Automated Data Analysis

Short Examples Series

using

Risk Simulator

For more information please visit:
www.realoptionsvaluation.com
or contact us at:
admin@realoptionsvaluation.com

Analytics – Statistical Analysis

File Name: *Analytics – Statistical Analysis*
Location: *Modeling Toolkit | Analytics | Statistical Analysis*
Brief Description: *Applying the Statistical Analysis Tool to determine the key statistical characteristics of your data set, including linearity, nonlinearity, normality, distributional fit, distributional moments, forecastability, trends, and stochastic nature of the data*
Requirements: *Modeling Toolkit, Risk Simulator*

This model provides a sample data set on which to run the Statistical Analysis tool in order to determine the statistical properties of the data. The diagnostics run include checking the data for various statistical properties. This means all you have to do is to select your existing data set and run the statistical analysis tool, all but taking a few seconds, and out comes many detailed reports of the characteristics of your data set. This provides a very powerful automated analytical tool which simply provides a wealth of knowledge of your data set.

Procedure

To run the analysis, follow the instructions below:

1. Go to the *Data* worksheet and select the data including the variable names (cells **C5:E55**).

2. Click on **Risk Simulator | Tools | Statistical Analysis** (Figure 1).
3. Check the *data type*, 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.
4. *Choose the statistical tests* you wish performed. 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.

The analysis reports include the following statistical results:

- Descriptive Statistics: Arithmetic and geometric mean, trimmed mean (statistical outliers are excluded in computing its mean value), standard error and its corresponding statistical confidence intervals for the mean, median (the 50th percentile value), mode (most frequently occurring value), range (maximum less

minimum), standard deviation and variance of the sample and population, confidence interval for the population standard deviation, coefficient of variability (sample standard deviation divided by the mean), first and third quartiles (25th and 75th percentile value), skewness and excess kurtosis.

- Distributional Fit: Fitting the data to all 24 discrete and continuous distributions in Risk Simulator to determine which theoretical distribution best fits the raw data, and proving it with statistical goodness-of-fit results (Kolmogorov-Smirnov and Chi-Square tests' p-value results).
- Hypothesis Tests: Single variable one-tail and two-tail tests to see if the raw data is statistically similar or different from a hypothesized mean value.
- Nonlinear Extrapolation: Tests for nonlinear time-series properties of the raw data, to determine if the data can be fitted to a nonlinear curve.
- Normality Test: Fits the data to a normal distribution using a theoretical fitting hypothesis test to see if the data is statistically close to a normal distribution.
- Stochastic Calibration: Using the raw data, various stochastic processes are fitted (Brownian motion, jump-diffusion, mean-reversion, and random walk processes) and the levels of fit as well as the input assumptions are automatically determined.
- Autocorrelation and Partial Autocorrelation: The raw data is tested to see if it is correlated to itself in the past by applying some econometric estimations and tests of autocorrelation and partial autocorrelation coefficients.
- Time-Series Forecasting: Eight most-commonly used time-series decomposition models are applied to determine if the raw data set follows any trend and seasonality, and whether the time-series is predictable.
- Trend Analysis: A linear time-trend is tested to see if the data has any appreciable trend, using a linear regression approach.

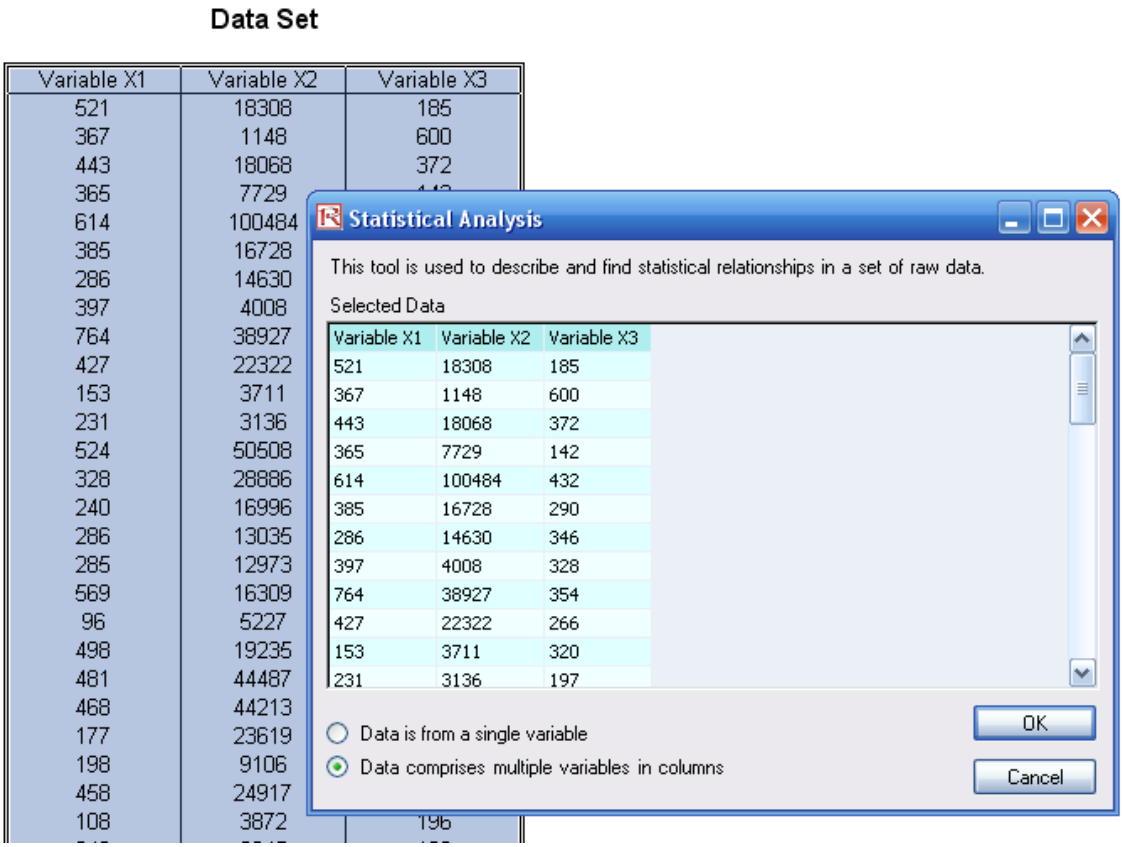


Figure 1: Running the statistical analysis tool

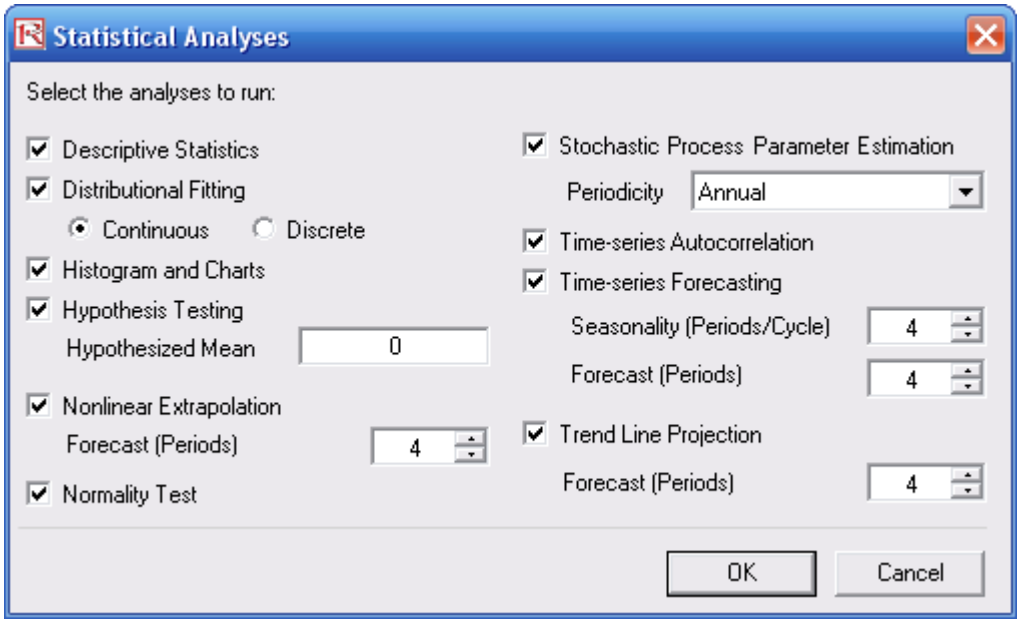


Figure 2: Statistical tests

Figure 3 shows a sample report generated by Risk Simulator that analyzes the statistical characteristics of your data set, providing all the requisite distributional moments and statistics to help you determine the specifics of your data, including the skewness and extreme events (kurtosis and outliers). The descriptions of these statistics are listed in the report for your review. Each variable will have its own set of reports.

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 report on descriptive statistics

Figure 4 shows the results of taking your existing data set and creating a distributional fit on 24 distributions. The best-fitting distribution (after Risk Simulator goes through multiple iterations of internal optimization routines and statistical analyses) is shown in the report, including the test statistics and requisite p-values, indicating the level of fit. For instance, Figure 4's example data set shows a 99.54% fit to a normal distribution with a mean of 319.58 and a standard deviation of 172.91. In addition, the actual statistics from your dataset are compared to the theoretical statistics of the fitted distribution, providing yet another layer of comparison. Using this methodology, you can take a large dataset and collapse it into a few simple distributional assumptions that can be simulated, thereby vastly reducing the complexity of your model or database while at the same time adding an added element of analytical prowess to your model by including risk analysis.

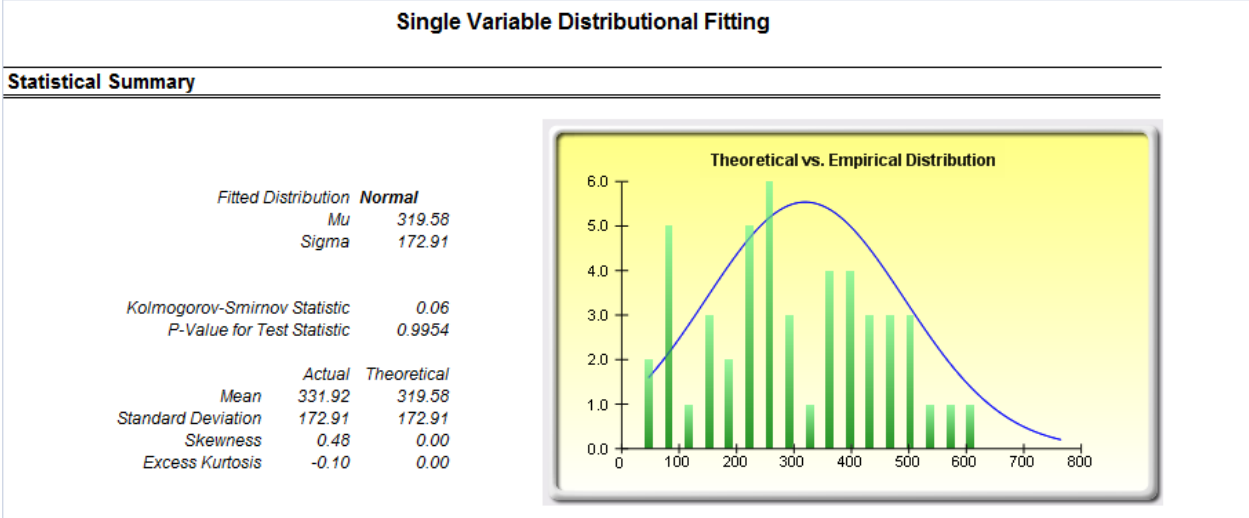


Figure 4: Sample report on distributional fitting

Sometimes, you might need to determine if the dataset's statistics are significantly different than a specific value. For instance, if the mean of your dataset is 0.15, is this statistically significantly different than say, zero? What about if the mean was 0.5 or 10.5? How far enough away does the mean have to be from this hypothesized population value to be deemed statistically significantly different? Figure 5 shows a sample report of such a hypothesis test.

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

Statistical Summary

Statistics from Dataset:		*Calculated Statistics:*	
Observations	*50*	*t-Statistic*	*13.5734*
Sample Mean	*331.92*	*P-Value (right-tail)*	***0.0000***
Sample Standard Deviation	*172.91*	*P-Value (left-tailed)*	***1.0000***
		P-Value (two-tailed)	***0.0000***
User Provided Statistics:			
		Null Hypothesis (Ho):	*μ = Hypothesized Mean*
Hypothesized Mean	*0.00*	*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 5: Sample report on theoretical hypothesis tests

Figure 6 shows the test for normality. In certain financial and business statistics, there is a heavy dependence on normality (e.g., asset distributions of option pricing models, normality of errors in a regression analysis, hypothesis tests using t-tests, z-tests, analysis of variance, and so forth). This theoretical test for normality is automatically computed as part of the statistical analysis tool.

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

Figure 6: Sample report on testing for normality

If your dataset is a time-series variable (i.e., data that has an element of time attached to them, such as interest rates, inflation rates, revenues, and so forth, that are time-dependent) then the Risk Simulator data analysis reports shown in Figures 7 to 11 will help in identifying the characteristics of this time-series behavior, including the identification of nonlinearity (Figure 7) versus linear trends (Figure 8), or a combination of both where there might be some trend and nonlinear seasonality effects (Figure 9). Sometimes, a time-series variable may exhibit relationship to the past (autocorrelation). The report shown in Figure 10 analyzes if these autocorrelations are significant and useful in future forecasts, that is, to see if the past can truly predict the future. Finally, Figure 11 illustrates the report on nonstationarity to test if the variable can or cannot be readily forecasted with conventional means (e.g., stock prices, interest rates,

foreign exchange rates are very difficult to forecast with conventional approaches and require stochastic process simulations) and identifies the best-fitting stochastic models such as a Brownian motion random walk, mean-reversion and jump-diffusion processes, and provides the estimated input parameters for these forecast processes.

Nonlinear Extrapolation

Statistical Summary

Extrapolation involves making statistical projections by using historical trends that are projected for a specified period of time into the future. It is only used for time-series forecasts. For cross-sectional or mixed panel data (time-series with cross-sectional data), multivariate regression is more appropriate. This methodology is useful when major changes are not expected, that is, causal factors are expected to remain constant or when the causal factors of a situation are not clearly understood. It also helps discourage the introduction of personal biases into the process. Extrapolation is fairly reliable, relatively simple, and inexpensive. However, extrapolation, which assumes that recent and historical trends will continue, produces large forecast errors if discontinuities occur within the projected time period. That is, pure extrapolation of time series assumes that all we need to know is contained in the historical values of the series that is being forecasted. If we assume that past behavior is a good predictor of future behavior, extrapolation is appealing. This makes it a useful approach when all that is needed are many short-term forecasts.

This methodology estimates the f(x) function for any arbitrary x value, by interpolating a smooth nonlinear curve through all the x values, and using this smooth curve, extrapolates future x values beyond the historical data set. The methodology employs either the polynomial functional form or the rational functional form (a ratio of two polynomials). Typically, a polynomial functional form is sufficient for well-behaved data, however, rational functional forms are sometimes more accurate (especially with polar functions, i.e., functions with denominators approaching zero).

Period	Actual	Forecast Fit	Extrapolation Error
1	47.00		
2	68.00	521.00	
3	87.00	283.27	
4	96.00	454.02	
5	102.00	320.15	
6	108.00	539.49	
7	114.00	344.12	
8	127.00	164.52	
9	153.00	339.71	
10	177.00	849.31	
11	186.00	373.33	
12	188.00	-232.05	
13	198.00	233.26	
14	222.00	692.53	
15	231.00	296.82	
16	240.00	168.77	
17	246.00	269.54	
18	251.00	282.63	
19	265.00	206.08	
20	280.00	127.22	
21	285.00	264.80	
22	286.00	480.49	
23	286.00	456.24	
24	291.00	542.93	
25	303.00	196.88	
26	311.00	482.25	
27	312.00	135.43	

Error Measurements	
RMSE	1093.8463
MSE	1196499.7684
MAD	440.2718
MAPE	269.18%
Theil's U	9.4724

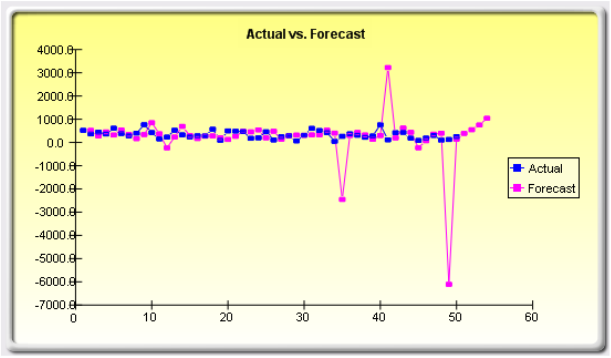


Figure 7: Sample report on nonlinear extrapolation forecast (nonlinear trend detection)

Linear Trend Line Projection

Regression Statistics

R-Squared (Coefficient of Determination)	*0.1193*
Adjusted R-Squared	*0.1009*
Multiple R (Multiple Correlation Coefficient)	*0.3454*

The R-Squared or Coefficient of Determination indicates the percent variation in the data set that can be explained and accounted for by the linear trend line alone, whereas the Adjusted R-Squared takes into account the limited data set and exogenous variables and adjusts this R-Squared value to a more accurate view of the explanatory power of the trend line in isolation.

The Multiple Correlation Coefficient (Multiple R) measures the correlation between the actual data and the fitted forecast of using a trend line. This is also the square root of the Coefficient of Determination (R-Squared).

Linear Trend Line Analysis Results

	Intercept	*Trend*
Coefficients	*436.3910*	*-4.0969*
Standard Error	*47.0778*	*1.6067*
t-Statistic	*9.2696*	*-2.5498*
p-Value	*0.0000*	***0.0140***
Lower 5%	*341.7347*	*-7.3275*
Upper 95%	*531.0473*	*-0.8663*

The Trend Line coefficients provide the estimated intercept and trend. The Standard Error measures how accurate the predicted Coefficients are, and the t-Statistics are the ratios of each predicted Coefficient to its Standard Error.

The t-Statistic is used in hypothesis testing, where we set the null hypothesis (Ho) such that the real mean of the Coefficient = 0, and the alternate hypothesis (Ha) such that the real mean of the Coefficient is not equal to 0. A t-test is is performed and the calculated t-Statistic is compared to the critical values at the relevant Degrees of Freedom for Residual. The t-test is very important as it calculates if the trend line is statistically significant.

The Linear Trend Line is statistically significant and correct if its calculated t-Statistic exceeds the Critical t-Statistic at the relevant degrees of freedom (df). The three main confidence levels used to test for significance are 99%, 95% and 90%. If a Coefficient's t-Statistic exceeds the Critical level, it is considered statistically significant. Alternatively, the p-Value calculates each t-Statistic's probability of occurrence, which means that the smaller the p-Value, the more significant the Coefficient. The usual significant levels for the p-Value are 0.01, 0.05, and 0.10, corresponding to the 99%, 95%, and 90% confidence levels.

The Coefficients with their p-Values highlighted in blue indicate that they are statistically significant at the 90% confidence or 0.10 alpha level, while those highlighted in red indicate that they are not statistically significant at any other alpha levels.

Forecasting

Period	Actual (Y)	Forecast (F)	Error (E)
1	521	432.2941	88.7059
2	367	428.1972	(61.1972)
3	443	424.1003	18.8997
4	365	420.0034	(55.0034)
5	614	415.9065	198.0935
6	385	411.8096	(26.8096)
7	286	407.7127	(121.7127)
8	397	403.6158	(6.6158)
9	764	399.5189	364.4811
10	427	395.4220	31.5780
11	153	391.3251	(238.3251)
12	231	387.2282	(156.2282)
13	524	383.1313	140.8687
14	328	379.0344	(51.0344)
15	240	374.9375	(134.9375)
16	286	370.8406	(84.8406)
17	285	366.7437	(81.7437)
18	569	362.6468	206.3532
19	96	358.5499	(262.5499)
20	498	354.4530	143.5470
21	481	350.3561	130.6439

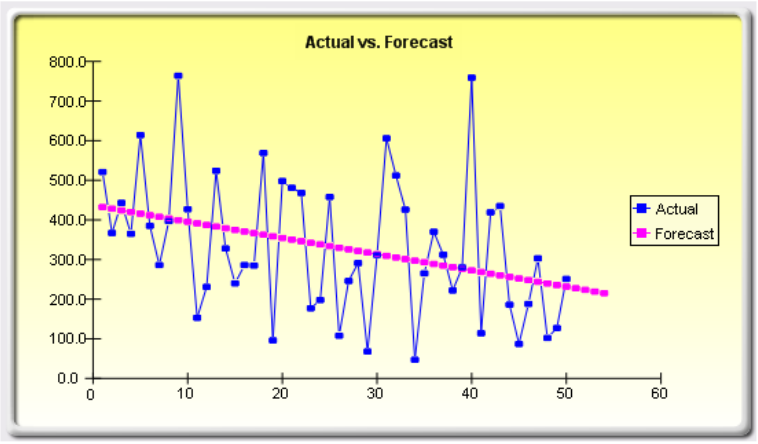


Figure 8: Sample report on trend line forecasts (linear trend detection)

Time-Series Forecasting

Time-Series Analysis Summary

Time-series forecasting is used to forecast the future based on historical data, through the decompositon of the historical data into the baseline

The best-fitting test for the moving average forecast uses the root mean squared errors (RMSE). The RMSE calculates the square root of the average squared deviations of the fitted values versus the actual data points.

Mean Squared Error (MSE) is an absolute error measure that squares the errors (the difference between the actual historical data and the forecast-fitted data predicted by the model) to keep the positive and negative errors from canceling each other out. This measure also tends to exaggerate large errors by weighting the large errors more heavily than smaller errors by squaring them, which can help when comparing different time-series models. Root Mean Square Error (RMSE) is the square root of MSE and is the most popular error measure, also known as the quadratic loss function. RMSE can be defined as the average of the absolute values of the forecast errors and is highly appropriate when the cost of the forecast errors is proportional to the absolute size of the forecast error. The RMSE is used as the selection criteria for the best-fitting time-series model.

Mean Absolute Percentage Error (MAPE) is a relative error statistic measured as an average percent error of the historical data points and is most appropriate when the cost of the forecast error is more closely related to the percentage error than the numerical size of the error. Finally, an associated measure is the Theil's U statistic, which measures the naivety of the model's forecast. That is, if the Theil's U statistic is less than 1.0, then the forecast method used provides an estimate that is statistically better than guessing.

The analysis was run with periodicity = 12

Period	Actual	Forecast Fit
1	47.00	
2	68.00	
3	87.00	
4	96.00	
5	102.00	
6	108.00	
7	114.00	
8	127.00	
9	153.00	
10	177.00	
11	186.00	
12	188.00	
13	198.00	412.75
14	222.00	413.00
15	231.00	409.75
16	240.00	392.83
17	246.00	386.25
18	251.00	358.83
19	265.00	374.17
20	280.00	358.33
21	285.00	366.75
22	286.00	343.17
23	286.00	346.58
24	291.00	348.58
25	303.00	345.83
26	311.00	340.33
27	312.00	322.00
28	328.00	322.50
29	365.00	322.92
30	367.00	304.83

Error Measurements	
RMSE	169.4315
MSE	28707.0345
MAD	138.4189
MAPE	85.13%
Theil's U	0.4781

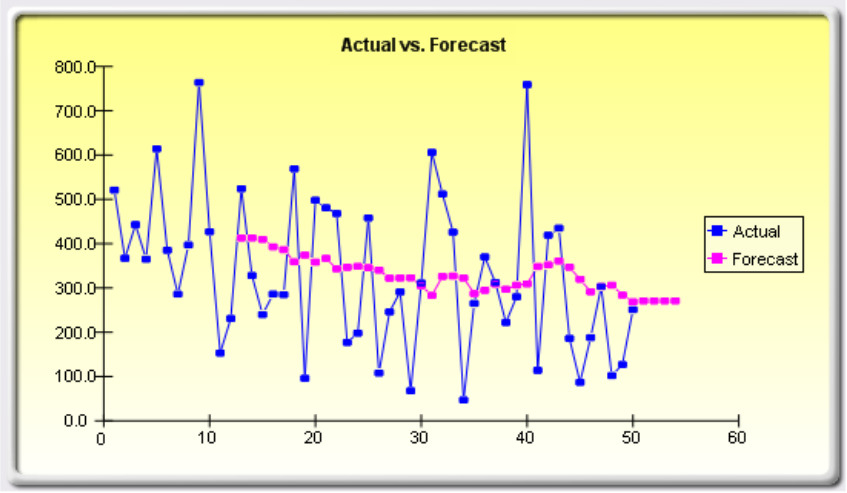


Figure 9: Sample report on time-series forecasting (seasonality and trend detection)

Autocorrelation

Autocorrelation

If autocorrelation *AC(1)* is nonzero, it means that the series is first order serially correlated. If *AC(k)* dies off more or less geometrically with increasing lag , it implies that the series follows a low-order autoregressive process. If *AC(k)* drops to zero after a small number of lags, it implies that the series follows a low-order moving-average process. Partial Autocorrelation *PAC(k)* measures the correlation of values that are *k* periods apart after removing the correlation from the intervening lags. If the pattern of autocorrelation can be captured by an autoregression of order less than *k*, then the partial autocorrelation at lag k will be close to zero. Ljung-Box Q-statistics and their p-values at lag *k* has the null hypothesis that there is no autocorrelation up to order *k*. The dotted lines in the plots of the autocorrelations are the approximate two standard error bounds. If the autocorrelation is within these bounds, it is not significantly different from zero at the 5% significance level.

Time Lag	AC	PAC	Lower Bound	Upper Bound	Q-Stat	Prob
1	0.0580	0.0580	-0.2828	0.2828	0.1752	0.6755
2	-0.1213	-0.1251	-0.2828	0.2828	0.9574	0.6196
3	0.0590	0.0756	-0.2828	0.2828	1.1464	0.7659
4	0.2423	0.2232	-0.2828	0.2828	4.4070	0.3537
5	0.0067	-0.0078	-0.2828	0.2828	4.4095	0.4921
6	-0.2654	-0.2345	-0.2828	0.2828	8.5034	0.2035
7	0.0814	0.0939	-0.2828	0.2828	8.8978	0.2601
8	0.0634	-0.0442	-0.2828	0.2828	9.1427	0.3304
9	0.0204	0.0673	-0.2828	0.2828	9.1688	0.4218
10	-0.0190	0.0865	-0.2828	0.2828	9.1920	0.5140
11	0.1035	0.0790	-0.2828	0.2828	9.8960	0.5398
12	0.1658	0.0978	-0.2828	0.2828	11.7535	0.4657
13	-0.0524	-0.0430	-0.2828	0.2828	11.9440	0.5322
14	-0.2050	-0.2523	-0.2828	0.2828	14.9439	0.3820
15	0.1782	0.2089	-0.2828	0.2828	17.2766	0.3026
16	-0.1022	-0.2591	-0.2828	0.2828	18.0670	0.3200
17	-0.0861	0.0808	-0.2828	0.2828	18.6463	0.3492
18	0.0418	0.1987	-0.2828	0.2828	18.7870	0.4050
19	0.0869	-0.0821	-0.2828	0.2828	19.4161	0.4304
20	-0.0091	-0.0269	-0.2828	0.2828	19.4233	0.4945

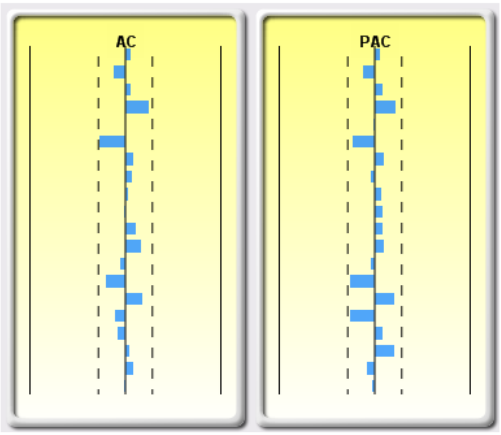


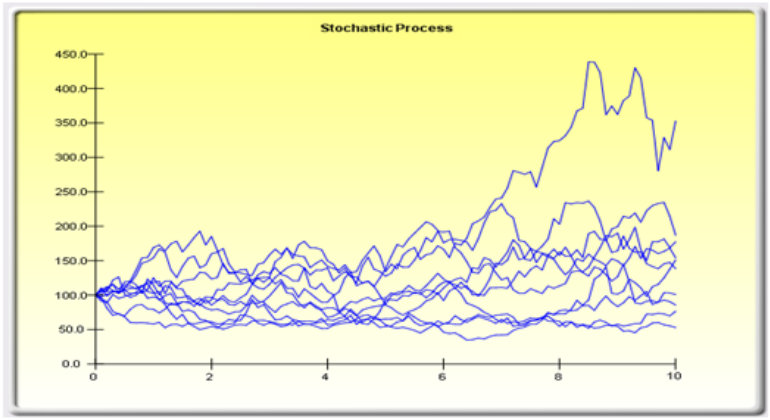
Figure 10: Sample report on autocorrelation (past relationship and correlation detection)

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	*-1.48%*	*Reversion Rate*	*283.89%*	*Jump Rate*	*20.41%*
Volatility	*88.84%*	*Long-Term Value*	*327.72*	*Jump Size*	*237.89*

Probability of stochastic model fit: *46.48%*

Figure 11: Sample report on stochastic parameter calibration (nonstationarity detection)