Agribusiness Analysis and Forecasting

Univariate Probability Distributions

Henry Bryant

Texas A&M University

Parametric vs. Non-Parametric Distributions

- Parametric Distributions
  - Fixed form, shape dependent on parameters.
  - Uniform, Normal, Beta, and Bernoulli.
- Non-Parametric Probability Distributions-not a fixed form that is parameter dependent, for example:
  - Discrete Empirical
  - Empirical

Discrete Empirical

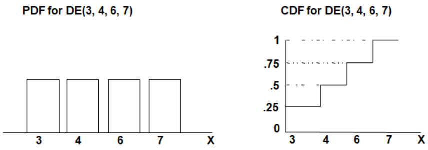
Discrete Empirical distribution is used where only fixed values can occur.

- Each value has a probability of being drawn equal to its historical rate of occurrence.
- No interpolation between observed values.

Examples:

- Number of students present for class
- Simulating a die: 1, 2, 3, 4, 5, 6
- Number of births per year

Discrete Empirical Distribution



PDF and CDF for a Discrete Uniform Distribution.

- Use function DEMPIRICAL in Simetar
- Argument is a range of cells containing historical observations $(x_1, x_2, x_3, \ldots, x_n)$

Row	A	B	C
1	10		
2	12		
3	20	=DEMPIRICAL (A1:A5)	function in Simetar
4	15		
5	13		

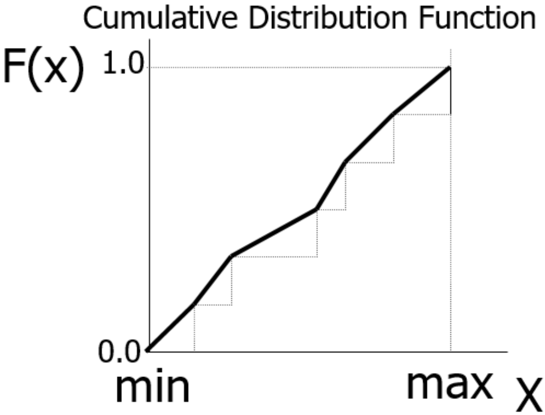
(Continuous) Empirical Distribution

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

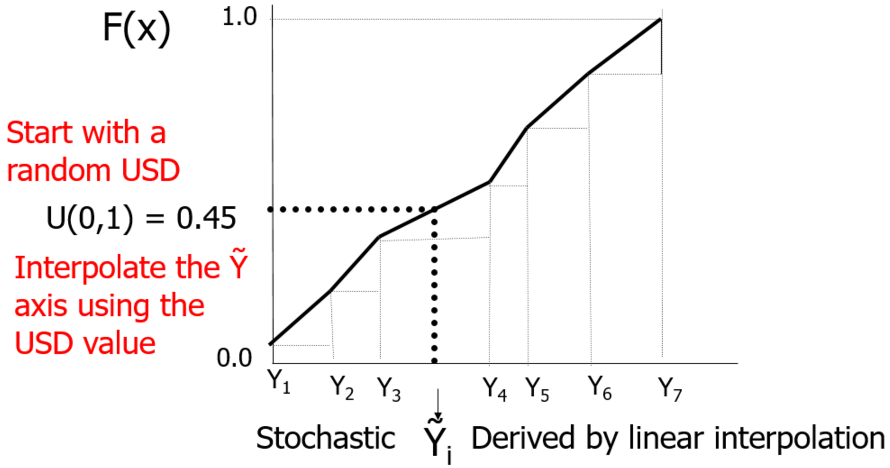
Steps to simulate an empirical distribution.

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 unifrom RV to use in the interpolation.
5. In Simetar: =EMPIRICAL($x_1, x_2, x_3, \ldots$)

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 20 observations for the variable, and/or.
  - You cannot easily estimate parameters for a parametric dist.
- Example: simulate crop yields given fewer than 20 historical values.
- 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

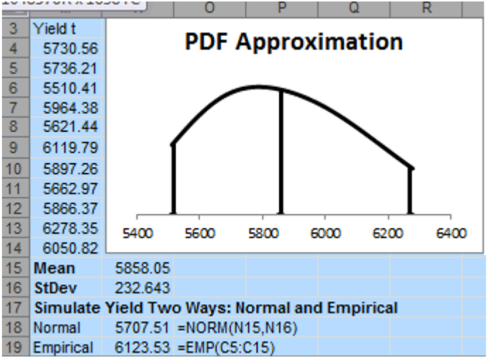
Advantages of EMP Distribution

- 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.

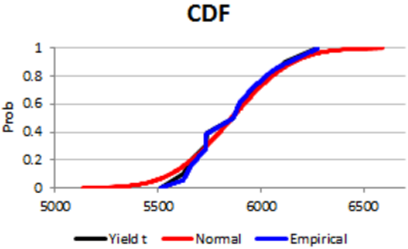
Disadvantages of EMP Distribution

- 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

Empirical Dist. vs. True Population Dist.

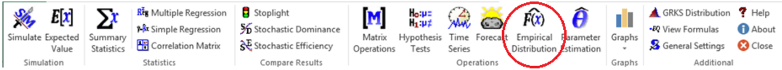


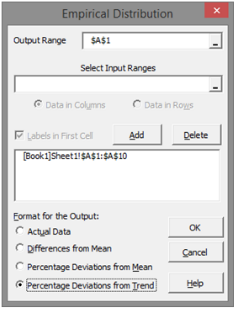
3	Yield t		
4	5730.56		
5	5736.21		
6	5510.41		
7	5964.38		
8	5621.44		
9	6119.79		
10	5897.26		
11	5662.97		
12	5866.37		
13	6278.35		
14	6050.82		
15	**Mean**	5858.05	
16	**StDev**	232.643	
17	**Simulate Yield Two Ways: Normal and Empirical**		
18	Normal	5707.51	=NORM(N15,N16)
19	Empirical	6123.53	=EMP(C5:C15)



Warning

Do NOT use Simetar's "Empirical Distribution" button. The purpose of this is to allow extra options for EMPIRICAL to work around a non-constant variance (e.g., assuming only a constant CV, not a constant variance) and/or non-constant mean. This is confusing at best, and hiding inappropriate methodology at worst. Work only with covariance stationary RVs.





Sorted Deviations from Trend as a Percent of Predicted

F(x)	Data
0	-0.61246
0.05	-0.6124
0.15	-0.51895
0.25	-0.39698
0.35	-0.20561
0.45	-0.17857
0.55	-0.09589
0.65	0.174699
0.75	0.316489
0.85	0.650367
0.95	0.864407
1	0.864493

Compound Growth and Stationarity

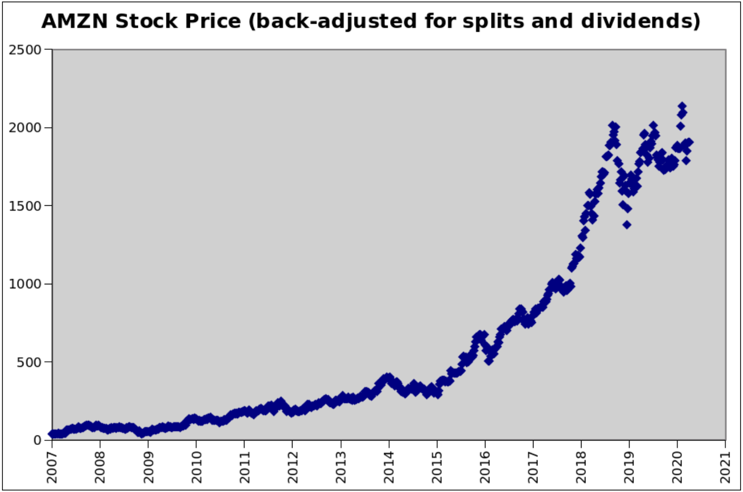
Many economic and biophysical phenomena reflect compound growth

- Measures of economic growth
- Financial prices

This results in heteroskedacity (non-constant variance) of deviations for a RV with compound growth. This causes econometric problems (for example, for fitting a trend), and means we do not have a constant variance for simulating the deviations.

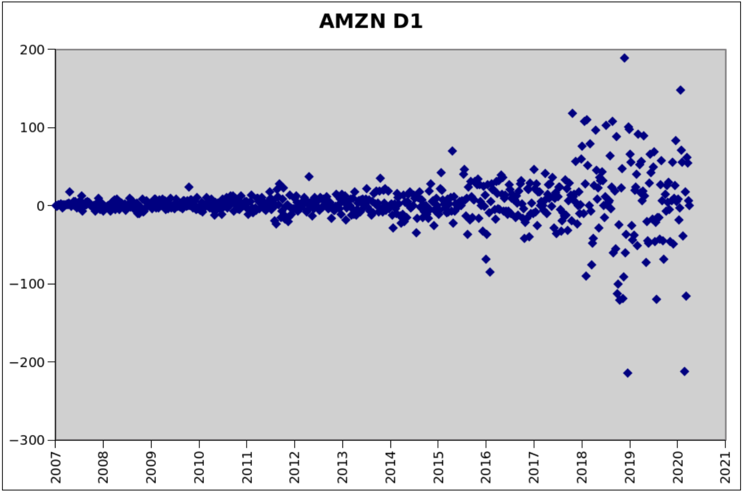
Standard solution: transform such variables using the natural logarithm function.

Example: AMZN stock price



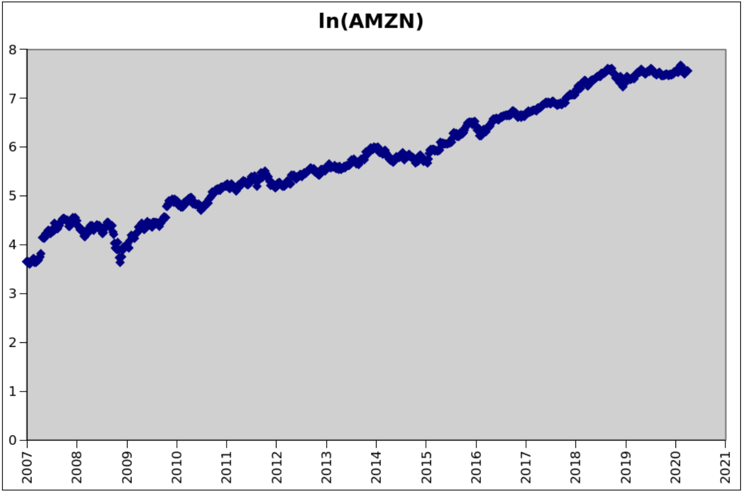
Notice the distinctly non-linear trend

Example: AMZN stock price



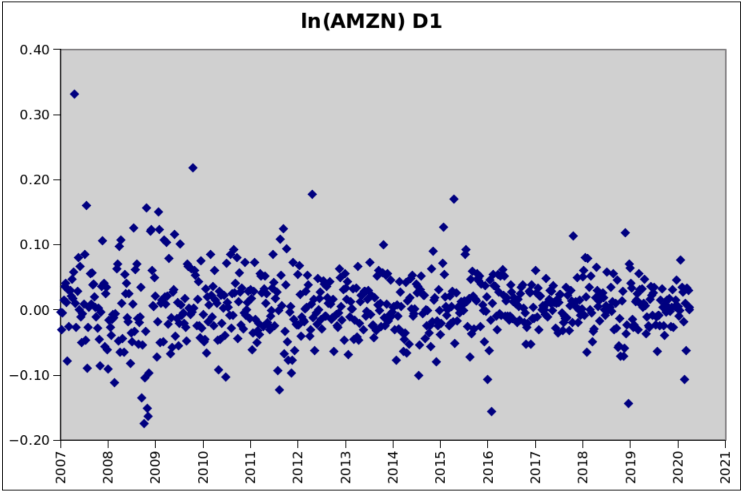
Behold the heteroskedasticity

Example: AMZN stock price



DF stat for manually de-trended series: -3.49 !!!

Example: AMZN stock price



Much more homoskedastic

Example: AMZN stock price

Lessons:

- Always plot your data
- Compound/exponential growth typically leads to non-constant variance
- Percentage deviations from a linear trend do not appropriately fix this problem
- Natural logarithm transformation often a good approach
- Always be sure that the random variables you are simulating are covariance stationary; do not rely on (potentially inappropriate) automated band-aids
- Understand exactly what your software is doing