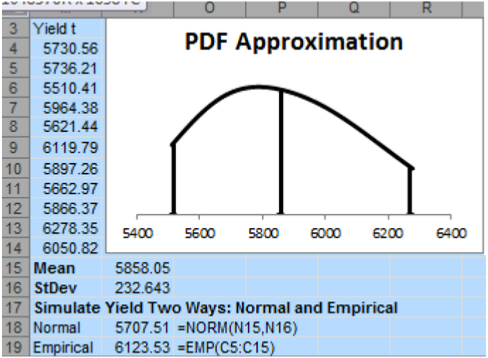
Advantages of EMP Distribution

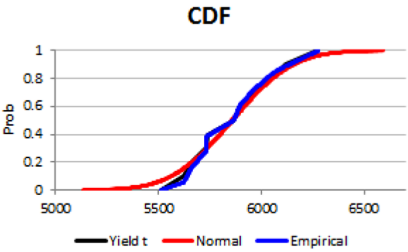
- It lets the data define the shape of the distribution.
- Does not risk assuming an incorrect parametric distribution.
- The larger the number of observations in the sample, the closer EMP will approximate the "true" distribution.

Disadvantages of EMP Distribution

- Small samples will, to some unknown extent, misrepresent the true shape of the population distribution.
- It has finite min and max values; quite possibly missing the tails of the actual underlying population distribution.
- May overfit the data.

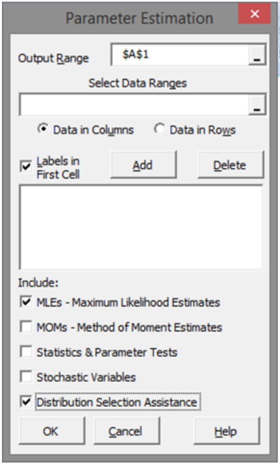
Empirical Dist. vs. True Population Dist.

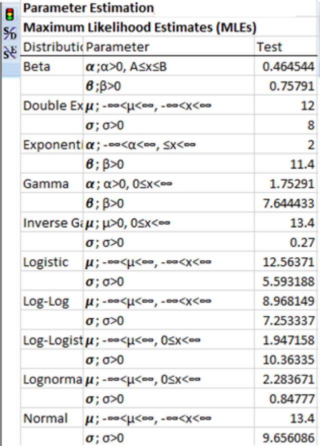




Parameter Estimation with Theta

Select MLE or MOM

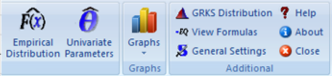




Parameter Estimation

Maximum Likelihood Estimates (MLEs)

Distributi	Parameter	Test
Beta	α;α>0, A≤x≤B	0.464544
	β;β>0	0.75791
Double Ex	μ; -∞<μ<∞, -∞<x<∞	12
	σ; σ>0	8
Exponenti	α; -∞<α<∞, ≤x<∞	2
	β; β>0	11.4
Gamma	α; α>0, 0≤x<∞	1.75291
	β; β>0	7.644433
Inverse Ga	μ; μ>0, 0≤x<∞	13.4
	σ; σ>0	0.27
Logistic	μ; -∞<μ<∞, -∞<x<∞	12.56371
	σ; σ>0	5.593188
Log-Log	μ; -∞<μ<∞, -∞<x<∞	8.968149
	σ; σ>0	7.253337
Log-Logist	μ; -∞<μ<∞, 0≤x<∞	1.947158
	σ; σ>0	10.36335
Lognorma	μ; -∞<μ<∞, 0≤x<∞	2.283671
	σ; σ>0	0.84777
Normal	μ; -∞<μ<∞, -∞<x<∞	13.4
	σ; σ>0	9.656086

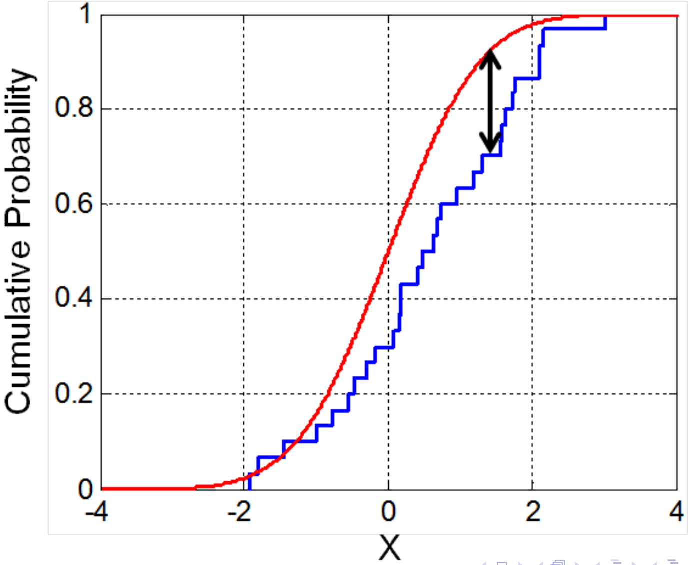


Comparing distributions

- CDFDEV is a Simetar function to compare the CDFs of two data samples
- CDFDEV calculates the integral between two distributions with a penalty for the two distributions being different.

$$\int_{-\infty}^{\infty} (F_1(x) - F_2(x))^2 + w(x)dx$$

Comparing distributions



Calculating CDFDEV

- Create simulated samples from candidate distributions
- Use CDFDEV to compare those simulated samples to the historical data sample: =CDFDEV(historical_sample, simulated_values)
- Select the distribution that has the lowest CDFDEV value.

Random Variables (MLE)

Distribution	Random Va
Beta (MLE)	2.0697062
Double Exponential (ML	2.1953671
Exponential (MLE)	1.8618588
Gamma (MLE)	2.1173946
Inverse Gaussian (MLE	2.1360436
Logistic (MLE)	2.1454888
Log-Log (MLE)	2.0783888
Log-Logistic (MLE)	2.1389988
Lognormal (MLE)	2.1076268
Normal (MLE)	2.1367503
Pareto (MLE)	1.8124891
Uniform (MLE)	2.0971311
Weibull (MLE)	2.1500771
Binomial (MLE)	2
Geometric (MLE)	2
Poisson (MLE)	2
EMP	2

Distribution	CDFDEV
Beta (MLE)	0.0178771
Double Exponential (MLE)	0.3655717
Exponential (MLE)	2.7087809
Gamma (MLE)	0.067896
Inverse Gaussian (MLE)	0.1142559
Logistic (MLE)	0.1606926
Log-Log (MLE)	0.4418301
Log-Logistic (MLE)	0.3453194
Lognormal (MLE)	0.100735
Normal (MLE)	0.0497376
Pareto (MLE)	79.214829
Uniform (MLE)	0.0324191
Weibull (MLE)	0.0811444
Binomial (MLE)	0.9668741
Geometric (MLE)	59.505363
Poisson (MLE)	6.3777462
EMP	0.0003111

What is the Next Step?

After choosing a distribution:

- It's a good idea to validate that the characteristics of the simulated data match those of the original historical data.
- Use statistical tests to check that the means and variances are not significantly different from one another.
- Check if the minimum and maximum values are realistic.
- Can also visually check the shape of the CDF and PDF.

Statistical Tests for Validation

Student t test

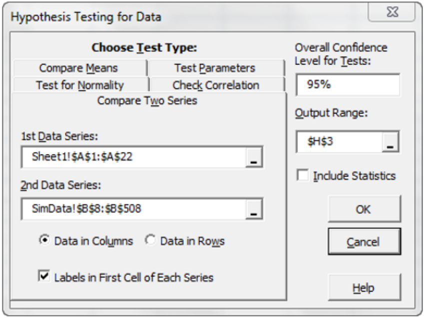
- H_o: Historical Mean $=$ Simulated Mean.
- H_a: Historical Mean $\neq$ Simulated Mean.

F test

- H_o: Historical Variance $=$ Simulated Variance
- H_a: Historical Variance $\neq$ Simulated Variance.

Validation Tests in Simetar

- Compare Two Series: Historical Data vs. Simulated Values
  - 1st Data Series is history
  - 2nd Data Series is simulated
- Simetar Icon is



Distribution Comparison of Normal Corn price & Corn price				
Confidence Level		95.0000%		
	Test Value	Critical Val	P-Value	
2 Sample t Test	0.00	2.69	0.999	*Fail to Reject the Ho that the Means are Equal*
F Test	1.00	1.90	0.437	*Fail to Reject the Ho that the Variances are Equal*