Monte Carlo Simulation (ARIMA Time Series Models)

statgraphics 18 centurion

Revised: 10/11/2017

Summary

This procedure generates random samples from ARIMA time series models. The general form of an ARIMA model is most easily expressed in terms of the backwards operator *B*, which operates on the time index of a data value such that $B^j Y_t = Y_{t-j}$. Using this operator, the model takes the form

$$\left(1-B-B^2-\ldots-B^p\right)\left(1-B^s-B^{2s}-\ldots-B^{Ps}\right)(1-B)^d(1-B^s)^D Z_t$$

$$=\left(1-B-B^2-\ldots-B^q\right)\left(1-B^s-B^{2s}-\ldots-B^{Qs}\right)a_t \qquad (1)$$

where

$$Z_t = Y_t - \mu \qquad (2)$$

© 2017 by Statgraphics Technologies, Inc

and a_t is a random error or shock to the system at time *t*, usually assumed to be random observations from a normal distribution with mean 0 and standard deviation σ. For a stationary series, μ represents the process mean. Otherwise, it is related to the slope of the forecast function. μ is sometimes assumed to equal 0.

The above model is often referred to as an ARIMA(*p*,*d*,*q*)x(*P*,*D*,*Q*)*s* model. It consists of several terms:

1. A nonseasonal autoregressive term of order *p*.
2. Nonseasonal differencing of order *d*.
3. A nonseasonal moving average term of order *q*.
4. A seasonal autoregressive term of order *P*
5. Seasonal differencing of order *D*.
6. A seasonal moving average term of order *Q*.

While the general model looks formidable, the most commonly used models are relatively simple special cases. These include:

<u>AR(1) – autoregressive of order 1</u>
The observation at time *t* is expressed as a mean plus a multiple of the deviation from the mean at the previous time period plus a random shock:

$$Y_t = \mu + \phi_1 (Y_{t-1} - \mu) + a_t \tag{3}$$

<u>AR(2) – autoregressive of order 2</u>
The observation at time *t* is expressed as a mean plus multiples of the deviations from the mean at the 2 previous time periods plus a random shock:

$$Y_t = \mu + \phi_1 (Y_{t-1} - \mu) + \phi_2 (Y_{t-2} - \mu) + a_t \tag{4}$$

<u>MA(1) – moving average of order 1</u>
The observation at time *t* is expressed as a mean plus a random shock at the current time period plus a multiple of the random shock at the previous time period:

$$Y_t = \mu + a_t - \theta_1 a_{t-1} \tag{5}$$

<u>MA(2) – moving average of order 2</u>
The observation at time *t* is expressed as a mean plus a random shock at the current time period plus multiples of the random shocks at the 2 previous time periods:

$$Y_t = \mu + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} \tag{6}$$

ARMA(1,1) – mixed model with 2 first order terms
The observation at time *t* is expressed as a mean plus a multiple of the deviation from the mean at the previous time period plus a random shock at the current time period plus a multiple of the random shock at the previous time period:

$$Y_t = \mu + \phi_1 (Y_{t-1} - \mu) + a_t - \theta_1 a_{t-1} \quad (7)$$

ARIMA(0,1,1) – moving average of order 1 applied to the first differences
The difference between the current period and the previous period is expressed as a random shock at the current time period plus a multiple of the random shock at the previous time period:

$$Y_t - Y_{t-1} = a_t - \theta_1 a_{t-1} \quad (8)$$

It can be shown that this model is equivalent to the *Simple Exponential Smoothing* model.

ARIMA(0,2,2) – moving average of order 2 applied to the second differences
The difference of the differences is expressed as a random shock at the current time period plus multiples of the random shocks at the 2 previous time periods:

$$(Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} \quad (9)$$

This model is equivalent to the *Holt's Linear Exponential Smoothing* model.

ARIMA(0,1,1)x(0,1,1)s – seasonal and nonseasonal MA terms of order 1
The observation at time *t* is expressed as a combination of the observation one season ago plus the difference between the observation last period and its counterpart one season ago plus multiple of the shocks to hit the system this period, last period, and two periods one season ago:

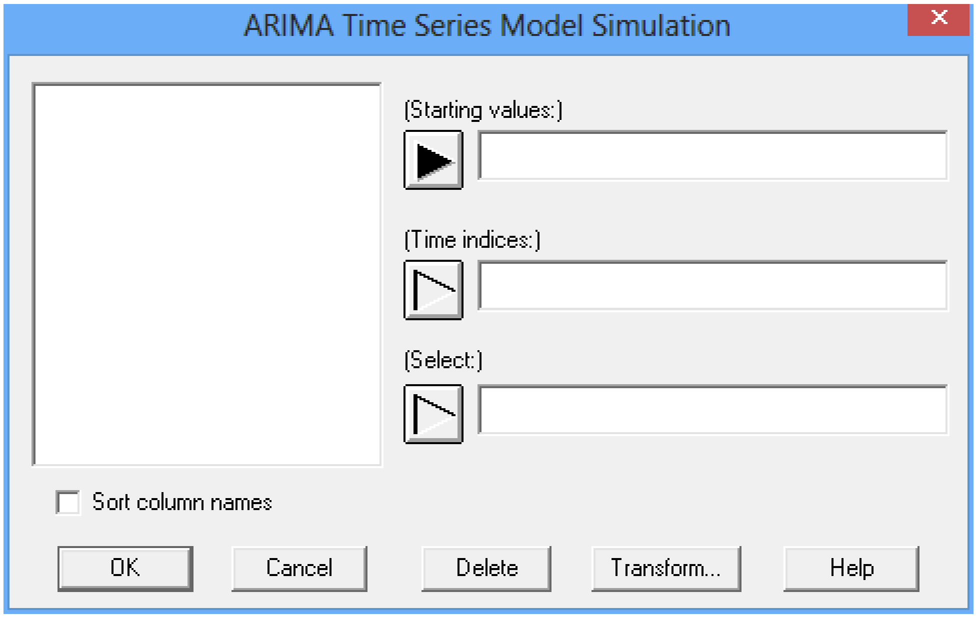
$$Y_t = Y_{t-s} + Y_{t-1} - Y_{t-s-1} + a_t - \theta_1 a_{t-1} - \Theta_1 a_{t-s} + \theta_1 \Theta_1 a_{t-s-1} \quad (10)$$

Many economic time series with a seasonal component can be well represented by this model.

Sample StatFolio: *monte4.sgp*

Data Input

The initial data input dialog box allows the analyst to enter a historical time series that will be used to set the starting values for the simulation:



Starting values: optional data to be used to set the starting values. The simulated data is assumed to begin immediately after the end of this data.

Time indices: optional values indicating the time at which each of the starting values was recorded. If supplied, these values will be used to scale the plot of the simulated data.

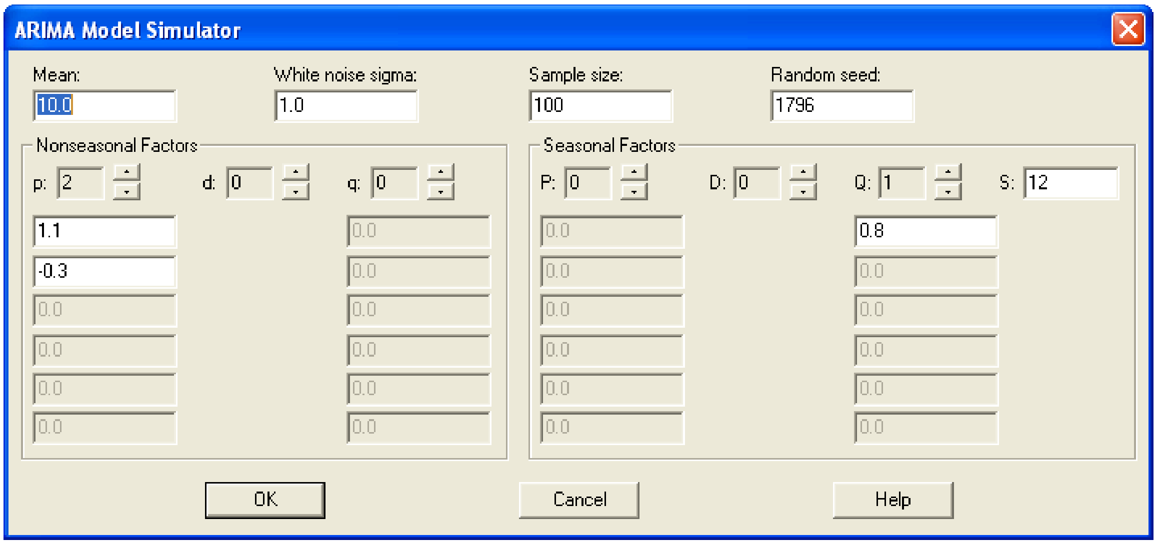
Select: optional subset selection.

If no starting values are supplied, the simulation will generate random starting values by simulating twice as much data as requested and discarding the first half.

© 2017 by Statgraphics Technologies, Inc

Analysis Options

The Analysis Options dialog box is used to specify the model from which the desired time series should be generated. For example, the dialog box below requests data from a (2,0,0)x(0,0,1)12 model:



- **Mean:** the value of the mean μ.
- **White noise sigma**: the standard deviation of the random shocks σ.
- **Sample size:** *n*, the length of the time series to be generated.
- **Random seed:** the seed for the random number generator. The initial default value is set based on the time of day. If you use the same seed more than once, you will get the same results.
- **Nonseasonal factors:** the order of the nonseasonal AR factor (*p*), the order of nonseasonal differencing (*d*), and the order of the nonseasonal MA factor *(q)*. The values of the AR and MA parameters are entered in the corresponding edit fields.
- **Seasonal factors:** the order of the seasonal AR factor (*P*), the order of seasonal differencing (*D*), the order of the seasonal MA factor *(Q)*, and the length of seasonality (*s*). The values of the AR and MA parameters are entered in the corresponding edit fields.

When the *OK* button is pressed, a random time series is generated from the specified model. To initialize the series, all values of Y_t for $t < 1$ are set equal to the mean, while all values for a_t for $t < 1$ are set equal to 0. A total of *2n* observations are generated, but only the last *n* are retained.

Analysis Summary

The *Analysis Summary* displays the requested model:

ARIMA Model Simulation
Sample size: 100
Seed for random number generator: 2993

Mean: 10.0
Sigma: 1.0

Nonseasonal Factors

	Order	*Parameters*
AutoRegressive	p=2	1.1,-0.3
Differencing	d=0	
Moving Average	q=0	

Seasonal Factors (S=12)

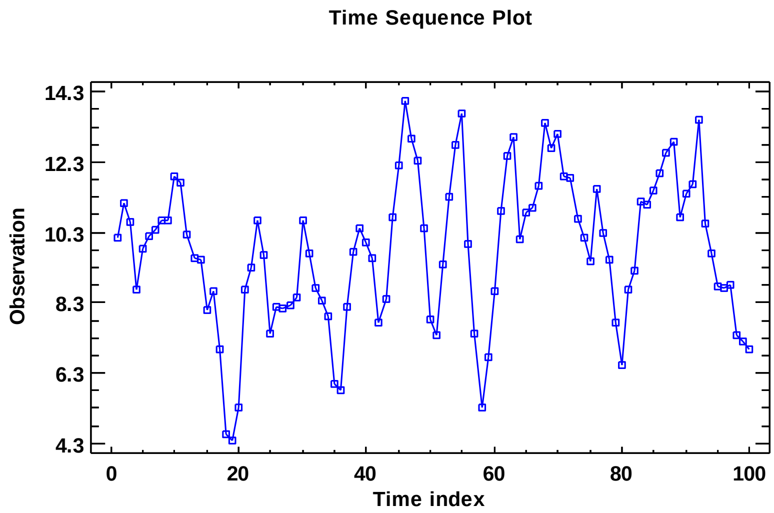
	Order	*Parameters*
AutoRegressive	P=0	
Differencing	D=0	
Moving Average	Q=1	0.8

If you wish to generate the same time series again, record the seed of the random number generator and use it the next time you generate the series. Otherwise, each time a series is generated, it will be different.

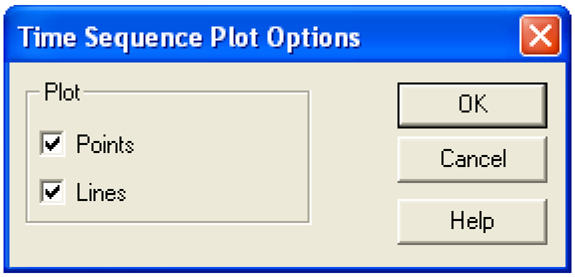
© 2017 by Statgraphics Technologies, Inc

Time Sequence Plot

The *Time Sequence Plot* displays the generated time series data in sequential order:



Pane Options



- **Points**: plot point symbols at the location of each observation.
- **Lines**: connect the observations with a line.

Autocorrelations

An important tool in modeling time series data is the autocorrelation function. The autocorrelation at lag *k* measures the strength of the correlation between observations *k* time periods apart. The sample lag *k* autocorrelation is calculated from

$$r_k = \frac{\sum_{t=1}^{n-k} (y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{n} (y_t - \bar{y})^2} \tag{11}$$

The *Autocorrelations* pane displays the sample autocorrelations together with large lag standard errors and probability limits:

© 2017 by Statgraphics Technologies, Inc

Autocorrelations

			Lower 95.0%	*Upper 95.0%*
Lag	*Autocorrelation*	*Stnd. Error*	*Prob. Limit*	*Prob. Limit*
1	0.768758	0.1	-0.195997	0.195997
2	0.430287	0.147715	-0.289517	0.289517
3	0.105649	0.159758	-0.313121	0.313121
4	-0.0422804	0.160455	-0.314487	0.314487
5	-0.0184315	0.160567	-0.314706	0.314706
6	0.0922117	0.160588	-0.314747	0.314747
7	0.19766	0.161117	-0.315783	0.315783
8	0.158793	0.163523	-0.320501	0.320501
9	0.0316962	0.165058	-0.323509	0.323509
10	-0.175929	0.165119	-0.323628	0.323628
11	-0.31828	0.166983	-0.327281	0.327281
12	-0.370336	0.172943	-0.338963	0.338963
13	-0.183741	0.1807	-0.354166	0.354166
14	0.0663615	0.182558	-0.357809	0.357809
15	0.247179	0.1828	-0.358281	0.358281
16	0.302024	0.186112	-0.364773	0.364773
17	0.205659	0.19095	-0.374256	0.374256
18	0.101909	0.193153	-0.378573	0.378573
19	0.0277422	0.193689	-0.379625	0.379625
20	0.0775123	0.193729	-0.379703	0.379703
21	0.183506	0.194039	-0.38031	0.38031
22	0.282405	0.195767	-0.383697	0.383697
23	0.280609	0.199799	-0.3916	0.3916
24	0.179063	0.203702	-0.399249	0.399249

The standard error for r_k is calculated on the assumption that the autocorrelations have "died out" by lag k and are equal to 0 at all lags greater or equal to k. It is calculated from:

$$se[r_k] = \sqrt{\frac{1}{n}\left\{1 + 2\sum_{i=1}^{k-1} r_k^2\right\}} \quad (12)$$

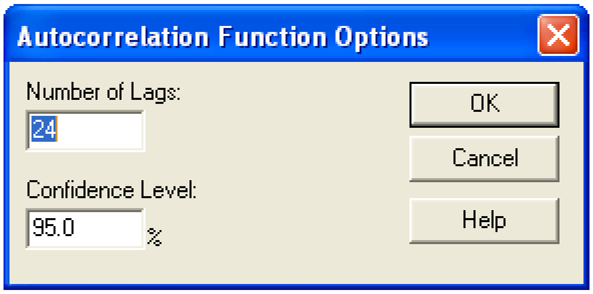
This standard error is then used to calculate *100(1-α)%* probability limits around zero, using a critical value of the standard normal distribution:

$$0 \pm z_{\alpha/2} se[r_k] \quad (13)$$

If $\alpha = 0.05$, any sample autocorrelations that fall outside these limits are statistically significantly different from 0 at the 5% significance level. The StatAdvisor highlights any such autocorrelations in red.

For the sample data, note that there are significant values for the first 2 lags and also at lag 12.

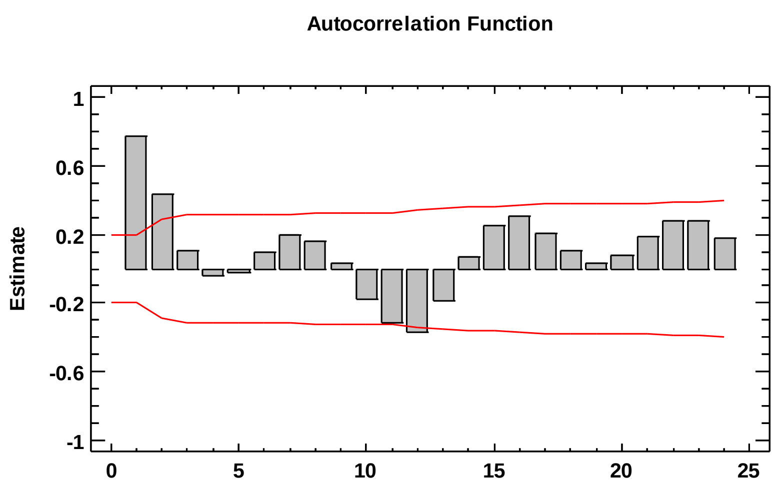
Pane Options



- **Number of lags**: maximum lag *k* at which to calculate the autocorrelation.
- **Confidence level:** value of 100(1-α)% used to calculate the probability limits.

Autocorrelation Function

The *Autocorrelation Function* plot displays the sample autocorrelations and probability limits:



Bars extending beyond the upper or lower limit correspond to statistically significant autocorrelations.

Partial Autocorrelations

Another important tool in modeling time series data is the partial autocorrelation function. The partial autocorrelations are used to help identify the proper order of autoregressive model to use to describe an observed time series. The sample lag *k* partial autocorrelation $\hat{\phi}_{kk}$ is calculated from the sample autocorrelations using:

$$\hat{\phi}_{kk} = \begin{cases} r_1 & k = 1 \\ \dfrac{r_k - \sum_{j=1}^{k-1} \hat{\phi}_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} \hat{\phi}_{k-1,j} r_j} & \text{for } k > 1 \end{cases} \tag{14}$$

where

$$\hat{\phi}_{kj} = \hat{\phi}_{k-1,j} - \hat{\phi}_{kk}\hat{\phi}_{k-1,k-j} \quad \text{for } j = 1, 2, \ldots, k\text{-}1 \tag{15}$$

The *Partial Autocorrelations* pane displays the sample partial autocorrelations together with large lag standard errors and probability limits:

Partial Autocorrelations

	Partial		*Lower 95.0%*	*Upper 95.0%*
Lag	*Autocorrelation*	*Stnd. Error*	*Prob. Limit*	*Prob. Limit*
1	0.768758	0.1	-0.195997	0.195997
2	-0.392904	0.1	-0.195997	0.195997
3	-0.1534	0.1	-0.195997	0.195997
4	0.199171	0.1	-0.195997	0.195997
5	0.136975	0.1	-0.195997	0.195997
6	0.0326462	0.1	-0.195997	0.195997
7	0.038875	0.1	-0.195997	0.195997
8	-0.217107	0.1	-0.195997	0.195997
9	-0.036418	0.1	-0.195997	0.195997
10	-0.193988	0.1	-0.195997	0.195997
11	-0.0270899	0.1	-0.195997	0.195997
12	-0.0823734	0.1	-0.195997	0.195997
13	0.419143	0.1	-0.195997	0.195997
14	0.00270772	0.1	-0.195997	0.195997
15	-0.0287618	0.1	-0.195997	0.195997
16	0.0996899	0.1	-0.195997	0.195997
17	-0.00248173	0.1	-0.195997	0.195997
18	0.101966	0.1	-0.195997	0.195997
19	0.0462495	0.1	-0.195997	0.195997
20	-0.00310214	0.1	-0.195997	0.195997
21	0.0590796	0.1	-0.195997	0.195997
22	-0.0531857	0.1	-0.195997	0.195997
23	-0.100044	0.1	-0.195997	0.195997
24	-0.0560427	0.1	-0.195997	0.195997

The standard error for $\hat{\phi}_{kk}$ is calculated from:

© 2017 by Statgraphics Technologies, Inc

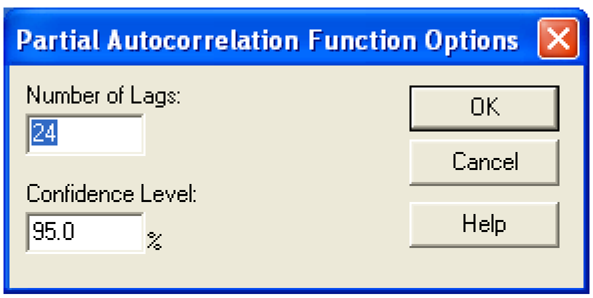
$$se[\hat{\phi}_{kk}] = \sqrt{\frac{1}{n}} \qquad (16)$$

This standard error is then used to calculate *100(1-α)%* probability limits around zero, using a critical value of the standard normal distribution:

$$0 \pm z_{\alpha/2} se[\hat{\phi}_{kk}] \qquad (17)$$

If α = 0.05, any sample partial autocorrelations that fall outside these limits are statistically significantly different from 0 at the 5% significance level. The StatAdvisor highlights any such partial autocorrelations in red.

Pane Options

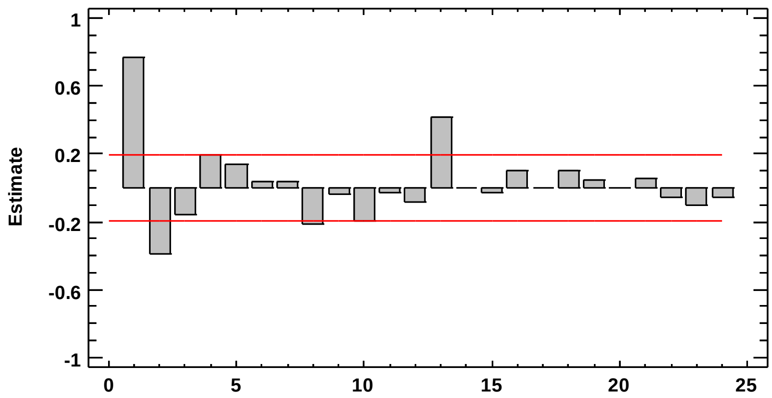


- **Number of lags**: maximum lag *k* at which to calculate the partial autocorrelation.
- **Confidence level:** value of 100(1-α)% used to calculate the probability limits.

Partial Autocorrelation Function

The *Partial Autocorrelation Function* plots the sample partial autocorrelations and probability limits:

Partial Autocorrelation Function



Bars extending beyond the upper or lower limit correspond to statistically significant partial autocorrelations.

Periodogram

The autocorrelations and partial autocorrelations describe the behavior of the data in the *time domain*, i.e., by estimating statistics based on the amount of time between observations. It is also useful to examine the data in the *frequency domain*, by considering how much variability exists at different frequencies. It has been shown that any discrete time series can be represented as the sum of a set of sines and cosines at a set of frequencies called the Fourier frequencies. A typical component has the form

$$a_i \cos(2\pi f_i t) + b_i \sin(2\pi f_i t) \tag{18}$$

where f_i is the *i-th* Fourier frequency. The *i-th* Fourier frequency is

$$f_i = \frac{i}{n} \tag{19}$$

for $i = 0, 1, \ldots, n/2$ if n is even and $i = 0, 1, \ldots, (n\text{-}1)/2$ if n is odd.

The periodogram calculates the power in the data at each Fourier frequency by calculating:

$$I(f_i) = \frac{n}{2}\left(a_i^2 + b_i^2\right) \tag{20}$$

which is scaled so that the sum of the periodogram ordinates across all of the Fourier frequencies except for $i = 0$ yields the sum of squared deviations of the time series about its mean, i.e., $\sum_{i=1}^{n}(y_i - \bar{y})^2$. In effect, the periodogram generates an analysis of variance by frequency.

The *Periodogram* pane displays the following table:

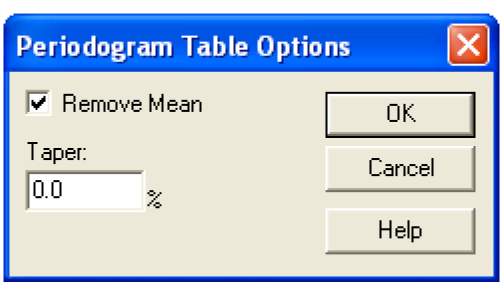
Periodogram Table

i	*Frequency*	*Period*	*Ordinate*	*Cumulative Sum*	*Integrated Periodogram*
0	0.0		0.0	0.0	0.0
1	0.01	100.0	74.6069	74.6069	0.168192
2	0.02	50.0	7.39836	82.0052	0.18487
3	0.03	33.3333	5.53374	87.539	0.197345
4	0.04	25.0	9.42166	96.9606	0.218585
5	0.05	20.0	128.435	225.396	0.508127
6	0.06	16.6667	6.60897	232.005	0.523026
7	0.07	14.2857	1.63018	233.635	0.526701
8	0.08	12.5	1.85305	235.488	0.530878
9	0.09	11.1111	6.17935	241.668	0.544809
10	0.1	10.0	14.8544	256.522	0.578296
11	0.11	9.09091	8.60954	265.132	0.597705
…	…	…	…	…	…

The table includes:

- **Frequency**: the *i-th* Fourier frequency $f_i = i/n$.
- **Period**: the period associated with the Fourier frequency, given by $1/f_i$. This is the number of observations in a complete cycle at that frequency.
- **Ordinate**: the periodogram ordinate $I(f_i)$.
- **Cumulative Sum**: the sum of the periodogram ordinates at all frequencies up to and including the *i-th*.
- **Integrated Periodogram**: the cumulative sum divided by the sum of the periodogram ordinates at all of the Fourier frequencies. This column represents the proportion of the power in the time series at or below the *i-th* frequency.

Pane Options

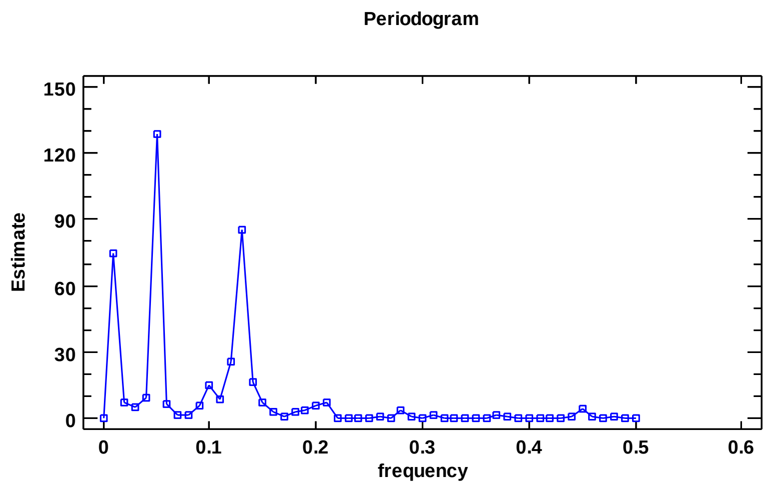


- **Remove mean**: check to subtract the mean from the time series before calculating the periodogram. If the mean is not removed, the ordinate at $i = 0$ is likely to be very large.
- **Taper**: percent of the data at each end of the time series to which a data taper will be applied before the periodogram is calculated. Following Bloomfield (2000), STATGRAPHICS uses

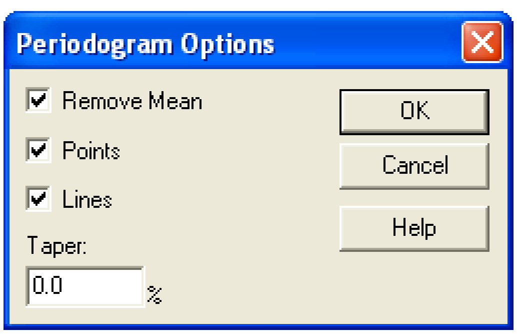
a cosine taper that downweights observations close to $i = 1$ and $i = n$. This is useful for correcting bias if the periodogram ordinates are to be smoothed in order to create an estimate of the underlying spectral density function.

Periodogram Plot

The *Periodogram Plot* displays the periodogram ordinates:



Pane Options



- **Remove mean**: check to subtract the mean from the time series before calculating the periodogram.

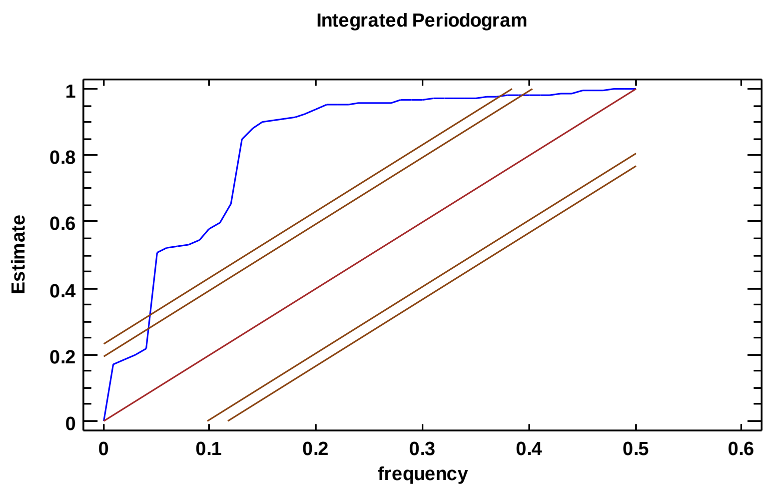
- **Points**: if checked, point symbols will be displayed.

- **Lines**: if checked, the ordinates will be connected by a line.

- **Taper**: percent of the data at each end of the time series to which a data taper will be applied before the periodogram is calculated.

Integrated Periodogram

The *Integrated Periodogram* displays the cumulative sums of the periodogram ordinates, divided by the sum of the ordinates over all of the Fourier frequencies:



A diagonal line is included on the plot, together with 95% and 99% Kolmogorov-Smirnov bounds. If the time series is purely random, the integrated periodogram should fall within those bounds 95% and 99% of the time.

Save Results

The generated data may be saved to a datasheet by pressing the *Save Results* button on the analysis toolbar.

© 2017 by Statgraphics Technologies, Inc