**Time series** is a collection of data points collected at **constant time intervals**(days, weeks, months, quarters, annual). These are analyzed to determine the long term trend so as to forecast the future or perform some other form of analysis.

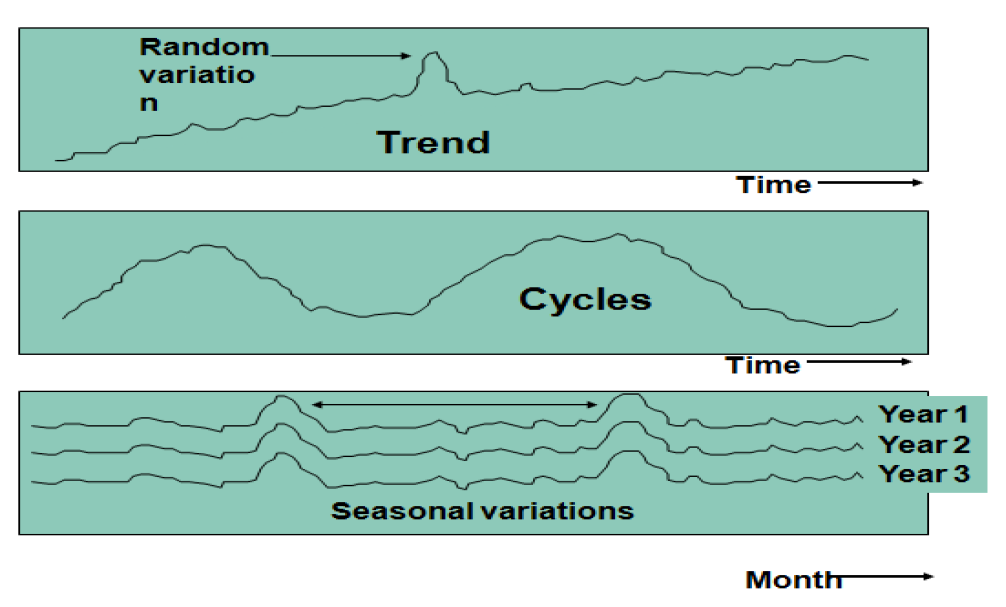
* **Time series analysis**is the use of statistical methods to **analyze time series** data and extract meaningful statistics and characteristics about the data.
* The analysis of time series is based on the assumption that successive values in the data file represent consecutive measurements taken at equally spaced time intervals.
* Time series analysis accounts for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for.
* Time series data is indexed in time order.
* There are two main goals of time series analysis: (a) identifying the nature of the phenomenon represented by the sequence of observations, and (b) forecasting (predicting future values of the time series variable).

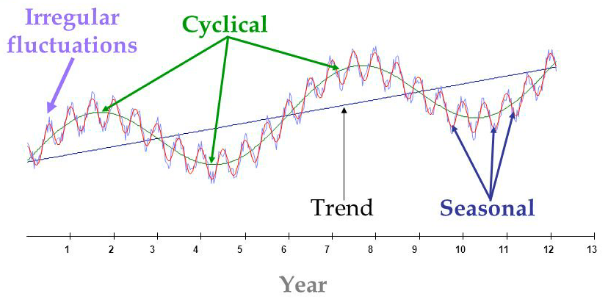
**Applications**

* Sales Forecasting
* Budgetary Analysis
* Stock Market Analysis
* Yield Projections
* Process and Quality Control
* Inventory Studies
* Workload Projections
* Utility Studies
* Census Analysis

**Components**

* **Trend(T):** Gradual long term movement (up or down) in data. Easiest to detect. Eg: Population growth In India
* **Seasonal Pattern(S):** Results from events that are periodic and recurrent in nature. An up-and-down repetitive movement within a trend occurring periodically. Often weather-related data could be daily or weekly occurrence. Short-term regular variations in data. Eg. Sales in festive seasons
* **Cyclical Patterns(C):** Results from events recurrent but not periodic in nature. An up-and-down repetitive movement in demand. Repeats itself over a long period of time. Wavelike variations of more than one year’s duration. Eg. Recession in Economy
* **Irregular Component(I):** Disturbances or residual variation that remain after all the other behaviors have been accounted for. Erratic movements that are not predictable because they do not follow a pattern. Caused by chance and unusual circumstances. Eg. Earthquake





**Time Series Forecasting Techniques**

* Simple moving average smoothing
* Weighted moving average smoothing
* Single Exponential smoothing
* Holt’s linear trend method
* Holt’s Winter seasonal method
* ARIMA

**Smoothing Techniques**

Statistical **technique** for removal of short-term irregularities in a time-series data to improve the accuracy of forecasts.

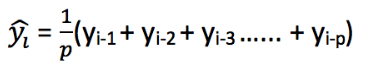
It is of 2 types:

* Moving Average
* Exponential

**Simple average**

**Simple Moving Average Smoothing**

The target value is equal to the average of all the previous observations.



**Weighted Moving Average Smoothing**

The target value is equal to the weighted average of all the previous observations with most recent observation carrying higher weight.



**Exponential Smoothing**

Exponential Smoothing is one of the more popular smoothing techniques due to its flexibility, ease in calculation, and good performance.

Exponential Smoothing uses a simple average calculation to assign exponentially decreasing weights starting with the most recent observations.

New observations are given relatively more weight in the average calculation than older observations.

**Single Exponential Smoothing**

The target value is equal to the weighted average of all the previous observations but here weights are assigned in an exponentially decreasing manner starting with the most recent observations.

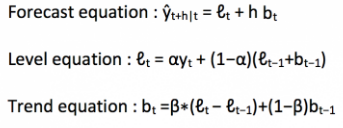


Where, 0≤ α ≤1 is the **smoothing** parameter

It should only be used when there is **no seasonality or trend** in the data.

**Holt’s Linear Trend Method(Double Exponential Smoothing)**

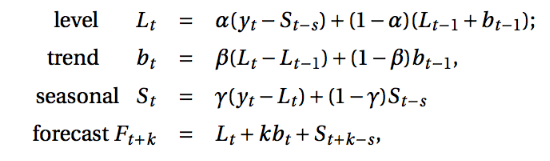
It is used when there is **trend in the data but no seasonality**.



Where, 0≤ β≤1 is the factor which accounts for trend in the data.

**Holt’s Winter Seasonal Method (Triple Exponential Smoothing)**

It is used when there is **both trend & seasonality in the data**.

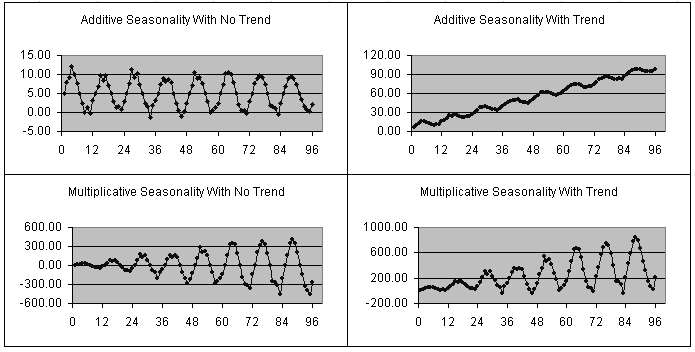


Where s is the length of the seasonal cycle, for 0 ≤ α ≤ 1, 0 ≤ β ≤ 1 and 0 ≤ γ ≤ 1

0≤ γ ≤1 is the factor which accounts for seasonality in the data.

**Decomposition Methods**

* Additive:  X = Trend + Seasonal + Random(Irregular)
* Multiplicative:  X = Trend \* Seasonal \* Random(Irregular)

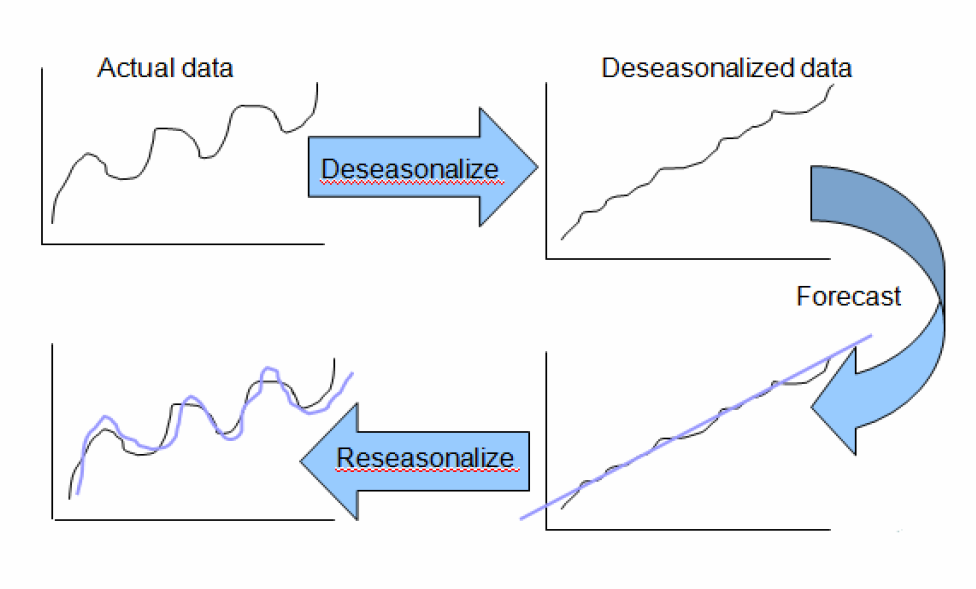
****

**Additive vs. Multiplicative Seasonality**

* In additive seasonal model, the appropriate seasonal index is *added* to the base of the forecast.
* In a multiplicative model, the appropriate seasonal index is *multiplied* with the base of the forecast.
* In many time series, the amplitude of both the seasonal and irregular variations increase as the level of the trend rises. In this situation, a multiplicative model is usually appropriate.
* In some time series, the amplitude of both the seasonal and irregular variations do not change as the level of the trend rises or falls. In such cases, an additive model is appropriate.

**Forecasting Process**

* Calculate seasonal index.
* Remove seasonality from the data by dividing actual values with seasonal index.
* Forecast new values based on the new data set obtained after removing seasonality.
* Add back the seasonality to final forecasted values.



**Stationary Time Series**

* A stationary time series is said to be stationary if it has no trend & no seasonality.
* Its properties (mean, standard deviation) are independent of time.

**Methods to check Stationarity of Time Series**

**Augmented Dickey-Fuller Test (ADF Test)**

It is a type of statistical test called as a unit root test.

In probability theory of statistics, a unit root is a feature of stochastic processes that can cause problems in statistical inference involving time series models.

In simple terms, unit root is non-stationary.

**ADF test assumptions:**

Null Hypothesis: Series is non-stationary or series has a unit root

Alternate Hypothesis: Series is stationary or series has no unit root

Conditions to Reject Null Hypothesis:

If p-value<0.05, **Reject Null Hypothesis**, i.e., time series does not have a unit root, meaning it is stationary.

**ACF**

**Autocorrelation** is a measure of the internal correlation within a **time series**. It is a way of measuring and explaining internal association between observations in a **time series**.

A correlogram (also called Auto Correlation Function **ACF Plot** or Autocorrelation **plot**) is a visual way to show serial correlation in data that changes over time (i.e. time series data).

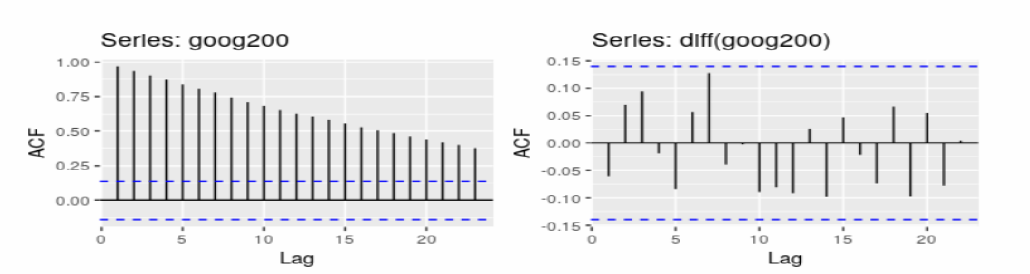
It is **auto-correlation** of any time series with its **lagged** values.

ACF describes the strength of the relationships between different points in the series.

Autocorrelations for consecutive lags are formally dependent.

It describes how well the present value of the series is related with its past values.

ACF considers all components (trend, seasonality, cyclic and residual) while finding correlations.



**PACF**

**Partial autocorrelation function** (**PACF**) give correlation between time series & residuals of previous lags.

Residuals are the values which remain after removing the effects already explained by the earlier lags with the next lag value. Hence it is partial & not complete as we have removed already found variations before we found next correlation.

It is correlation between a tome series & its error values.

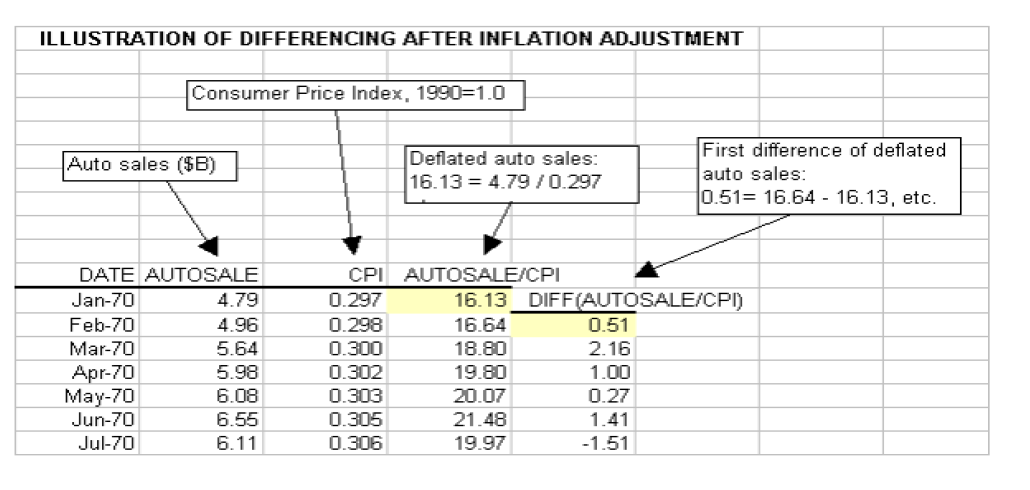
**Methods to make time series stationary**

**Differencing**: Compute the differences between consecutive observations. Differencing can help stabilise the mean of a time series by removing changes in the level of a time series, and therefore eliminating (or reducing) trend and seasonality.

**Transformations**: Logarithmic transformations can help to stabilise the variance of a time series & removes exponential trends.

For non-constant variance, taking the logarithm or square root of the series may stabilize the variance.

For negative data, you can add a suitable constant to make all the data positive before applying the transformation. This constant can then be subtracted from the model to obtain predicted (i.e., the fitted) values and forecasts for future points.



**ARIMA**

ARIMA stands for **Auto-Regressive Integrated Moving Average**.

Lags of the stationary series in the forecasting equation are called **autoregressive** terms

Lags of the forecast errors are called **moving average** terms, and

A time series which needs to be differenced to be made stationary is said to be an **integrated** version of a stationary series.

A non seasonal ARIMA model is classified as an **ARIMA(p,d,q)** model, where:

**p is the number of autoregressive terms,**

**d is the number of non seasonal differences needed for stationarity, and**

**q is the number of lagged forecast errors in the prediction equation**

**AR:Autoregressive**

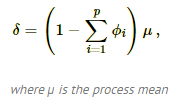
A model that uses the dependent relationship between an observation and some number of lagged observations.

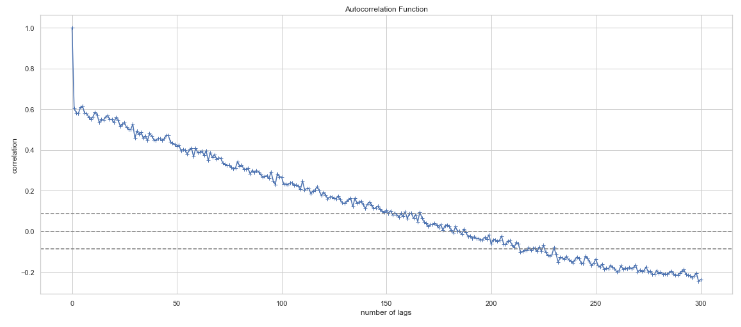
**The AR model is defined by the equation:**

yt= δ + a1yt-1+ a2yt-2+ … + apyt-p+ et

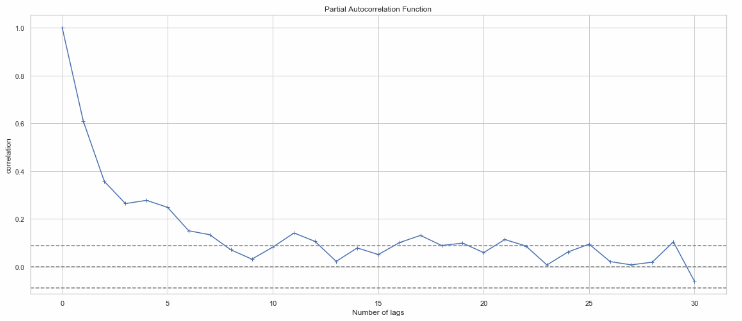
Where:

* yt-1, yt-2…yt-p are the past series values (lags),
* et  is white noise (i.e. randomness),
* and δ is defined by the following equation:





In the above correlation plot dotted lines represent the confidence band, with center dotted line represents mean and upper and lower dotted line represent boundries based on 95% confidence interval.  
Notice that we have good positive correlation with the lags upto 150, this is the point where ACF plot cuts the upper confidence threshold. Although we have good correlation upto 150th lag we cannot use all of them as it will create multi-collinearity problem, thats why we turn to PACF plot to get only the most relevant lags!



In the above plot we can see that lags upto 7 have good correlation before the plot first cuts the upper confidence interval. This is our p value i.e the order of our AR process. We can model given AR process using linear combination of first 7 lags.

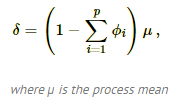
**MA**: Moving Average

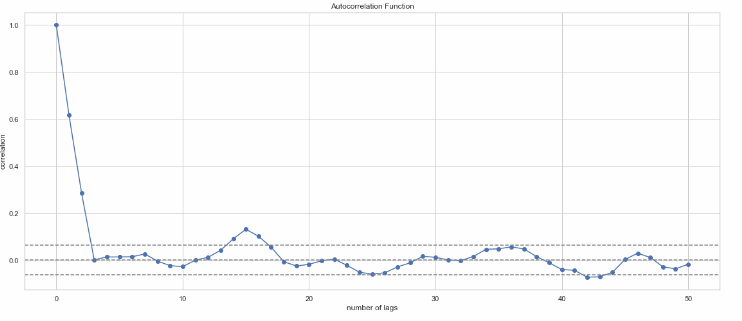
A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

**The MA model is defined by the equation:**  
yt = δ + et - b1et-1

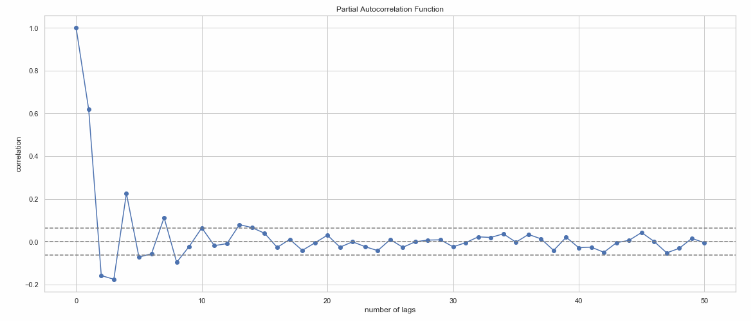
Where:

* yt-1, yt-2…yt-p are the past series values (lags),
* et  is white noise (i.e. randomness),
* and δ is defined by the following equation:





As per above plot we have good correlation upto 2nd lag, this is the lag after which plot cuts the upper confidence interval. Order q of series obtained by the plot is 2, which is correct as we had defined our series with linear combination of residuals upto lag 2.  
Thus this proves that ACF correctly predicted order of our MA series.

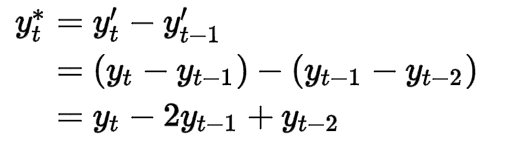


If we see PACF plot there are many instances where correlation is above upper confidence band as PACF calculates correlations of lags of time series with residuals and our series itself is linear combination of residual and its lagged values. Hence we can get good correlation for near as well as past lags.

**I** : (Integrated)

A model that uses the **differencing** of raw observations.

**Differencing** in statistics is a transformation applied to time-series data in order to make it stationary.



**Seasonal ARIMA (SARIMA) models : This model is used when the time series exhibits seasonality.**

**ARIMA(p,d,q)(P, D, Q)m**

**where,**

* **p — the number of autoregressive**
* **d — degree of differencing**
* **q — the number of moving average terms**
* **m — refers to the number of periods in each season**
* **(P, D, Q )— represents the (p,d,q) for the seasonal part of the time series**
* **In Purely seasonal AR model, ACF decays slowly while PACF cuts off to zero.**
* **In Purely seasonal MA model, ACF cuts off to zero and PACF decays slowly.**

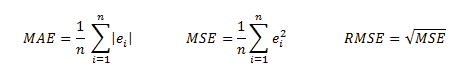
**Forecast Accuracy**

If y1, …, y*n* represents a time series, then ŷ*i* represents the *i*th forecasted value, where *i ≤ n*.

For *i ≤ n*, the*i*th error*ei* is then

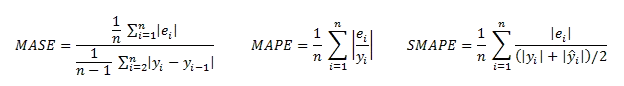
*ei=* y*i –* ŷ*i*

Our goal is to find a forecast that minimize the errors. A number of measures are commonly used to determine the accuracy of a forecast, **including the mean absolute error (MAE), mean squared error (MSE) and root mean squared error (RMSE).**

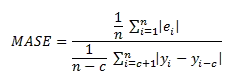
****

**MAE is also commonly called mean absolute deviation (MAD).**

Some other measurements are **mean absolute percentage error (MAPE),** **mean absolute scaled error (MASE)** and **symmetric mean absolute percentage error (SMAPE)**.



For data with seasonality where the periodicity is*c*, the formula for *MASE* becomes



**Box Jenkins Procedure**

**The steps are**

* Plot the time series data to check for trend and seasonality.
* Checking for stationarity or non-stationarity.
* Transforming the data, if necessary.
* Plot ACF, and PACF.
* Identification of a suitable ARIMA model (AR, MA, or ARIMA).
* Estimation of the parameters of the chosen model.
* Diagnostic checking of model adequacy.
* Forecasting.

