Exploring Data Patterns

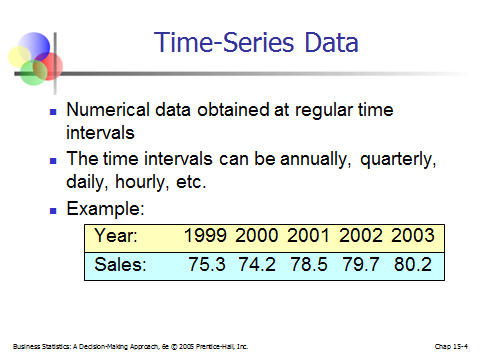
The most sophisticated forecasting model will fail if it is applied to unreliable data.

1. Data should be reliable and accurate (موثوقة و دقيقة).
2. Data should be relevant (لها صلة بالموضوع). The data should be representative of the circumstances for which they are being used.
3. Data should be consistent (ملائمة). When definition concerning data collection changes, adjustment need to be made to retain consistency in historical patterns. For example, when government agencies change the mix or the "market basket" used in determining a cost-of-living index. Years ago personal computers were not part of the mix of products being purchased by consumers; now they are.
4. Data should be timely (لها علاقة بالفترة المدروسة).

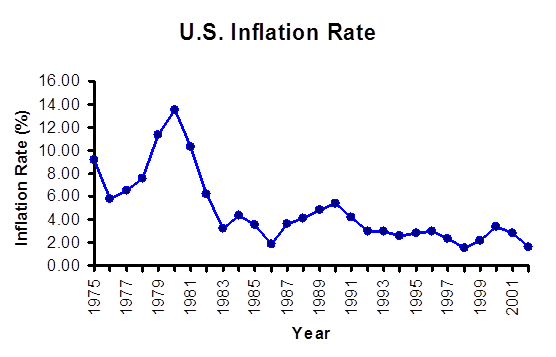
Generally, two types of data: cross-sectional data and a time series data.

**Cross-sectional data:** are observations collected at single point in time.

**A time series data:** are collected over successive increments of time

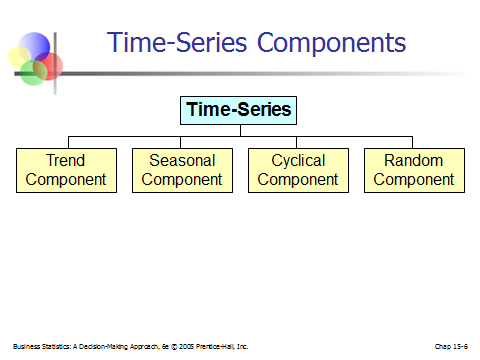


**';vl';df**

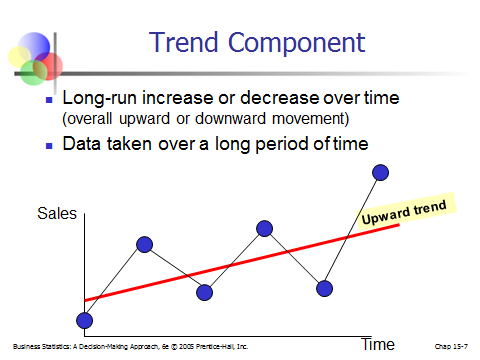


**Exploring time series Data Patterns:**

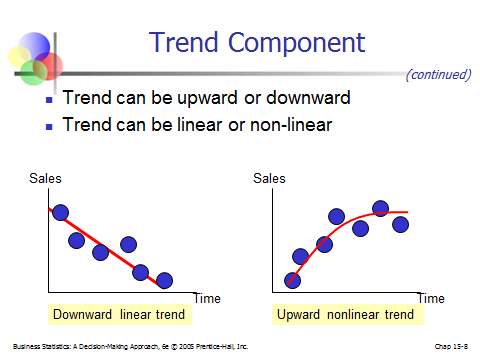
There are typically four general types of patterns: horizontal, trend, seasonal, and cyclical.



When data grow or decline over several time periods, a trend pattern exists. The following Figures show the trend component:

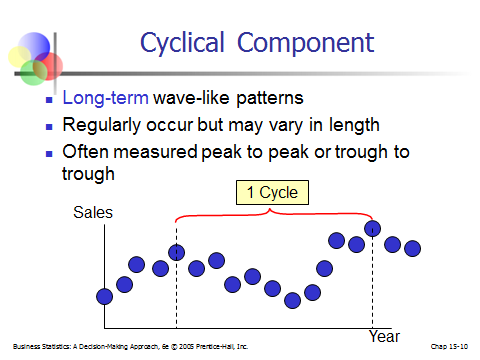


The **trend** is a long term component that represents the growth or decline in the time series over an extended period of time.

****

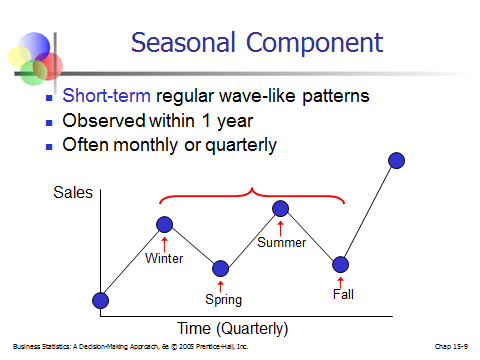
**Time Series Plot**

The **cyclical** component is the wavelike fluctuation around the trend.



When data collected over time fluctuate around a constant level or mean, a horizontal pattern exists. This type of series is to be stationary in its mean. Monthly sales for a food product that do not increase or decrease consistently over an extended period would be considered to have a horizontal pattern.

The **seasonal** component is a pattern that repeats itself year after year.



**Exploring time Patterns with autocorrelation analysis:**

Autocorrelation is the correlation between a variable lagged one or more time periods and itself.

 k = 0, 1, 2......

Where

the autocorrelation coefficient for a lag of k periods

the mean of the values of the series

the observation in time period t

the observation k time periods earlier or at time period t - k

The following is monthly sales data: table 3-1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| t | Month |  |  |  |  |  |  |  |
| 1 | Jan | 123 |  |  | -19 |  | 361 |  |
| 2 | Feb | 130 | 123 |  | -12 | -19 | 144 | 228 |
| 3 | Mar | 125 | 130 | 123 | -17 | -12 | 289 | 204 |
| 4 | Apr | 138 | 125 | 130 | -4 | -17 | 16 | 68 |
| 5 | May | 145 | 138 | 125 | 3 | -4 | 9 | -12 |
| 6 | Jun | 142 | 145 | 138 | 0 | 3 | 0 | 0 |
| 7 | Jul | 141 | 142 | 145 | -1 | 0 | 1 | 0 |
| 8 | Aug | 146 | 141 | 142 | 4 | -1 | 16 | -4 |
| 9 | Sep | 147 | 146 | 141 | 5 | 4 | 25 | 20 |
| 10 | Oct | 157 | 147 | 146 | 15 | 5 | 225 | 75 |
| 11 | Nov | 150 | 157 | 147 | 8 | 15 | 64 | 120 |
| 12 | Dec | 160 | 150 | 157 | 18 | 8 | 324 | 144 |
|  |  |  |  |  | **0** |  | **1474** | **843** |

Example 3.1: compute the lag 1 autocorrelation coefficient and the lag 2 autocorrelation coefficient.



We can say that there is a positive lag 1 autocorrelation in this time series. It is .572. This means that the successive monthly sales are somewhat correlated with each other.

Figure 3-4 Scatter Diagram for Example 3. 1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| t | Month |  |  |  |  |  |  |  |
| 1 | Jan | 123 |  |  | -19 |  | 361 |  |
| 2 | Feb | 130 | 123 |  | -12 |  | 144 |  |
| 3 | Mar | 125 | 130 | 123 | -17 | -19 | 289 | 323 |
| 4 | Apr | 138 | 125 | 130 | -4 | -12 | 16 | 48 |
| 5 | May | 145 | 138 | 125 | 3 | -17 | 9 | -51 |
| 6 | Jun | 142 | 145 | 138 | 0 | -4 | 0 | 0 |
| 7 | Jul | 141 | 142 | 145 | -1 | 3 | 1 | -3 |
| 8 | Aug | 146 | 141 | 142 | 4 | 0 | 16 | 0 |
| 9 | Sep | 147 | 146 | 141 | 5 | -1 | 25 | -5 |
| 10 | Oct | 157 | 147 | 146 | 15 | 4 | 225 | 60 |
| 11 | Nov | 150 | 157 | 147 | 8 | 5 | 64 | 40 |
| 12 | Dec | 160 | 150 | 157 | 18 | 15 | 324 | 270 |
|  |  |  |  |  | **0** |  | **1474** | **682** |



We can say that there is a positive moderate lag 2 autocorrelation in this time series. It is .463. Generally, as the number of time lag (k) increase, the magnitude of the autocorrelation decrease.

Autocorrelation coefficient for different time lags for a variable can be used to answer the following questions about a time series:

1. Are the data random?
2. Do the data have a trend (are they nonstationary)?
3. Are the data stationary?
4. Are the data seasonal?

Are the data random?

If a series is random, the autocorrelation between and  for any time lag k are close to zero. The successive values of a time series are not related to each other.

Do the data have a trend (are they nonstationary)? Are the data stationary?

If a series has a trend, successive observations are highly correlated, and the autocorrelation coefficients typically are significantly different from zero for the first several time lags and then gradually drop toward zero as the number of lags increase.

Are the data seasonal?

If a series has a seasonal pattern, a significant autocorrelation coefficient will occur at the seasonal time lag or multiples of the seasonal lag. The seasonal lag is 4 for quarterly data and 12 for monthly data.

How does an analyst determine whether an autocorrelation coefficient is significantly different from zero for data of table 3-1?

Quenouille (1949) and others have demonstrated that autocorrelation coefficients of random data have a sampling distribution that can be approximated by a normal curve with a mean of zero and an approximated standard deviation of . Knowing this, the analyst can compare the sample autocorrelation coefficients with this theoretical sampling distribution and determine whether, for given time lags, they come from a population whose mean is zero.

Actually, some software packages use a slightly different formula, as shown in Equation 3.2, to compute the standard deviation (or standard error) of the autocorrelation coefficients. This formula assumes that any autocorrelation before time lag k is different from zero and any autocorrelation at time lags greater than or equal to k is different from zero. For an autocorrelation at time lag 1, the standard error is used.

 (3.2)

Where

the standard error (estimated standard deviation) of the autocorrelation at time lag k

the autocorrelation at time lag *i*

the time lag

the number of observations in the time series

This computation will be demonstrated in Example 3.2

If the series is truly random, almost all of the sample autocorrelation coefficients should lie within a range specified by zero, plus or minus a certain number of standard error.

At a specified confidence level, a series can be considered random if each of the calculated autocorrelation coefficients is within the interval about 0 given by, where the multiplier t is an appropriate percentage point of a *t* distribution.

Although testing each  to see if it is individually significantly different from 0 is useful, it is also good practice to examine a set of consecutive's as group. We can use a portmanteau test to see whether the test, say, of the first 10  values, is significantly different from a set in which all values are zero.

One common portmanteau test is based on the Ljung-Box Q statistic:

 (3.3)

Where

the number of observations in the time series

the time lag

the number of time lags to be tested

the sample autocorrelation function of the residuals lagged k time periods

Are the data random?

A simple random model, often called a white noise model, is displayed in Equation 3.4. Observation  is composed of two parts: c, the overall level, and, which is the random error component. It is important to note that the  component is assumed to be uncorrelated from period to period.

 (3.4)

Are the data in table 3-1 consistent with this model? This issue will be explored in Example 3.2 and 3.3.

Example 3.2

A hypothesis test is developed to determine whether a particular autocorrelation coefficient is significantly different from zero for the correlation shown in Figure 3-5. The null and alternative hypotheses for testing the significance of the lag 1 population autocorrelation coefficient are



If the null and hypothesis is true, the test statistic



Has t distribution with df = n-1. Here, n-1 = 12-1 = 11, so for a 5% significance level, the decision rule is as follows:

If t < 2.2 or t > 2.2, reject and conclude the lag 1 autocorrelation is significantly different from 0.

The standard error of  is, and the value of the test statistic becomes



Since -2.2 <1.98<2.2, we cannot reject.

To test for zero autocorrelation at time lag 2, we consider



and test statistic



Has t distribution with df = n-1. Here, n-1 = 12-1 = 11, so for a 5% significance level, the decision rule is as follows:

If t < 2.2 or t > 2.2, reject and conclude the lag 1 autocorrelation is significantly different from 0.

The standard error of  is, and the value of the test statistic becomes



Since -2.2 <1.25<2.2, we cannot reject.

An alternative way to check for significant autocorrelation is to construct, say, 95% confidence limits centered at 0. The limits for time lags 1 and 2 are as follows:

Lag 1:  or   (-.636, .636)

Lag 2:  or   (-.816, .816)

Autocorrelation significantly different from 0 is indicated whenever a value for  falls outside the corresponding confidence limits.

Example 3. 3

Do the Data Have a Trend?

If a series has a trend, a significance relationship exists between successive series values. The autocorrelation coefficients are typically large for the first several time lags and then gradually drop toward zero as the number of lags increases.

A stationary time series is one whose basic statistical properties, such as the mean and variance remain constant over time. Consequently a series that varies about a fixed level (no growth or decline) over time is said to be stationary. A series that contains trend is said to be nonstationary. The autocorrelation coefficients for a stationary series decline to zero fairly rapidly, generally after the second or third time lag. On the other hand, sample autocorrelation for nonstationary series remain fairly large for several time periods. Often, to analyze nonstationary series, the trend is removed before additional modeling occurs. The procedures discussed in chapter 9 use this approach.

A method called *differencing* can be used to remove the trend from nonstationary series. The data originally presented in table 3-1 are shown again in figure 3-8, column A. the  values are lagged one period, , are shown in column B. the differences,  (column A – column B), are shown in column C. for example the first value of differences is = 130-123=7. Note the upward growth or trend of the VCR data shown in Figure 3-9, Plot A. Now observe the stationary pattern of the differenced data in figure 3-9, plot B. Differencing the data has removed the trend.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | E | F | E |
| 1 |  |  | Differences |  |  |  |  |
| 2 | 123 |  |  |  |  |  |  |
| 3 | 130 | 123 | 7 |  |  |  |  |
| 4 | 125 | 130 | -5 |  |  |  |  |
| 5 | 138 | 125 | 13 |  |  |  |  |
| 6 | 145 | 138 | 7 |  |  |  |  |
| 7 | 142 | 145 | -3 |  |  |  |  |
| 8 | 141 | 142 | -1 |  |  |  |  |
| 9 | 146 | 141 | 5 |  |  |  |  |
| 10 | 147 | 146 | 1 |  |  |  |  |
| 11 | 157 | 147 | 10 |  |  |  |  |
| 12 | 150 | 157 | -7 |  |  |  |  |
| 13 | 160 | 150 | 10 |  |  |  |  |

**Figure 3-8 Excel Results of Differencing VCR Data of Example 3.1**

**Figure 3-9 Time series plots the VCR Data and Differenced VCR Data for Example 3-1**

Example 3.4

An analyst for Sears is assigned the task of forecasting operating revenue for 2005. She gathers the data for the years 1955 to 2004, shown in table 3-4. The data are plotted as time series in figure 3-10. Notice that, although Sears operating revenues were declining over the 2000-2004 period, the general trend over the entire 1955-2004 time frame is up. First, the analyst computes a 95% confidence interval for the autocorrelation coefficients at time lag 1 using  where, for large samples, the standard normal .025 points has replaced the corresponding *t* distribution percentage point:





**Table 3-4 Yearly operating revenue for Sears, 1955-2004, for Example 3.4**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year |  | Year |  | Year |  | Year |  | Year |  |
| 1955 | 3307 | 1967 | 7296 | 1979 | 17514 | 1991 | 57242 | 2003 | 23253 |
| 1956 | 3556 | 1968 | 8178 | 1980 | 25195 | 1992 | 52345 | 2004 | 19701 |
| 1957 | 3601 | 1969 | 8844 | 1981 | 27357 | 1993 | 50838 |  |  |
| 1958 | 3721 | 1970 | 9251 | 1982 | 30020 | 1994 | 54559 |  |  |
| 1959 | 4036 | 1971 | 10006 | 1983 | 35883 | 1995 | 34925 |  |  |
| 1960 | 4134 | 1972 | 10991 | 1984 | 38828 | 1996 | 38236 |  |  |
| 1961 | 4268 | 1973 | 12306 | 1985 | 40715 | 1997 | 41296 |  |  |
| 1962 | 4578 | 1974 | 13101 | 1986 | 44282 | 1998 | 41322 |  |  |
| 1963 | 5093 | 1975 | 13639 | 1987 | 48440 | 1999 | 41071 |  |  |
| 1964 | 5716 | 1976 | 14950 | 1988 | 50251 | 2000 | 40937 |  |  |
| 1965 | 6357 | 1977 | 17224 | 1989 | 53794 | 2001 | 36151 |  |  |
| 1966 | 6769 | 1978 | 17946 | 1990 | 55972 | 2002 | 30762 |  |  |

welfjk

Next the analyst runs the data on Minitab and produces the autocorrelation function shown in Figure 3-11. Upon the examination, the analyst notices that the autocorrelation for the first four time lags are significantly different from zero (.96, .92, .87, and .81) and that the values then gradually drop to zero. As a final check, the analyst looks at the Q statistic for 10 times lags. He LBQ is 300.56, which is greater than the chi-square value 18.3 (the upper .05 point of a chi-square distribution with 10 degrees of freedom). This result indicates the autocorrelation for the first 10 lags as a group are significantly different from zero. The analyst decides that the data are highly autocorrelated and exhibit trendlike behavior.

The analyst suspects that the series can be differenced to remove the trend and to create a stationary series. He differences the data (see Minitab applications section at end of the chapter), and the results are shown in Figure 3-12. The differenced series shows no evidence of a trend, and the autocorrelation function, shown in Figure 3-13, appears to support this conclusion. Examining Figure 3-13, the analyst notes that the autocorrelation coefficient at time lag 3, .32 is significantly different from zero (tested at the .05 significance level). The autocorrelations at lags other than lag 3 are small, and the LBQ statistics for 10 lags is also relatively small, so there is little evidence to suggest the differenced data are autocorrelated. Yet the analyst wonders whether there is some pattern in these data can be modeled by one of the more advanced forecasting techniques discussed in Chapter 9.

Are the Data Seasonal?

If quarterly data with a seasonal pattern are analyzed, first quarters tend to look alike, second quarters tend to look alike, and so forth, and a significant autocorrelation coefficient will appear at lag 4. If monthly data are analyzed, a significant autocorrelation coefficient will appear at lag 12. That is January will correlate with other Januarys, February will correlate with other Februarys, and so on. Example 3.5 discusses a series that is seasonal.

Example 3.5

Perkin is an analyst for the Costal Marine Corporation. Perkin gathers the data shown in table 3-5 for the quarterly sales of the corporation from 1994 to 2006 and plots them as the time series

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 3-5** | **Quarterly sales for Costal Marine Corporation, 1994-2006, for example 3.5** | | | |
| year | 31-Dec | 31-Mar | 30-Jun | 30-Sep |
| 1994 | 147.6 | 251.8 | 273.1 | 249.1 |
| 1995 | 139.3 | 221.2 | 260.2 | 259.5 |
| 1996 | 140.5 | 245.5 | 298.8 | 287.0 |
| 1997 | 168.8 | 322.6 | 393.5 | 404.3 |
| 1998 | 259.7 | 401.1 | 464.6 | 479.7 |
| 1999 | 264.4 | 402.6 | 411.3 | 385.9 |
| 2000 | 232.7 | 309.2 | 310.7 | 293.0 |
| 2001 | 205.1 | 234.4 | 285.4 | 285.7 |
| 2002 | 193.2 | 263.7 | 292.5 | 315.2 |
| 2003 | 178.3 | 274.5 | 295.4 | 286.4 |
| 2004 | 190.8 | 263.5 | 318.8 | 305.5 |
| 2005 | 242.6 | 318.8 | 329.6 | 338.2 |
| 2006 | 232.1 | 285.6 | 291.0 | 281.4 |

graph shown in figure 3-14. Next, he computes a large-sample

95% confidence interval for the autocorrelation coefficient at time lag 1: 



Then Perkin computes the autocorrelation coefficients shown in figure 3-15. He notes that the autocorrelation coefficients at time lags 1 and 4 are significantly different from zero (and ). He concludes that the corporation sales are seasonal on a quarterly basis.

Choosing a Forecasting Technique

Some of the questions that must be considered before deciding on the most appropriate forecasting technique for a particular problem follow:

* Why is a forecast needed?
* Who will use the forecast?
* What are the characteristics of the variable data?
* What time period is to be forecasted?
* What are the minimum data requirements?
* How much accuracy is desired?
* What will the forecast cost?

To select the appropriate forecasting technique properly, the forecaster must be able to accomplish the following:

* Define the nature of the forecasting problem.
* Explain the nature of the data under investigation.
* Describe the capability and limitation of potentially useful forecasting technique.
* Develop some predetermined criteria on which the selection decision can be made.

A major factor influencing the selection of a forecasting technique is the identification and understanding of historical patterns of the data. If trend, cyclical, or seasonal patterns can be recognized, then techniques that are capable of effectively extrapolating these patterns can be selected.

**Forecasting Technique for Stationary Data**

A Stationary Series is one whose mean value is not changing over time.

It is important to recognize that stationary data do not necessarily vary randomly about the mean level. Stationary series can be autocorrelated.

Stationary forecasting technique are used in the following circumstances:

* The forces generating a series have stabilized, and the environment in which the series is relatively unchanging.

Example:

The number of breakdowns per week on an assembly line having a uniform production rate.

* A very simple model is needed because of a lack of data or for ease of explanation or implementation.

Example:

When a business is new and very few historical data are available.

* Stability may be obtained by making simple corrections for factors such as population growth or inflation.

Example:

Changing income to per capita income amounts.

* The series may be transformed into a stable one.

Example:

Transforming a series by taking logarithms, square roots, or differences.

* The series is a set of forecast errors from a forecasting technique that is considered adequate.

Example:

(See Example 3.7 on p. 85.)

Techniques that should be considered when forecasting stationary series include naive method, simple averaging methods, moving averages, and autoregressive moving averages (ARMA) models (Box-Jenkins methods).

**Forecasting Technique for Data with trend**

A trend in time series is a persistent, long-term growth or decline.

For a tending time series, the level of the series is not constant. It is common for economic time series to contain a trend.

Forecasting technique for trending data are used in the following circumstances:

* Increased productivity and new technology lead to change in lifestile.

Example:

The demand for electronic components, which increased with advent of the computer.

* Increasing population causes increases in demand for goods and services.

Example:

Increases of the sales revenues of computer goods.

* The purchasing power of the dollar affects the economic variables due to inflation.

Example:

Salaries, production costs, and prices.

Techniques that should be considered when forecasting tending series include moving averages, holt's linear exponential smoothing, simple regression, growth curves, exponential models, and autoregressive integrated moving average (ARIMA) models (Box-Jenkins methods).

**Forecasting Technique for Seasonal Data**

A time series is a series with a pattern of change that repeat itself year after year.

Forecasting technique for seasonal data are used in the following circumstances:

* Weather influences the variable of interest.

Example:

Summer and winter influence activities (e.g., sports such as skiing), and clothing.

* The annual calendar influences the variable of interest.

Example:

Retail sales influenced by holidays.

Techniques that should be considered when forecasting seasonal series include classical decomposition, Census X-12, Winter's exponential smoothing, multiple regression, and autoregressive integrated moving average (ARIMA) models (Box-Jenkins methods).

**Forecasting Technique for Cyclical Data**

A cyclical effect is as wavelike fluctuation around the trend.

Decomposition methods can be extended to analyze cyclical data.

Forecasting technique for seasonal data are used in the following circumstances:

* The business cycle influences the variable of interest.

Example:

Variables like economic, market, and competition factors.

* Shifts in popular tastes occur.

Example:

Shifts like fashions, music, and food.

* Shifts in population occur.

Example:

Shifts like because of war, famines, epidemics and natural disasters.

Techniques that should be considered when forecasting cyclical series include classical decomposition, economic indicators, econometric models, multiple regression, and autoregressive integrated moving average (ARIMA) models (Box-Jenkins methods).