Stationarity and Transformation of Time Series

STA 3050: Time Series and Forecasting

Introduction

In the previous two units of this block, we have seen that time series can be decomposed into four components, i.e., trend, seasonal, cycle, and irregular components. We have also discussed some methods for estimating trend, seasonal and cyclic components and then how to use them for the forecast. Due to several features of the time series, this approach is not necessarily the best one.

According to the modern approach, we try to fit a time series model so that we can forecast the observations. But one of the essential elements of time series modelling is stationarity. A stationary time series is unaffected by the instant at which it is viewed. Most time series forecasting models assume that the underlying time series is stationary. In this unit, you will learn some fundamental concepts that are necessary for a proper understanding of time series modelling. We begin with a simple introduction of stationary and nonstationary time series. Since stationarity is one of the essential elements of a time series, therefore, we discuss various methods of detecting stationarity. As stationarity is necessary to model a time series and if a time series shows a particular type of non-stationarity, then some simple transformation makes it stationary and then we can model them. Therefore, we explain various methods of transforming a nonstationary time series into a stationary one. In the next unit, you will study the concept of correlation in time series.

Expected Learning Outcomes

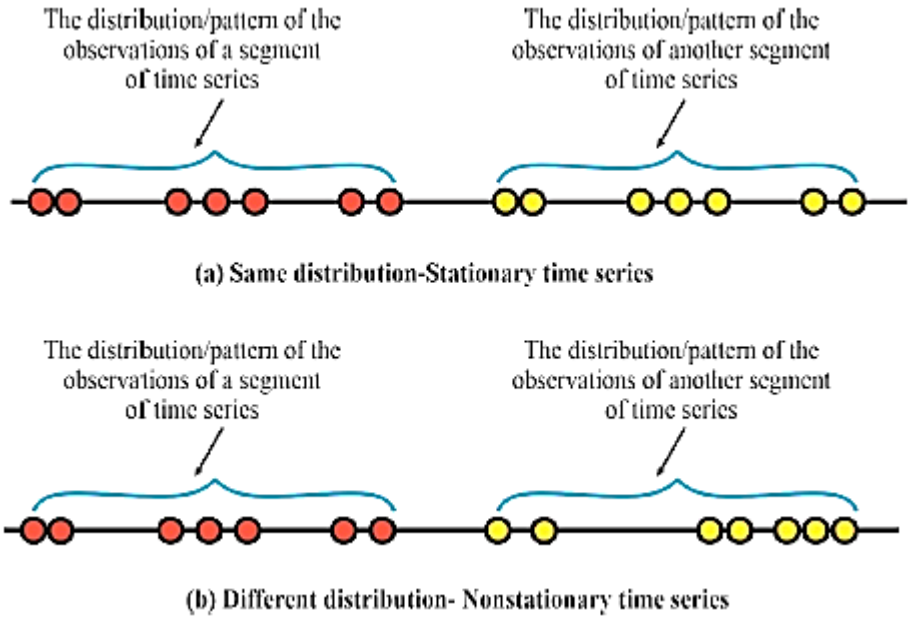
After studying this unit, you would be able to:

- Describe a very useful class of time series, which is called stationary time series.
- Define weak and strict stationary time series.
- Distinguish between stationary and nonstationary time series.
- Apply various methods for detecting stationarity.
- Transform a nonstationary time series to a stationary time series.

Stationary and Nonstationary Time Series

One of the essential elements of time series analysis is stationarity. In simple words, a time series is stationary if it is unaffected by the instant at which it is viewed. Most time series forecasting models assume that the underlying time series is stationary. Think in your mind, “Why is stationarity necessary in time series analysis?” To give the answer to this question, first, we try to understand stationary and nonstationary time series. Therefore, let’s take a moment to discuss the stationary process before moving to the time series models.

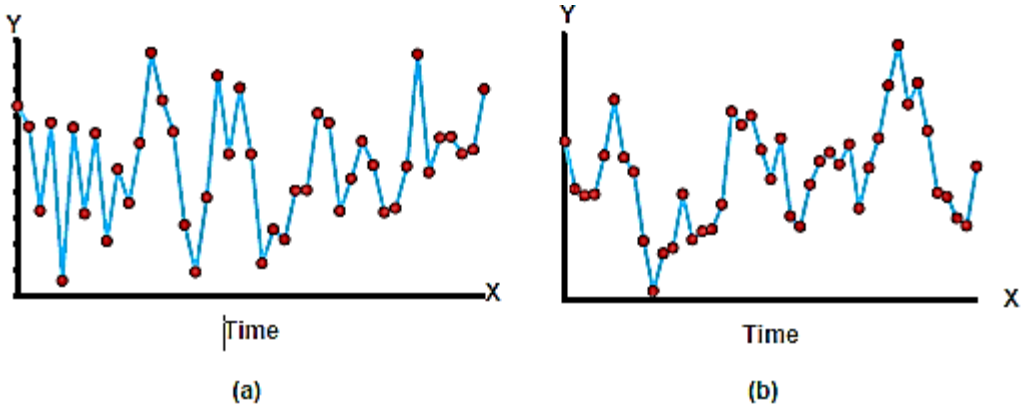
A time series is said to be stationary if the properties of one segment of the time series are similar to the other segment of the time series. In other words, a stationary time series is a series whose statistical properties such as mean, variance, etc., of one section are much like other sections. A time series whose statistical properties change over time is called a nonstationary time series.



Distributions of stationary and nonstationary time series

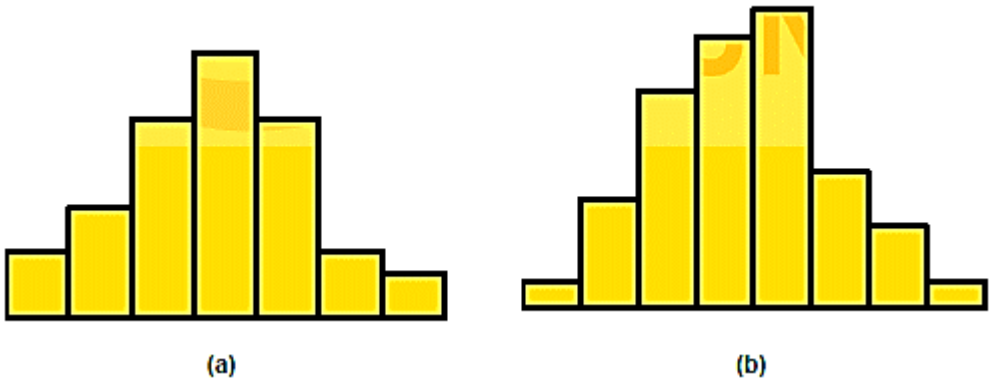
Part (a) of the figure shows that the distribution of the time series remains the same at different segments, therefore, it is a stationary time series whereas part (b) of the figure shows that the distribution of the time series changed with time segment, therefore, it is a nonstationary time series.

We now discuss which statistical properties are used to check the same. As you know that the mean and the variance of data are frequently used as basic statistics to capture characteristics of data. Also, to describe the broad characteristics of the data distribution, a histogram is used. Therefore, by obtaining the mean, the variance and the histogram, it is expected to capture some aspects or features of the data. but it is observed that for two different time series data, the shape of the histogram is almost similar. Consider the two segments of the sales data of a grocery shop in different months as shown below.



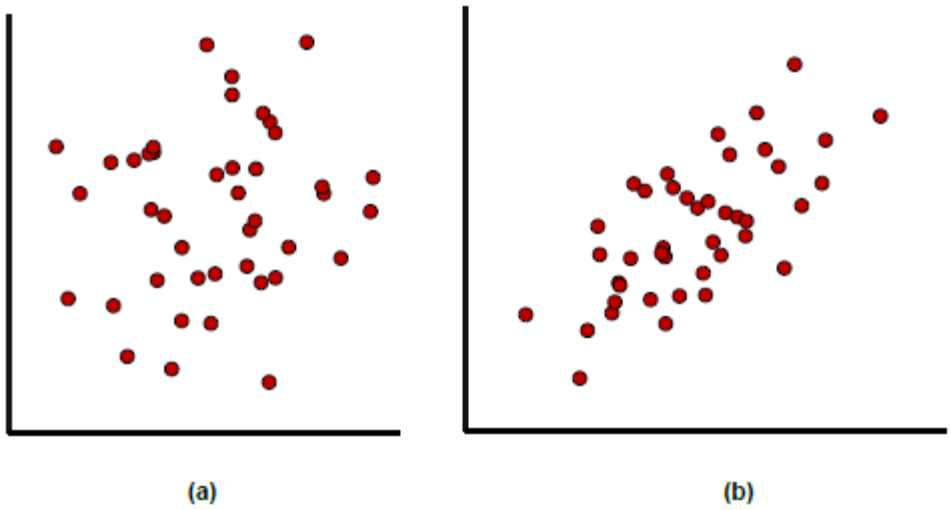
Two segments of the sales data

We can observe that the shape of both segments of the time series appeared differently. We now plot the histograms below.



Histograms of two segments of the sales data

The figure indicates that the histograms of both segments are quite similar, but our time series are quite different. Therefore, it indicates that a histogram is not sufficient to check whether the properties/distribution of one segment of the time series are similar to the other segment. In other words, we can say that only the univariate (marginal) distribution of the time series cannot check the same. Therefore, we move to the joint distribution, and we know that joint distribution describes the properties such as covariance and correlation coefficients. As you know that to get the idea of correlation, we plot the scatter diagram. Therefore, we plot the scatter diagrams of both segments of the time series by taking time series values, say, Y_t on the X-axis and its lag values, say, Y_{t+1} on the Y-axis (You will learn about the term lag in the next sections of this unit.)



Scatterplots with lag of two segments of the sales data

The scatterplot reveals that the data are uniformly distributed about the origin in a circle, which suggests that there is minimal connection between Y_t and Y_{t+1}. On the other hand, the scatterplot shown is concentrated around a line with a positive slope, showing that Y_t and Y_{t+1} have a strong positive correlation.

These examples demonstrate that it is important to consider not only the distribution of Y_t but also the joint distribution of a time series with its lags for checking the stationarity of a time series.

Weak Stationarity

A time series is said to be stationary if its mean, variance and covariance do not change over time. It means that if we find the mean, variance and covariance of one segment of the time series and they will approximately be the same for the other segment of the time series, i.e.,

$$\text{Mean}(Y_t) = \text{Mean}(Y_{t+k}) = \mu \text{ (constant)}$$

$$\text{Var}(Y_t) = \text{Var}(Y_{t+k}) = \sigma^2 \text{ (constant)}$$

$$\text{Cov}(Y_t, Y_s) = \text{Cov}(Y_{t+k}, Y_{s+k}) = \text{constant}$$

where Var and Cov represent variance and covariance, respectively. Then the time series is called stationary but in a strict sense it is called weak stationarity.

If time series data follows a normal distribution, then these properties (mean, variance and covariance) completely describe the distribution but if it is not then these properties are not sufficient to describe the distribution or check the stationarity. Therefore, a time series with these properties is called weakly stationary or covariance stationary.

Strict Stationarity

As we described that if the time series data do not follow the normal distribution, then the mean, variance and covariance do not capture the complete distribution. Therefore, it is necessary to check the joint distribution of the time series. In the strict sense, a time series is called stationary if the joint probability distribution of the observations, Yt, Yt+1, ..., Yt+n remains the same as another set of observations shifted by k $(k > 0)$, k is called lag, time units, that is, $Y_{t+k}, Y_{t+k+1}, \ldots, Y_{t+k+n}$. As a result, a time series is strictly stationary if all statistical measures (such as mean, variance, higher moments, etc.) are constant with respect to time, i.e., do not depend on t.

Non-Stationarity

In simple words, a time series is said to be nonstationary if the statistical properties of one segment of the time series are not similar to the other segment of the time series, that is, the mean, variance, and covariance of the time series change over time. Therefore, a time series which contains trend, seasonality, cycles, random walks, or combinations of these is nonstationary. Nonstationary data, as a rule, are unpredictable and cannot be modelled or forecasted. In order to receive consistent, reliable results, the nonstationary data needs to be transformed into stationary data.

Detecting Stationarity

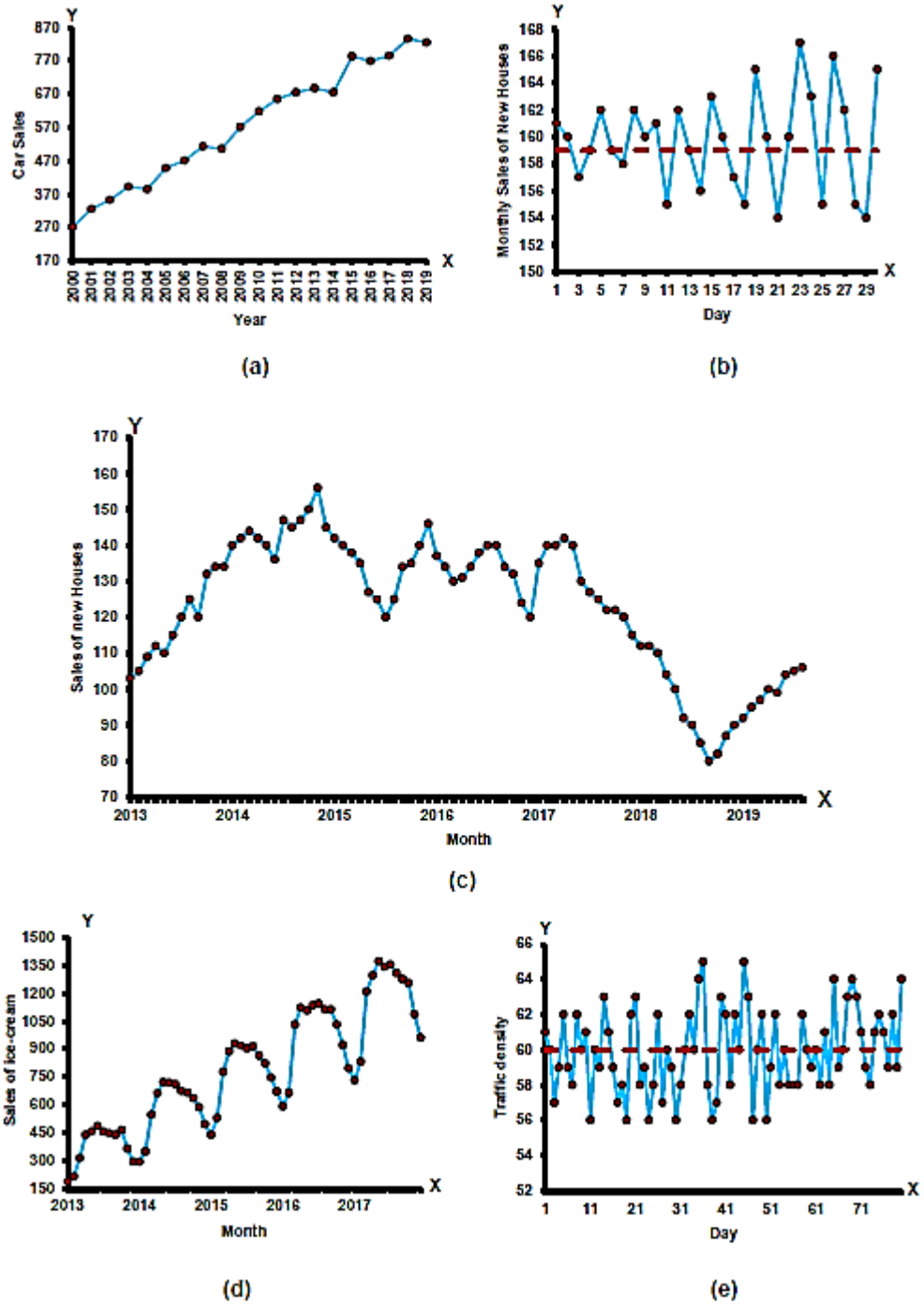
As we have seen in the above section, stationarity is necessary for reliable modelling and forecasting, therefore, it is very important to ascertain whether a given time series is stationary or not. We are describing some ways through which you can check whether a given time series is stationary or nonstationary.

Visualisation

The simplest way to check whether a given data comes from a stationary series or not is to plot the data or some function of it. Both stationary and nonstationary time series have some properties that can be detected very easily from the plot of the data.

Looking at the Data

First, we plot the data with respect to time and try to understand the pattern of mean and variance. If the plot shows roughly horizontal (although some cyclic behaviour is possible), with constant variance then this indicates that the series is stationary. The data points in a stationary series would constantly move in the direction of the long-run mean with a constant variance. If the data points might show some trend or seasonality, then these are an indication of a nonstationary series. For more explanation, we consider some time series plotted below.



Time series plots of various data

Which of these do you think are stationary? Let us discuss them one at a time.

If you look at the first plot (car sales) (a), we can see that the mean varies (increases) with time which results in an upward trend. Thus, this is a nonstationary series. For a series to be classified as stationary, it should not exhibit a trend.

Moving on to the second plot (b), we can see that there is no trend in the series, but the variance of the series increases with time. As mentioned previously, a stationary series must have a constant variance.

The third plot (c) shows that the sales of new houses first increase and then decrease over a span of time, therefore, it is not also a stationary series.

The fourth plot (d) shows the seasonality as well as the trend (upward) and a regularly repeating pattern of highs and lows related to months of the year. Therefore, it is not a stationary series.

If you look at the fifth plot (e), we can see that there is no consistent trend (upward or downward) over the entire time span. The series appears to slowly wander up and down. The horizontal line drawn at 60 indicates the mean of the series and we notice that the series tends to stay on the same side of the mean (above or below) for a while and then wanders to the other side. Also, the variance is constant. Almost by definition, there is no seasonality. So we can say that this time series is stationary.

This method is only used to get an idea about stationarity and is not completely reliable.

Autocorrelation Functions Plots

The time series plot gives only the idea of stationarity. Autocorrelation function plot also known as correlogram is another method through which we can look for the stationarity of a time series more accurately. We will describe it in the next section.

Summary Statistics

As we discussed, a stationary time series has a constant mean, variance, etc. over time. Therefore, we can use summary statistics like mean and variance to check whether a time series is stationary or not.

In this method, we divide the data into two or more random groups and for each group, we calculate the summary statistics as the mean or other moments. After that, we analyse the summary statistics of such groups. If the mean and variance of these groups are very close to each other, the series is stationary otherwise, it is not stationary.

For example, we split the data of the traffic intensity as discussed above into two halves and calculate the mean and variance of each group as:

$$\text{Mean of the first group} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i = 60$$

$$\text{Mean of the second group} = \frac{1}{n_2}\sum_{i=1}^{n_2} x_i = 60.5$$

$$\text{Variance of the first group} = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 = 5.69$$

$$\text{Variance of the second group} = \frac{1}{n_2}\sum_{i=1}^{n_2}(x_i - \bar{x}_2)^2 = 5.26$$

We can observe that the mean and variance of both groups are very close to each other, therefore, the series is stationary.

In a similar way, if we split the data of the monthly sales of new houses sold in the region into two halves and calculate mean and variance of each group as:

$$\text{Mean of the first group} = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i = 132$$

$$\text{Mean of the second group} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i = 114$$

$$\text{Variance of the first group} = \frac{1}{n_1}\sum_{i=1}^{n_1}(y_i - \bar{y}_1)^2 = 168.48$$

$$\text{Variance of the second group} = \frac{1}{n_2}\sum_{i=1}^{n_2}(y_i - \bar{y}_2)^2 = 366.45$$

We can observe that there is a big difference between the two variances and the two means' values. This implies that the series is not stationary.

If we want to apply a more effective and practical way to check whether the series is stationary or not, then we may use different statistical tests which are discussed in the next sub-sections.

Statistical Tests

There are both parametric and nonparametric tests that may be used to check whether a time series is stationary or not. For testing the stationarity, we formulate the hypotheses as:

- Null Hypothesis H0: The time series is not stationary.
- Alternative Hypothesis H1: The time series is stationary.

For testing the hypotheses, two of the most used tests to test for stationarity are the Augmented Dickey-Fuller test and the Kwatkowski-Phillips-Schmidt-Shin test. These tests are beyond the scope of this unit. We shall not discuss these here and if someone is interested in these, may refer to "Time Series Analysis Forecasting and Control", 4th Edition, written by Box, Jenkins and Reinsel.

Transforming Nonstationary Time Series into Stationary

In the previous sections, you learnt stationary and nonstationary time series and why stationarity is important in a time series. But if we observe a nonstationary time series then how can we draw a reliable forecast? One way for that is transforming a nonstationary time series to a stationary series. We can do that by identifying and removing trends and seasonal effects. There are following main methods for doing the same:

- Differencing
- Seasonal Differencing
- Log-Transformation
- Power Transformation
- Box-Cox transformation

Let us learn about these one at a time.

Differencing

It is one of the simplest methods for removing a systematic structure from the time series. This method is used when the series exhibits non-stabilise mean, that is, trend and seasonality. Therefore, it is typically performed to get rid of the varying mean. For example, a trend can be removed by subtracting the previous value from each value in the series. In this method, we compute the difference between consecutive observations in a series. Mathematically, it can be applied as:

$$Y_t' = Y_t - Y_{t-1}$$

where Y_t and Y_{t-1} are the values of the time series at time point t and $t-1$, respectively and Y_t' represents the first-order difference.

This is called first-order differencing. In some situations, a nonstationary time series will nevertheless convert into stationary from a single difference. In that case, a second-order differencing is required. Second-order differencing is the change between two consecutive data points in a first-order differenced time series. In general, differencing of order d is used to convert nonstationary time series to stationary time series. It is tricky to implement because the inverse operation of differencing is the cumulative sum. This is not as straightforward as a transformation because when we apply the inverse of the differenced data to our forecasts, we must add in the last known observation of our series in order to get the correct transformation.

Seasonal Differencing

If a time series shows a seasonality effect, then to remove this effect, we calculate the difference between an observation and a previous observation from the same season instead of calculating the difference between consecutive values. For example, if we observe a monthly seasonal effect then we subtract the observation, say, the month of July of a year from an observation taken in July of the previous year. If a season has a period m, then mathematically it can be written as:

$$Y_t' = Y_t - Y_{t-m}$$

If there is a seasonal component at the level of one week, then we can remove it on an observation today by subtracting the value from last week, that is, $m = 7$.

The method of differencing has the main drawback of losing one observation each time

when the difference is calculated and it does not remove a non-linear trend.

Log-Transformation

In time series analysis, the log-transformation is often used to stabilise the variance and remove the non-linearity trend from a time series. Time series with an exponential trend can be made linear by taking the logarithm of the values. If we denote the original observations as $Y_1, Y_2, \ldots, Y_N$ then we make the transformation as:

$$Z_t = \log(Y_t)$$

where $Z_1, Z_2, \ldots, Z_N$ denote the transformed observations and the logarithm is natural (i.e., to base e). Since the stationarity of the time series helps us construct the forecast model, therefore, it is important to apply the inverse of that transformation to the data in order to get back to the original scale. Therefore,

$$Y_t = \exp(Z_t)$$

Power Transformation

The method of log-transformation has the main drawback that it can be applied when the observations are positive and non-zero because the log of negative observation is not defined and the log of zero is infinite. Other transformations may also be used such as square roots and cube roots. They are also called power transformations because they take the form as follows:

$$Z_t = Y_t^p$$

The inverse of the transformation is:

$$Y_t = Z_t^{\frac{1}{p}}$$

Box-Cox Transformations

George Box and Sir David Cox proposed a very useful family of transformations called Box-Cox transformations. The beauty of this transformation is that it includes both logarithms and power transformations and is defined as follows:

$$Z_t = \begin{cases} \log(Y_t) & \text{if } \lambda = 0 \\ \dfrac{Y_t^{\lambda} - 1}{\lambda} & \text{otherwise} \end{cases}$$

In this transformation, the logarithm is natural (i.e., to base e). It depends on the parameter λ which varies from -5 to 5. If $\lambda = 0$, then this transformation uses the log-transformation whereas if $\lambda \neq 0$, then a power transformation. We consider all values of λ and select the optimal value for our data. The "optimal value" is the one which results in the variance stationary. We list some common values used for λ as follows:

- $\lambda = -1$ is a reciprocal transform.
- $\lambda = -0.5$ is a reciprocal square root transform.
- $\lambda = 0$ is a log transform.
- $\lambda = 0.5$ is a square root transform.
- $\lambda = 1$ is no transform.

In the case of the Box-Cox method, we can find the original values as:

$$Y_t = \begin{cases} \exp(Z_t) & \text{if } \lambda = 0 \\ (Z_t\lambda + 1)^{\frac{1}{\lambda}} & \text{otherwise} \end{cases}$$

Example 1:

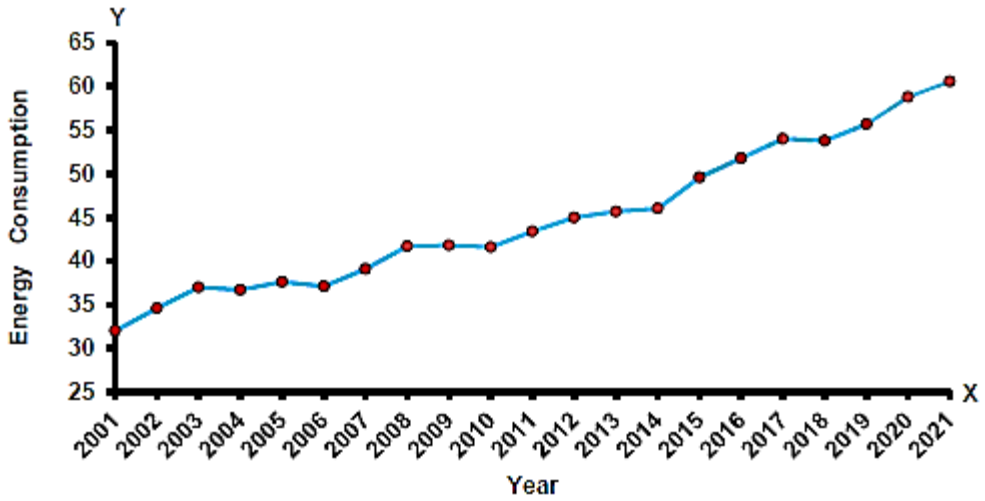
The data given below presents the information on total annual energy consumption (EC) in a particular area over different years.

Year	EC (in units)	Year	EC (in units)
2001	32.0	2012	45.0
2002	34.6	2013	45.7
2003	37.0	2014	46.0
2004	36.7	2015	49.6
2005	37.6	2016	51.8
2006	37.1	2017	54.0
2007	39.1	2018	53.8
2008	41.7	2019	55.7
2009	41.8	2020	58.8
2010	41.6	2021	60.6
2011	43.4		

(i) Plot the energy consumption data.
(ii) Is there an indication of nonstationary behavior in the time series?
(iii) If yes, calculate the first difference of the time series and plot it.
(iv) What impact has differencing had on the time series?

Solution

First, we plot the time series data by taking years on the X-axis and the energy consumption on the Y-axis. We get the time series plot as shown



From the plot, we can see that the mean energy consumption varies (increases) with time, which results in an upward trend. Thus, this is a non-stationary time series. For a series to be classified as stationary, it should not exhibit a trend.

To remove the trend, we obtain the first-order difference, that is, we compute the difference between consecutive observations in the series by subtracting the previous value from each value in the series. Mathematically, it can be expressed as:

$$Y_t' = Y_t - Y_{t-1}$$

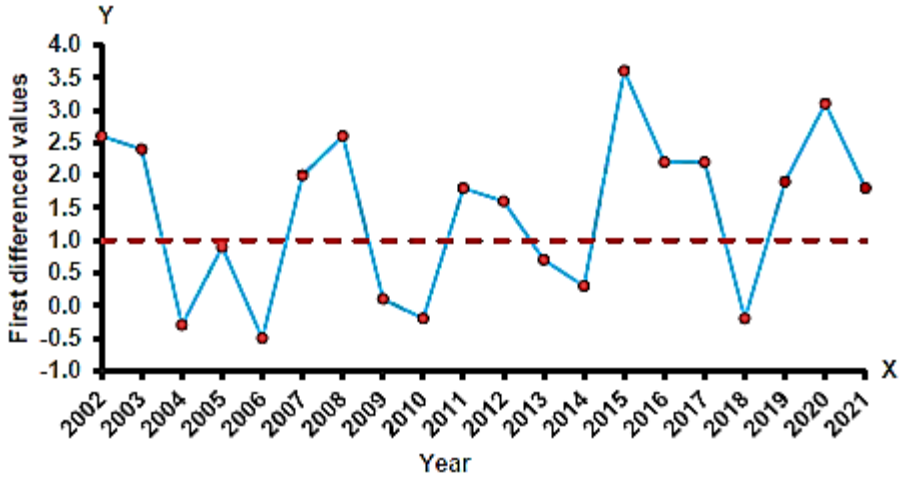
For example,

$$Y_{2002}' = Y_{2002} - Y_{2001} = 34.6 - 32.0 = 2.6$$

You calculate the rest of the values in a similar way, which are given in the following table:

Year	EC (in units)	Yt'	Year	EC (in units)	Yt'
2001	32.0	-	2011	43.4	1.8
2002	34.6	2.6	2012	45.0	1.6
2003	37.0	2.4	2013	45.7	0.7
2004	36.7	-0.3	2014	46.0	0.3
2005	37.6	0.9	2015	49.6	3.6
2006	37.1	-0.5	2016	51.8	2.2
2007	39.1	2.0	2017	54.0	2.2
2008	41.7	2.6	2018	53.8	-0.2
2009	41.8	0.1	2019	55.7	1.9
2010	41.6	-0.2	2020	58.8	3.1

To study the impact of the first-order differencing on the pattern of the time series, we plot the first-order difference values against time (years) in Fig. 12.8.



We observe that there is no consistent trend (upward or downward) over the entire time span. It means that the first-order difference removes the trend effect, and the time series becomes almost stationary.

Summary

In this unit, we have discussed:

- A time series is said to be stationary if the statistical properties of one segment of the time series are similar to the other segment of the time series otherwise it is called a nonstationary time series.
- Various methods for detecting stationarity such as visualisation, summary statistics and statistical tests.
- Various methods of transforming nonstationary time series to stationary such as differencing, seasonal differencing, log-transformation, power transformation, and Box-Cox transformation.

Questions

A researcher wants to study the pattern of sales of a new single house in a region. She collects the data of the number of new single house sales for 15 months in that region which are given as follows:

Month	Sales	Month	Sales
1	116	9	290
2	154	10	300
3	175	11	315
4	207	12	345
5	225	13	353
6	230	14	385
7	245	15	410

Month	Sales	Month	Sales
8	270		

For the data:

1. Plot the time series data and comment on any features of the data that you see.
2. If the plot will show non-stationarity then transfer the data using first difference and plot new time series data.
3. What impact has differencing had on the time series? Is the new time series stationary or nonstationary?