Probabilistic Approach to Oil and Gas Prospect Evaluation Using the Excel Spreadsheet

Introduction

Oil and gas exploration is arguably the riskiest of all commercial activities. As a result, the utilization of probability and statistics in the oil and gas industry is becoming widely accepted as a method to estimate oil and gas exploration prospect size. Analysis typically utilizes the Monte Carlo simulation method and one of the commercial statistical packages such as @Risk™ or Crystal Ball™ (for information on these statistical Excel™ add-in programs, see http://www.crystalball.com or http://www.palisade.com). An exhaustive discussion of the Monte Carlo method can be found in Decision Analysis for Petroleum Exploration by Paul Newendorp and John Schuyler. An interesting collection of articles addressing exploration risk can be found in The Business of Petroleum Exploration published by The American Association of Petroleum Geologists, Tulsa, Oklahoma.

These sophisticated commercial programs are necessary for complex statistical analysis. Many of the simple analyses desired for oil and gas exploration, however, can be accomplished using the Excel spreadsheet program. Explorationists typically use "normal", "lognormal", and "triangular" statistical distributions. Utilizing built in Excel functions and add-ins, and some customized Excel functions, the experienced spreadsheet user can perform simple Monte Carlo analysis without the expense and need to learn one of the commercial programs.

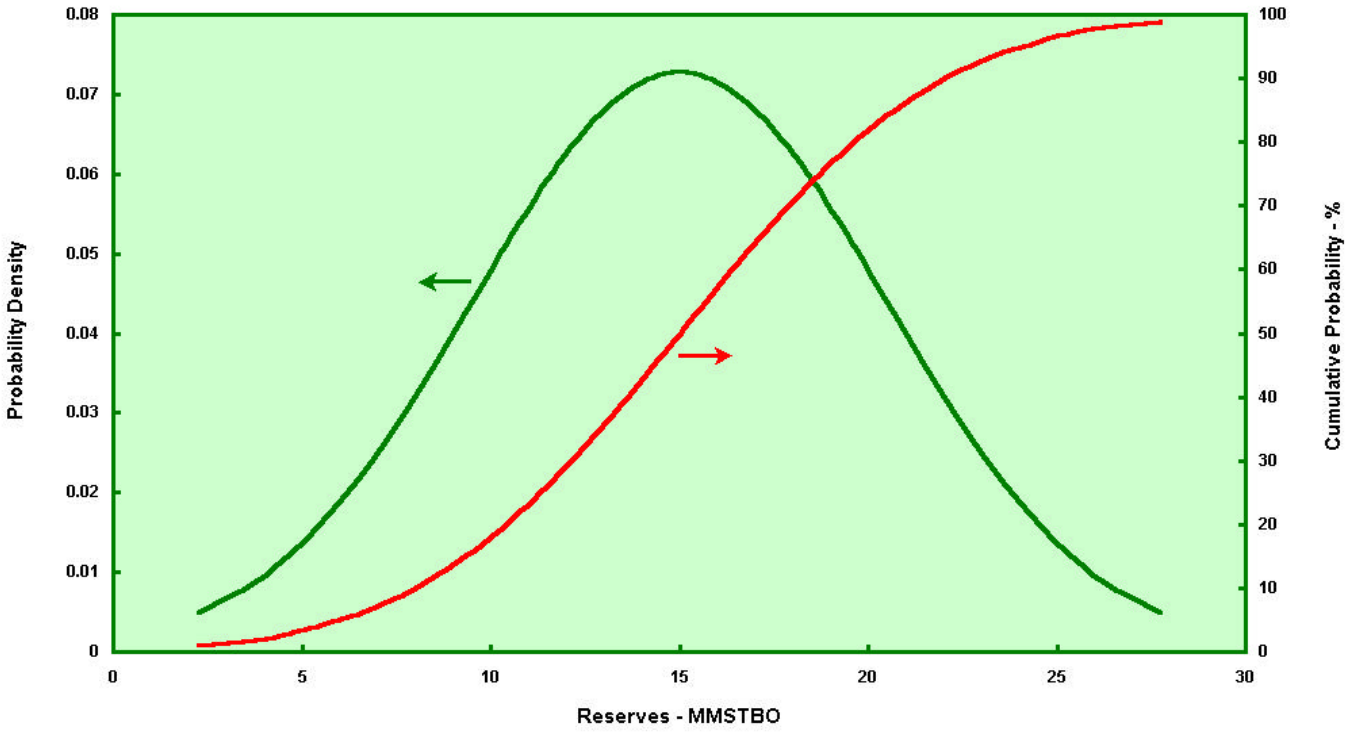
The Normal Distribution

Perhaps the most well known statistical distribution is the "bell shaped" normal distribution. The algebraic function, f_N, that yields the normal distribution is known as the probability density function. The "height" of the curve can be calculated by this equation:

$$f_N(x) = \frac{1}{s\sqrt{2p}} \times e^{\left\{-(x-m)^2/2s^2\right\}}$$

It is important to recognize that two parameters determine the distribution: the "mean" value of the function is represented by μ, and the variance represented by σ^2. Note that the square root of the variance is the standard deviation. As variance increases, the bell curve flattens resulting in more uncertainty (or a wider range of outcomes).

An example normal distribution is shown on the following page. This particular distribution describes the probability for a discovery volume with a "mean" of 15 million barrels. A complete discussion of the normal distribution can be found in any basic statistics textbook.



While the probability density function is useful in understanding the behavior of the distribution, Monte Carlo simulation utilizes the "S-shaped" cumulative probability curve. Each point on this curve is determined by dividing the area under the curve to the left of the "x" value of interest by the total area of the curve (which happens to be the square root of 2π). Entering our example distribution at a cumulative probability of say 20% yields a reserve value of 10.4 million barrels. This indicates that for this particular distribution, we have a 20% chance of discovering 10.4 million barrels or less. Put another way, this indicates that we have an 80% chance of discovering at least 10.4 million barrels.

The "area under the curve" is often estimated by integrating the function between limits. Unfortunately, *this function can't be analytically integrated*. The area under the curve can only be numerically estimated, for example, by breaking the area into numerous segments and using the trapezoidal rule to determine each segments area, which are subsequently summed. While this could be accomplished through custom functions written in basic, Excel comes to the rescue with two very useful built-in functions, NORMDIST and NORMINV. The function NORMDIST can be used to determine the cumulative probability (or probability density) for a given "x" value given a normal distribution with specified mean and standard deviation. (Remember that standard deviation is the square root of the variance.) NORMINV determines the inverse operation by returning the "x" value for a given cumulative probability and normal distribution with specified mean and standard deviation. Clearly, these Excel functions are useful, indeed!

Excel help files with syntax and an example for each of these two functions are as follows:

NORMDIST

See Also

Returns the normal cumulative distribution for the specified mean and standard deviation. This function has a very wide range of applications in statistics, including hypothesis testing.

Syntax

NORMDIST(x,mean,standard_dev,cumulative)

X is the value for which you want the distribution.

Mean is the arithmetic mean of the distribution.

Standard_dev is the standard deviation of the distribution.

Cumulative is a logical value that determines the form of the function. If cumulative is TRUE, NORMDIST returns the cumulative distribution function; if FALSE, it returns the probability mass function.

Remarks

- If mean or standard_dev is nonnumeric, NORMDIST returns the #VALUE! error value.
- If standard_dev ≤ 0, NORMDIST returns the #NUM! error value.
- If mean = 0 and standard_dev = 1, NORMDIST returns the standard normal distribution, NORMSDIST.
- The equation for the normal density function is:

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

Example

NORMDIST(42,40,1.5,TRUE) equals 0.908789

NORMINV

See Also

Returns the inverse of the normal cumulative distribution for the specified mean and standard deviation.

Syntax

NORMINV(probability,mean,standard_dev)

Probability is a probability corresponding to the normal distribution.

Mean is the arithmetic mean of the distribution.

Standard_dev is the standard deviation of the distribution.

Remarks

- If any argument is nonnumeric, NORMINV returns the #VALUE! error value.
- If probability < 0 or if probability > 1, NORMINV returns the #NUM! error value.
- If standard_dev ≤ 0, NORMINV returns the #NUM! error value.

NORMINV uses the standard normal distribution if mean = 0 and standard_dev = 1 (see NORMSINV).

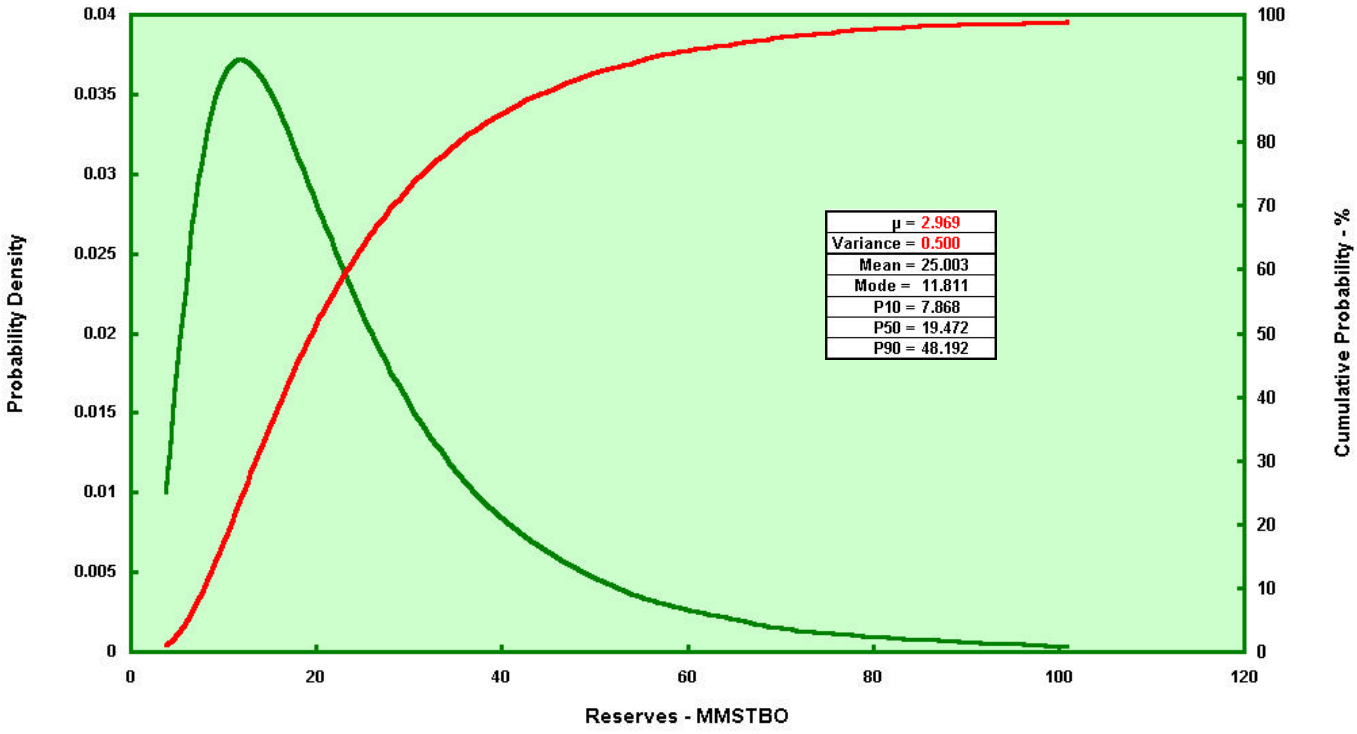
NORMINV uses an iterative technique for calculating the function. Given a probability value, NORMINV iterates until the result is accurate to within ± 3x10^-7. If NORMINV does not converge after 100 iterations, the function returns the #N/A error value.

Example

NORMINV(0.908789,40,1.5) equals 42

The Lognormal Distribution

Another well-known statistical distribution often used to describe quantities is the lognormal distribution. This type of distribution with the mode (or most likely value) skewed to the left is often observed in nature. The origin of this distribution comes from the "central limit theorem" which indicates that a log normal distribution always results when the observed quantity is the product of two or more independent distributions. An example of a lognormal distribution is illustrated below.



The equation for the probability density for the lognormal distribution is:

$$f_N(x) = \frac{1}{x s \sqrt{2p}} \times e^{\left\{-(\ln x - m)^2 / 2s^2\right\}}$$

While this equation appears very similar to the normal probability density function, the quantity m *is the mean of the natural logarithm (ln x) and* s^2 *is the variance of the natural logarithm*. The actual mean of the values can be calculated by:

$$M = e^{m + s^2/2}$$

The value of the mode can be determined by:

$$m = e^{m - s^2}$$

And finally, the value corresponding to a cumulative probability of 50% is known as the median. It can be calculated by the equation:

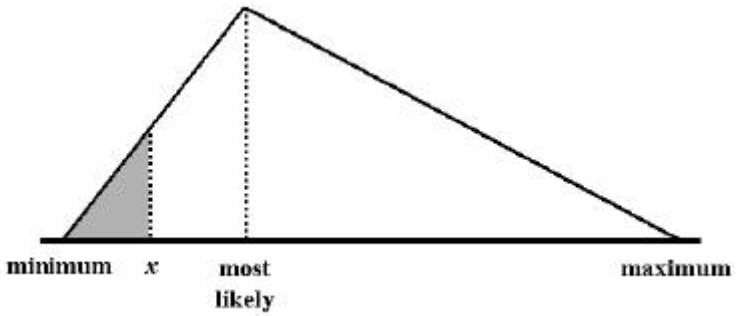
$$P_{50} = e^{m}$$

Note that only two of the parameters μ, variance, mean, mode, or median are required to define a unique lognormal distribution. It is also worth noting that in the symmetric normal distribution, only the mean was defined because the mean, mode, and median are all equal.

As in the case of the normal distribution, Excel provides useful built-in functions for the lognormal distribution. The function LOGNORMDIST can be used to determine the cumulative probability for a given "x" value given a specified μ and standard deviation. LOGINV determines the inverse operation by returning the "x" value for a given cumulative probability with specified μ and standard deviation. Remember, for the lognormal distribution, μ is the mean of the natural logarithm (ln x), σ^2 is the variance of the natural logarithm, and standard deviation is the square root of the variance. Excel help files for the lognormal functions are on the following page.

The Triangular Distribution

Another distribution (of arguable merit) that is often used in the oil and gas industry is the triangular distribution. In this distribution, minimum, maximum, and most likely values are specified. They are connected in a triangular fashion as illustrated below to create the probability density function.



The primary advantage of the triangular distribution is that the cumulative distribution can be easily expressed algebraically with a bit of geometry. Recall that the cumulative probability of the value "x" in the distribution above is the shaded area divided by the total area of the triangle.

An Excel function to calculate the value "x" for a given cumulative probability for a triangular distribution is included in the spreadsheet "Example Simulation.XLS". The syntax of the function is as follows:

Triangleinv(P, min, most, max)

P = cumulative probability (as a fraction between 0 and 1)
min = minimum value
most = most likely value
max = maximum value

An example of the use of this function will be discussed later.

LOGNORMDIST

See Also

Returns the cumulative lognormal distribution of x, where ln(x) is normally distributed with parameters mean and standard_dev. Use this function to analyze data that has been logarithmically transformed.

Syntax

LOGNORMDIST(x,**mean**,**standard_dev**)

X is the value at which to evaluate the function.

Mean is the mean of ln(x).

Standard_dev is the standard deviation of ln(x).

Remarks

- If any argument is nonnumeric, LOGNORMDIST returns the #VALUE! error value.
- If x ≤ 0 or if standard_dev ≤ 0, LOGNORMDIST returns the #NUM! error value.
- The equation for the lognormal cumulative distribution function is:

$$\text{LOGNORMDIST}(x,\mu,\sigma) = \text{NORMSDIST}\left(\frac{\ln(x)-\mu}{\sigma}\right)$$

Example

LOGNORMDIST(4,3.5,1.2) equals 0.039084

LOGINV

See Also

Returns the inverse of the lognormal cumulative distribution function of x, where ln(x) is normally distributed with parameters mean and standard_dev. If p = LOGNORMDIST(x,...) then LOGINV(p,...) = x.

Use the lognormal distribution to analyze logarithmically transformed data.

Syntax

LOGINV(**probability**,**mean**,**standard_dev**)

Probability is a probability associated with the lognormal distribution.

Mean is the mean of ln(x).

Standard_dev is the standard deviation of ln(x).

The inverse of the lognormal distribution function is:

$$\text{LOGINV}(p,\mu,\sigma) = e^{[\mu+\sigma\times(\text{NORMSINV}(p))]}$$

Remarks

- If any argument is nonnumeric, LOGINV returns the #VALUE! error value.
- If probability < 0 or probability > 1, LOGINV returns the #NUM! error value.
- If standard_dev <= 0, LOGINV returns the #NUM! error value.

Example

LOGINV(0.039084, 3.5, 1.2) equals 4.000014

Analysis of Data

Excel can be used to determine the mathematical description of data using either the normal, lognormal, or triangular distribution. An example of an analysis can be found in the spreadsheet "Example Analysis.XLS". This spreadsheet analyzes some porosity data from Canadian Fields presented by Ed Capen in the article "Dealing with Exploration Uncertainties" (The Business of Petroleum Exploration published by The American Association of Petroleum Geologists, Tulsa, Oklahoma, 1992, p. 39).

Statisticians often build probability plots by selecting appropriate "bins" and counting the occurrences in each bin (Excel actually can perform this operation somewhat automatically using the Data Analysis add-in). This type of analysis is only useful when large amounts of data are available. Data available for analysis in typical oil and gas exploration applications is often very limited and another method is employed. This method creates a *cumulative* probability distribution plot using the following steps:

1) The values are first arranged in ascending order.
2) The values are then numbered from 1 to the total number of samples, *n*.
3) The cumulative probability is then calculated by dividing the sample number by the *n+1*.
4) The cumulative probability data is then plotted on special probability paper for analysis.

While it may seem intuitive that the cumulative probability should be divided by the total number of samples instead of the total number of samples plus one, this approach allows for as much *probability space* above the highest value as below the lowest value.

A copy of a portion of the worksheet titled "Normal Analysis" within the spreadsheet "Example Analysis.XLS" is shown on the following page. The cumulative probability distribution is calculated for the 36 original porosity values using the procedure described above, and is the two columns highlighted with the red font.

The next step in the procedure requires plotting the cumulative probability on special cumulative probability graph paper for analysis. This paper is commercially available for both the normal and lognormal distributions. The probability scale is arranged so that a cumulative distribution will plot as a straight line if the data follows a normal distribution (or lognormal in the case of lognormal graph paper). Unfortunately, Excel does not allow probability as a scale option. In lieu of using paper and plotting by hand, the "Example Analysis.XLS" spreadsheet contains a basic function "XAxis" that provides an x-axis value between an absolute value of 0 ,which corresponds to a probability value of 0.1%, and a value of 2, which corresponds to a probability value of 99.9%. The 50% probability value logically has an absolute value of 1. Along with some clever data point labeling and the construction of gridlines (values on a hidden worksheet named "Gridlines"), the following cumulative probability plot was constructed.

Microsoft Excel - Example Analysis.XLS

File Edit View Insert Format Tools Data Window Help

Arial 10 B I U 100%

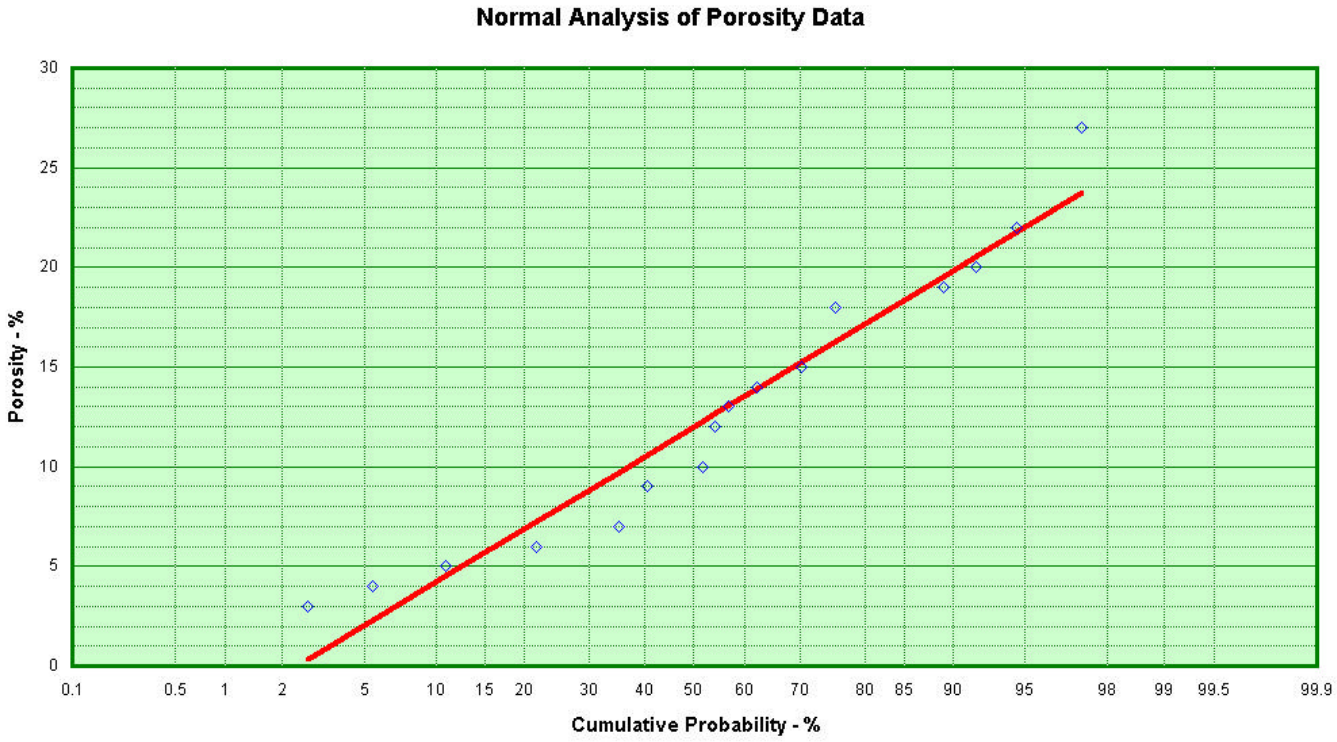
O4 =

Sample Average =	12.750

Mean = 12.052	Sum Sq = 0.682
Variance = 36.881	
P10 = 4.269	
P90 = 19.835	

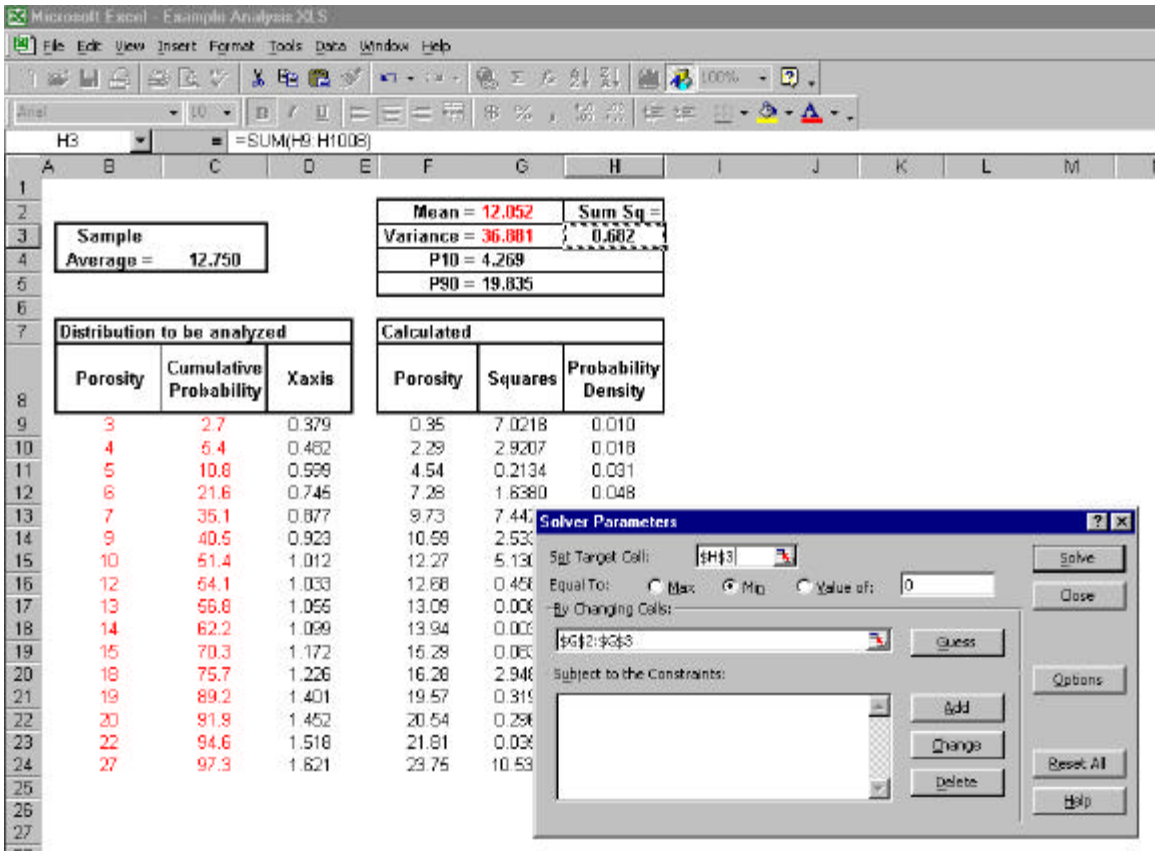
Distribution to be analyzed			Calculated		
Porosity	Cumulative Probability	Xaxis	Porosity	Squares	Probability Density
3	2.7	0.379	0.35	7.0218	0.010
4	5.4	0.482	2.29	2.9207	0.018
5	10.8	0.599	4.54	0.2134	0.031
6	21.6	0.745	7.28	1.6380	0.048
7	35.1	0.877	9.73	7.4429	0.061
9	40.5	0.923	10.59	2.5336	0.064
10	51.4	1.012	12.27	5.1302	0.066
12	54.1	1.033	12.68	0.4584	0.065
13	56.8	1.055	13.09	0.0085	0.065
14	62.2	1.099	13.94	0.0037	0.063
15	70.3	1.172	15.29	0.0836	0.057
18	75.7	1.226	16.28	2.9487	0.052
19	89.2	1.401	19.57	0.3199	0.031
20	91.9	1.452	20.54	0.2961	0.025
22	94.6	1.518	21.81	0.0351	0.018
27	97.3	1.621	23.75	10.5394	0.010

Normal Analysis | Normal Plot | Normal Density Plot | Lognormal Analysis | Lognormal Plot | Lognormal Density Plot

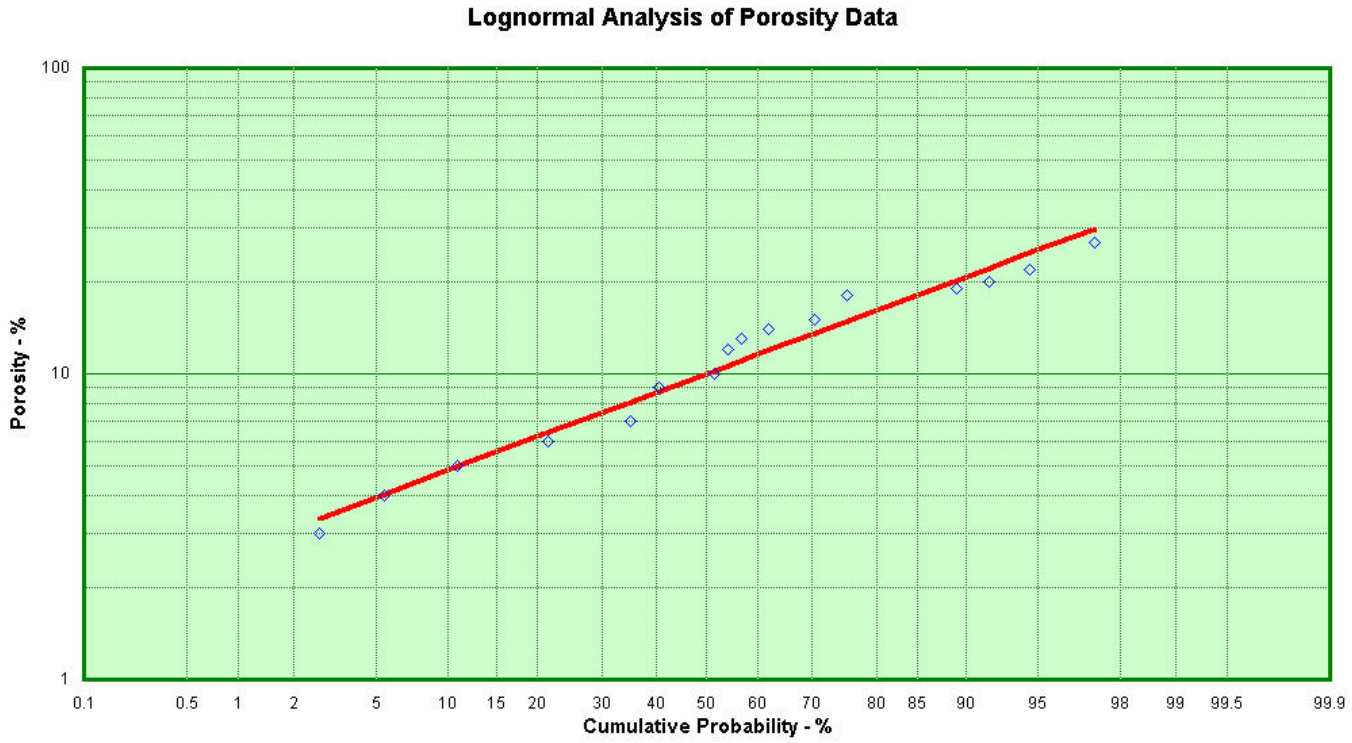


Care should be exercised when using the special probability scales used in the graphs of this spreadsheet. The "X" axis range must use the calculated value (using the XAxis function) instead of the actual cumulative probability value. Fonts and titles on the graph can be changed and the desired "Y" scale ranges can be changed. Note that once "Y" scale ranges are changed, any mouse movement repositions the gridlines and probability value labels. Plotted ranges can be changed, but the example range allows for 1,000 values, which should be adequate in most cases. Do not "reorder" the plotted ranges (the gridlines are the second sequence plotted) and do not rename the graph worksheets as these references will not adjust in the macros.

Note that a straight line is constructed through the discrete points on the previous cumulative normal probability plot. This line was calculated for the indicated mean and variance using the built-in Excel NORMINV function and the actual cumulative probability values. The powerful Excel add-in "Solver" can be used to determine a "least squares" fit of the data. As shown below, the column labeled "Squares" is the square of the difference between the actual and calculated porosity values. These quantities are then summed in the cell titled "Sum Sq =". Then the powerful Solver add-in is used to determine the values of mean and variance that yield the minimum sum of the squares. The usefulness of the Solver add-in goes well beyond statistics and is a valuable Excel tool useful for a wide variety of applications.



A similar analysis using the lognormal probability distribution is also included in the "Example Analysis.XLS" spreadsheet. Solver is once again used to determine a least squares best-fit line which is illustrated below. The primary difference is that instead of minimizing the square of the distances, the difference between the square of the distance between the natural logarithm of the actual and calculated values is minimized. This explains the additional columns on the worksheet "Lognormal Analysis". While some parameters that compare the quality of these regression analyses might be calculated, it is clear from visual inspection of the graph that the lognormal analysis most accurately describes this limited porosity data set.



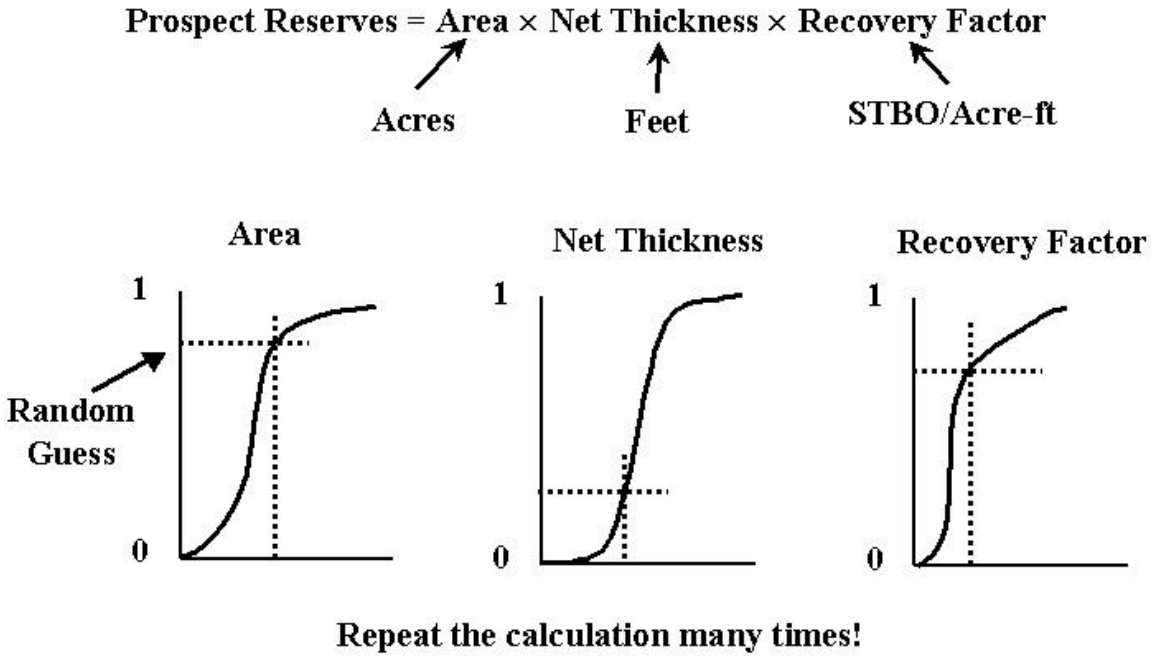
Add-ins in Excel must be initially installed by selecting "Tool" and then "Add-ins" from the drop down menu. Make sure that the Solver Add-in is checked. Excel may prompt for the initial installation diskettes. Once Solver is installed, it will be available in the future whenever Excel is started. For additional information on Excel Add-ins, consult the Excel help facility.

Monte Carlo Simulation

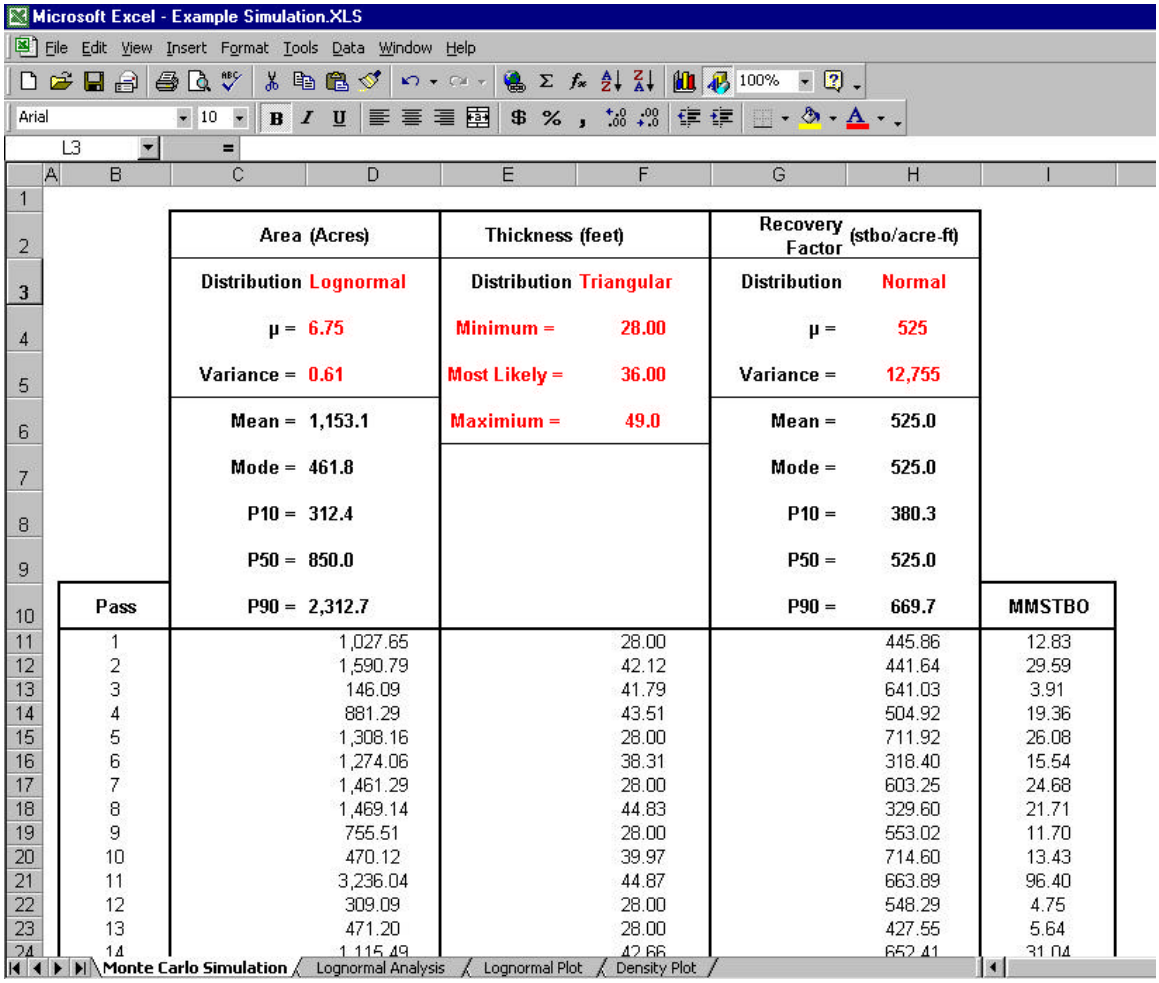
The procedure of Monte Carlo simulation is best understood with an example. Consider the simple estimation of prospect reserves that calculates ultimate recovery as the product of prospect size (acres), average thickness (feet), and recovery factor (STBO/acre-ft). The deterministic approach would simply multiply the "best estimate" for each of these quantities to obtain a single value.

But suppose that an *independent* cumulative probability distribution can be determined for each of the three quantities area, thickness, and recovery factor. These distributions could perhaps be determined using an involved regional study, or may be as simple as using what limited technical information is available and intuition to arrive at the best estimate of the probability distribution. The Monte Carlo method can make use of these distributions to arrive at a cumulative probability distribution for the estimate of prospect reserves. The determination of a probability distribution is superior to the single deterministic value in that insight is gained into the "upside", "downside", most likely outcome, and the mean level of reserves if a large number of similar prospects are drilled. *Prospect economics typically are based on mean reserve estimates, even though this reserve estimate exceeds the most likely outcome with the drilling of a single prospect!*

The Monte Carlo procedure is schematically illustrated below. A random number between zero and one is generated for each of the three parameters (easily accomplished using the Excel RAND() function). The cumulative probability distributions for each parameter are each then entered with the random number to determine a sample value which are then multiplied to yield a reserve estimate.



An individual calculation or iteration to estimate a prospect size in this fashion is known as a “pass” in Monte Carlo simulation parlance. By itself, the individual value is meaningless. However, when repeated numerous times, a cumulative distribution for the outcome emerges. An example using Monte Carlo simulation can be found in the Excel spreadsheet “Example Simulation.XLS”, a portion of which is shown below. In this example, prospect area uses a lognormal distribution, thickness uses a triangular distribution, and recovery factor uses a normal distribution. It is not suggested that these distributions are appropriate for each of these parameters, but are used to illustrate that any combination of distributions can be used in Monte Carlo simulation. Inspection of the formulas in this spreadsheet reveals that each parameter estimate is based upon a random number and changes with each recalculation (press F9 to recalculate the spreadsheet). The calculation is repeated 1000 times, which is typically adequate to sufficiently describe the resultant distribution considering the inaccurate nature of the input variables. While this sheet illustrates the simple operation of multiplying three values together, it is obvious that *the calculation can be as complicated as warranted!*



Microsoft Excel - Example Simulation.XLS

File Edit View Insert Format Tools Data Window Help

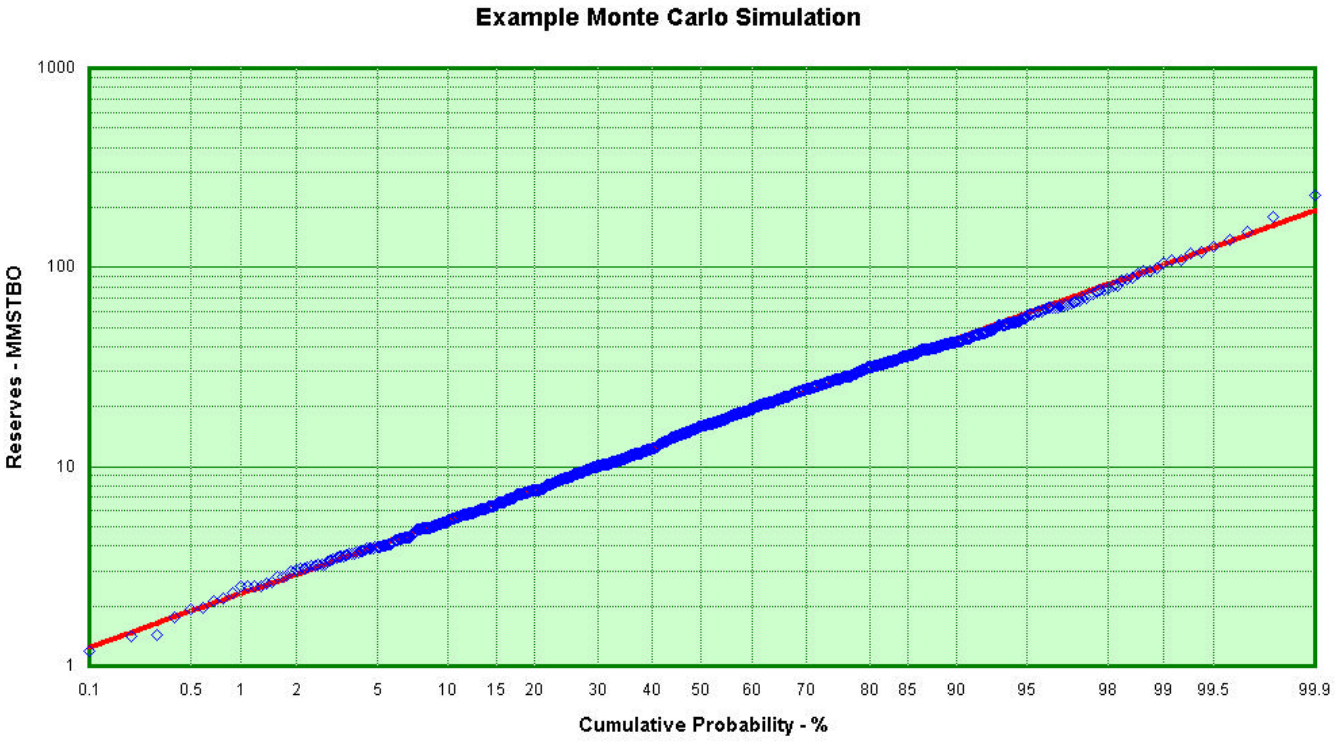
	Area (Acres)		Thickness (feet)		Recovery Factor (stbo/acre-ft)		
	Distribution	Lognormal	Distribution	Triangular	Distribution	Normal	
	μ =	6.75	Minimum =	28.00	μ =	525	
	Variance =	0.61	Most Likely =	36.00	Variance =	12,755	
	Mean =	1,153.1	Maximium =	49.0	Mean =	525.0	
	Mode =	461.8			Mode =	525.0	
	P10 =	312.4			P10 =	380.3	
	P50 =	850.0			P50 =	525.0	
Pass	P90 =	2,312.7			P90 =	669.7	MMSTBO
1		1,027.65		28.00		445.86	12.83
2		1,590.79		42.12		441.64	29.59
3		146.09		41.79		641.03	3.91
4		881.29		43.51		504.92	19.36
5		1,308.16		28.00		711.92	26.08
6		1,274.06		38.31		318.40	15.54
7		1,461.29		28.00		603.25	24.68
8		1,469.14		44.83		329.60	21.71
9		755.51		28.00		553.02	11.70
10		470.12		39.97		714.60	13.43
11		3,236.04		44.87		663.89	96.40
12		309.09		28.00		548.29	4.75
13		471.20		28.00		427.55	5.64
14		1,115.49		42.66		652.41	31.04

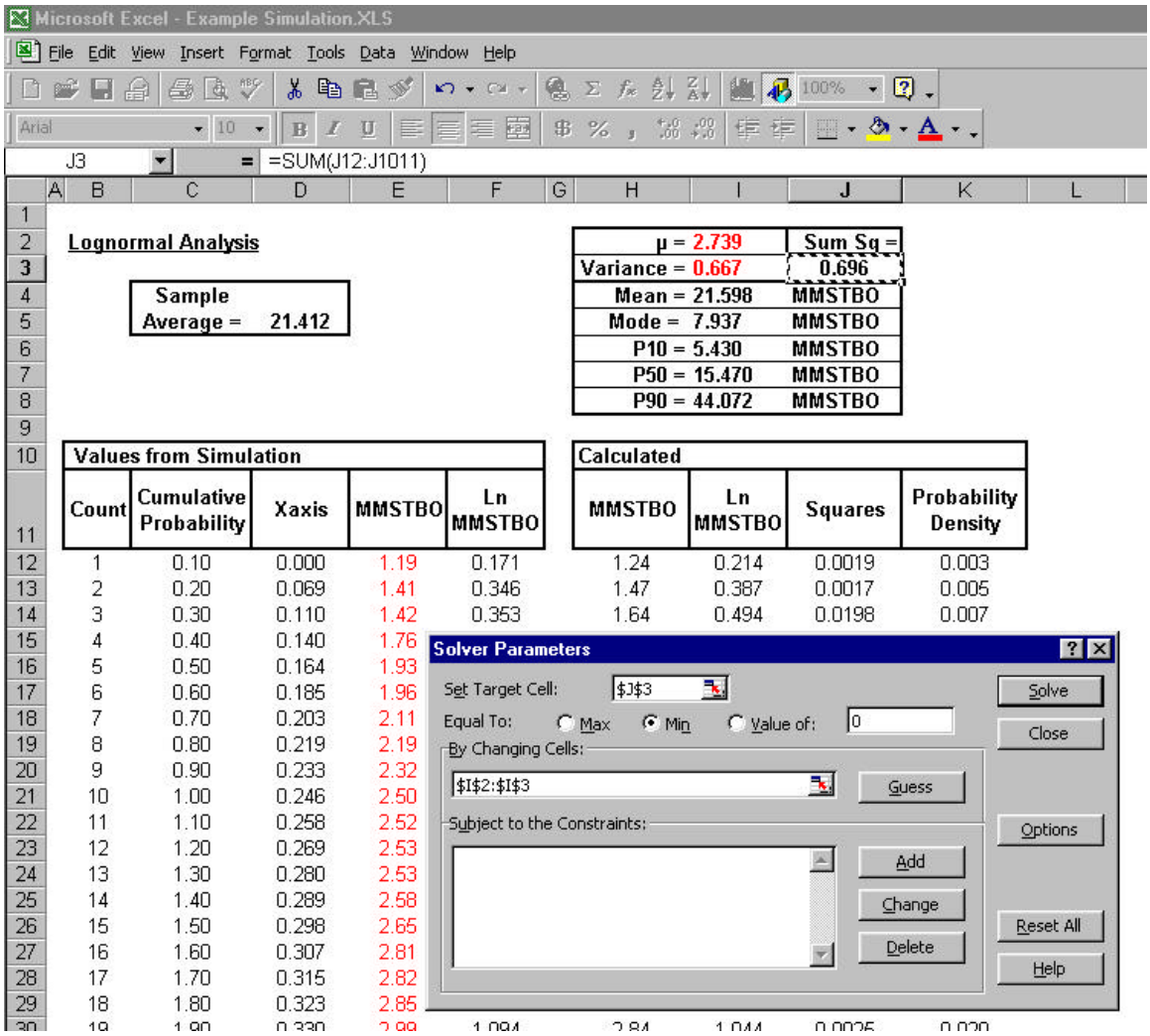
Monte Carlo Simulation | Lognormal Analysis | Lognormal Plot | Density Plot

Once the simulation is completed, the results must then be analyzed. This can be accomplished using a procedure similar to the previous discussion that described the procedure using the spreadsheet "Example Analysis.XLS". The analysis will be limited to a lognormal distribution (recall that the Central Limit Theorem indicates that the resultant distribution should follow a lognormal distribution). In an identical procedure, the 1000 passes are manipulated as follows:

1) The values are first arranged in ascending order.
2) The values are then numbered from 1 to the total number of samples, 1000.
3) The cumulative probability is then calculated by dividing the sample number by the total number of samples *plus one* (in this case, 1001).
4) The cumulative probability data is then plotted on special probability paper for analysis.

The values are first copied to a work area on the spreadsheet *as values* and not as formulas (use Paste Special, Values) before following this procedure. To streamline this operation, a keystroke macro (Ctrl+S) in the spreadsheet will effortlessly complete the entire procedure. The lognormal cumulative probability plot for this example simulation is shown below (select the "Lognormal Plot" worksheet. As previously discussed, the "Solver" add-in can be used to determine the best-fit least-squares line for the resultant function (shown on the following page). Be forewarned, however, that with 1000 points, the solver may take a few moments to complete. Progress can be monitored on the bottom left of the spreadsheet in an informational panel.





This analysis yields the statistical parameters desired for prospect analysis. These values should be expected to vary slightly with each simulation. An unacceptably large variation suggests that the number of passes should be increased. As previously mentioned, the mean reserve value is normally used for prospect economics and comparison to other prospects. In this case, the mean of the prospect based upon the best fit line is 21.6 million barrels. This compares favorably to the actual sample average of 21.4 million barrels. The upside and downside of the prospect can be assessed by the P10 value of 5.4 million barrels and the P90 value of 44.1 million barrels. (The P10 value suggests that there is only a 10% chance of discovering less than 5.4 million barrels. Put another way, there is a 90% chance of discovering at least 5.4 million barrels. For this reason, many companies chose to reverse the P10 and P90 values). The most likely outcome for this prospect (the mode) *is only 7.9 million barrels!* The use of the much greater mean value for justification of a prospect assumes that a large number of prospects will be drilled, with the occasional extremely large discovery.

μ = 2.739	Sum Sq =
Variance = 0.667	0.696
Mean = 21.598	MMSTBO
Mode = 7.937	MMSTBO
P10 = 5.430	MMSTBO
P50 = 15.470	MMSTBO
P90 = 44.072	MMSTBO