Statistical Analysis Tool

www.realoptionsvaluation.com

ROV Technical Papers Series: Volume 40

In This Issue

1. Learn how to determine the statistical properties of your data

*"What statistical tests can be performed using **Risk Simulator's** Statistical Analysis tool?"*

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

Procedure

- Open the example model (*Risk Simulator* | *Example Models* | *Statistical Analysis*), go to the *Data* worksheet, and select the data including the variable names (Figure 1).
- Click on *Risk Simulator* | *Tools* | *Statistical Analysis* (Figure 1).
- Check the data type, that is, whether the data selected are from a single variable or multiple variables arranged in rows. In our example, we assume that the data areas selected are from multiple variables. Click *OK* when finished.
- Choose the statistical tests you wish to perform. The suggestion (and by default) is to choose all the tests. Click *OK* when finished (Figure 2).

Spend some time going through the reports generated to get a better understanding of the statistical tests performed (sample reports are shown in Figures 3 through 6).

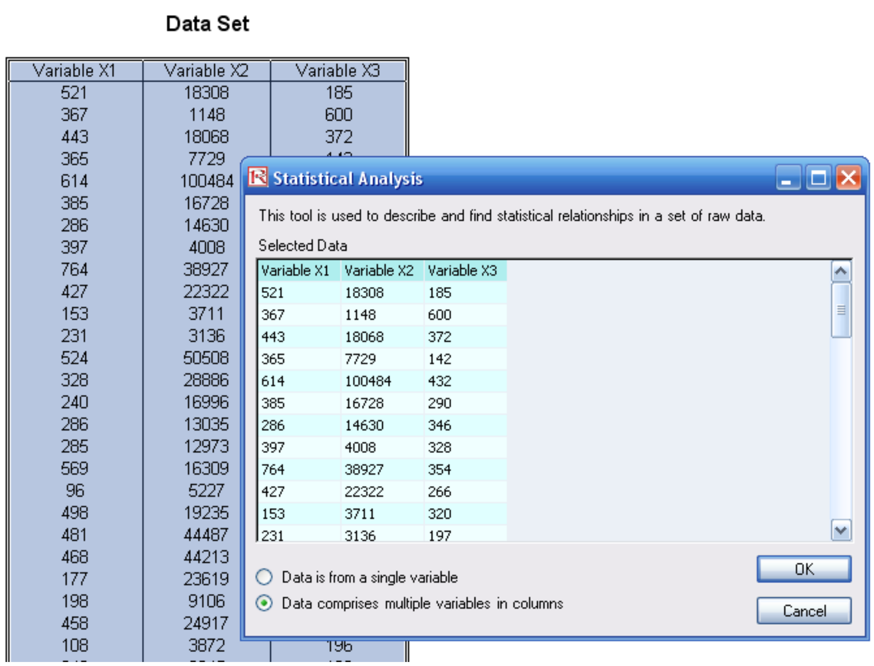


Figure 1. Running the Statistical Analysis Tool

Contact Us

Real Options Valuation, Inc.

4101F Dublin Blvd., Ste. 425, Dublin, California 94568 U.S.A.

admin@realoptionsvaluation.com
www.realoptionsvaluation.com
www.rovusa.com

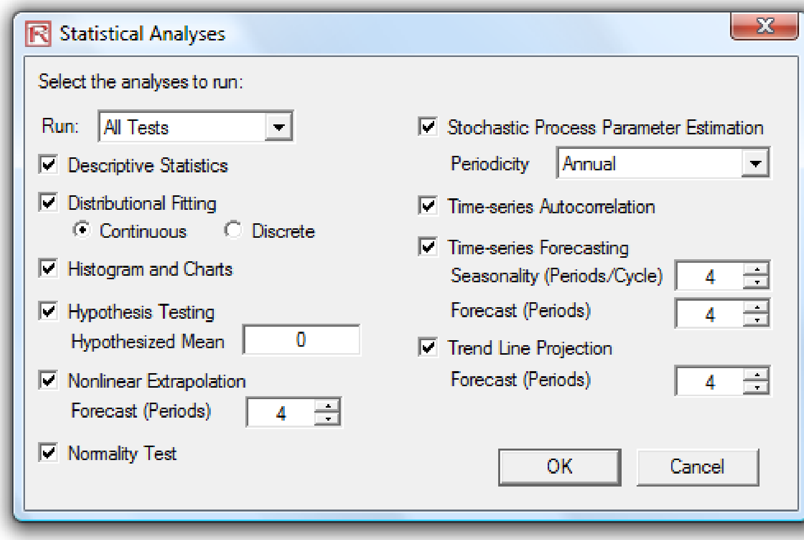


Figure 2. Statistical Tests

Descriptive Statistics

Analysis of Statistics

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

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

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

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

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

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

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

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

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

Summary Statistics

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

Figure 3. Sample Statistical Analysis Tool Report

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

Statistical Summary

Statistics from Dataset:

Observations	50
Sample Mean	331.92
Sample Standard Deviation	172.91

User Provided Statistics:

Hypothesized Mean	0.00

Calculated Statistics:

t-Statistic	13.5734
P-Value (right-tail)	**0.0000**
P-Value (left-tailed)	**1.0000**
P-Value (two-tailed)	**0.0000**

Null Hypothesis (Ho): μ = Hypothesized Mean
Alternate Hypothesis (Ha): μ <> Hypothesized Mean
Notes: "<>" denotes "greater than" for right-tail, "less than" for left-tail, or "not equal to" for two-tail hypothesis tests.

Hypothesis Testing Summary

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

Two-Tailed Hypothesis Test

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

Right-Tailed Hypothesis Test

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

Left-Tailed Hypothesis Test

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

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

Figure 4. Sample Statistical Analysis Tool Report (Hypothesis Testing of One Variable)

Test for Normality

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

Test Result

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

Null Hypothesis: The data is normally distributed.

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

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

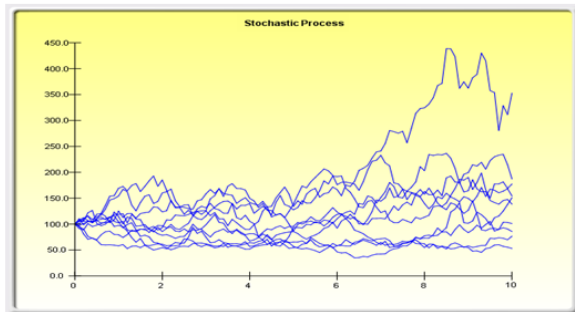
Figure 5. Sample Statistical Analysis Tool Report (Normality Test)

Stochastic Process - Parameter Estimations

Statistical Summary

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

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



Statistical Summary

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

(Annualized)

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

Probability of stochastic model fit: *4.63%*

Figure 6. Sample Statistical Analysis Tool Report (Forecasting)

Real Options Valuation Inc

