Module 1

Introduction:

1. Time Series Analysis Types:

1. **Classification:** Identifies and assigns categories to the data.
2. **Curve fitting:** Plots the data along a curve to study the relationships of variables within the data.
3. **Descriptive analysis:** Identifies patterns in time series data, like trends, cycles, or seasonal variation.
4. **Explanative analysis:** Attempts to understand the data and the relationships within it, as well as cause and effect.
5. **Exploratory analysis:** Highlights the main characteristics of the time series data, usually in a visual format.
6. **Forecasting:** Predicts future data. This type is based on historical trends. It uses the historical data as a model for future data, predicting scenarios that could happen along future plot points.
7. **Intervention analysis:** Studies how an event can change the data.
8. **Segmentation:** Splits the data into segments to show the underlying properties of the source information.

2. Stationarity vs Non-Stationarity:

In the context of time series analysis and statistics, the terms "stationary" and "non-stationary" are used to **describe the behavior of a series over time**.

A stationary time series is one **whose statistical properties, such as mean, variance, and autocovariance, remain constant over time**. In other words, the distribution of data points in a stationary series does not change with time. Stationary series are often **easier to analyze** and model because their properties are consistent.

On the other hand, a non-stationary time series is one that **exhibits trends, seasonality, or other systematic patterns that change over time**. The statistical properties of a non-stationary series vary with time, making it more challe**nging to analyze and model accurately**. Non-stationary series may show long-term trends, cyclical patterns, or irregular fluctuations.

The distinction between stationary and non-stationary series is important because many statistical techniques and models assume stationarity to provide reliable results. When dealing with non-stationary series, it is often necessary to apply transformations or differencing operations to make the series stationary before applying such models.

To determine whether a time series is stationary or non-stationary, statisticians often look at various statistical tests or visual inspection. Common tests include the Augmented Dickey-Fuller (ADF) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, which examine the presence of unit roots or trend-stationarity in the data.

It's worth noting that the concept of stationarity can vary depending on the context. For example, in the field of spatial statistics, stationarity refers to the assumption that statistical properties of a process do not vary across different locations in space.

In summary, stationary time series have constant statistical properties over time, while non-stationary time series exhibit changing properties.

3. Seasonality:

Seasonality in time series refers to a pattern or fluctuation that repeats at regular intervals within a fixed time frame. It is characterized by the presence of regular and predictable variations in the data, which occur within specific time periods, such as hours, days, weeks, months, or years. Seasonality can be observed in various domains, including economics, finance, weather forecasting, and sales forecasting.

The presence of seasonality in a time series can have significant implications for data analysis and forecasting. Understanding and accounting for seasonality is crucial for accurately analyzing trends, making forecasts, and identifying any underlying patterns or effects. Seasonal patterns can provide valuable insights into the behavior of the data and can help in making informed decisions.

There are generally two types of seasonality:

1. Additive Seasonality: In this type, the seasonal pattern is consistent throughout the time series, and the magnitude of the seasonal effect remains relatively constant over time. The seasonal component is added to the trend and error terms. For example, in retail sales, there might be an increase in sales during the holiday season each year.
2. Multiplicative Seasonality: In this type, the seasonal pattern is not consistent throughout the time series, and the magnitude of the seasonal effect varies with the level of the time series. The seasonal component is multiplied by the trend and error terms. For example, in tourism, the seasonal variation might be more pronounced during peak vacation months.

To detect and analyze seasonality in a time series, several methods can be employed, such as:

1. Visual Inspection: Plotting the time series data and visually examining the pattern for any regular and repetitive fluctuations.
2. Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF): These statistical tools can help identify the presence of seasonality by examining the correlation between the time series observations and their lagged values.
3. Decomposition: Time series decomposition techniques, such as the additive or multiplicative decomposition, can separate a time series into its trend, seasonal, and residual components. This can help identify the presence and characteristics of seasonality.

Once seasonality is identified, appropriate modeling techniques can be applied to capture and account for it, such as seasonal autoregressive integrated moving average (SARIMA) models or seasonal exponential smoothing methods. These models can incorporate the seasonal component to improve forecasting accuracy and make more reliable predictions.

It's important to note that seasonality is just one aspect of analyzing and forecasting time series data. Other factors like trends, irregular variations, and external influences should also be considered to build robust models and obtain accurate predictions.

4. **Stationary stochastic process:**

A stationary stochastic process is a type of time series where the statistical properties of the process remain constant over time. This means that the mean, variance, and autocovariance structure of the process do not depend on the specific time point within the series.

Mathematically, a stationary stochastic process can be defined as follows:

Strict Stationarity: A stochastic process $\{X_t\}$ is strictly stationary if, for any set of time indices $\{t_1, t_2, ..., t_n\}$ and any integer k, the joint distribution of $\{X_{t_1}, X_{t_2}, ..., X_{t_n}\}$ is the same as the joint distribution of $\{X_{t_1+k}, X_{t_2+k}, ..., X_{t_n+k}\}$. In other words, the joint distribution of the process remains invariant under time shifts.

Weak Stationarity (Second-Order Stationarity): A stochastic process $\{X_t\}$ is weakly stationary if the mean, variance, and autocovariance of the process remain constant over time. Mathematically, it can be defined as:

Mean Stationarity: $E[X_t] = \mu$ for all t, where $E[X_t]$ represents the expected value of X_t and μ is a constant.

Variance Stationarity: $Var[X_t] = \sigma^2$ for all t, where $Var[X_t]$ represents the variance of X_t and σ^2 is a constant.

Autocovariance Stationarity: $Cov[X_t, X_{t+h}] = Cov[X_s, X_{s+h}]$ for all t, s, and h, where $Cov[X_t, X_{t+h}]$ represents the autocovariance between X_t and X_{t+h} and $Cov[X_s, X_{s+h}]$ represents the autocovariance between X_s and X_{s+h}. This means that the autocovariance function is only a function of the time lag h and not the specific time points t or s.

The autocovariance function of a stationary stochastic process is often denoted as $\gamma(h)$ or $\gamma_X(h)$, where h represents the time lag. It can be calculated as:

$$\gamma(h) = Cov[X_t, X_{t+h}] = E[(X_t - \mu)(X_{t+h} - \mu)]$$

where $Cov[X_t, X_{t+h}]$ is the covariance between X_t and X_{t+h}, and E[.] represents the expected value.

In summary, a stationary stochastic process is characterized by constant mean, variance, and autocovariance over time. It is an important assumption in time series analysis, as many analytical techniques and models are based on the assumption of stationarity.

A stationary stochastic process refers to a random process where the statistical properties do not change over time. Here's an example of a stationary stochastic process in day-to-day life:

Example of Stationary stochastic process:

Let's consider the daily temperature recordings in a city. Suppose we have a dataset that contains the daily temperature measurements taken at the same time every day over several years. In this case, we can assume that the temperature process is stationary if the statistical properties of the temperature (mean, variance, correlation, etc.) remain constant over time.

In this scenario, we can observe the following characteristics of a stationary stochastic process:

Mean Temperature: The mean temperature remains relatively constant throughout the year. Even though the daily temperature fluctuates, the long-term average remains the same.

Variance: The temperature variability remains constant over time. While there may be fluctuations from day to day, the overall spread of the temperature measurements remains consistent.

Autocorrelation: The correlation between temperature measurements taken on different days remains constant. For example, the correlation between today's temperature and tomorrow's temperature is the same as the correlation between today's temperature and the temperature two weeks from now.

Probability Distribution: The probability distribution of temperature values remains unchanged. For instance, the likelihood of a specific temperature range occurring in a given day remains constant over time.

By assuming the daily temperature process satisfies these characteristics, we can model it as a stationary stochastic process. This allows us to apply various statistical techniques for analysis, prediction, and understanding of temperature patterns in the city.

5. Correlogram

An autocorrelation plot, also known as a correlogram, is a graphical tool used to visualize the correlation between a time series and its lagged values. It helps identify any repeating patterns or relationships within the data over time. The autocorrelation plot displays the correlation coefficient on the vertical axis and the lag (or time shift) on the horizontal axis.

In an autocorrelation plot, each point represents the correlation between the time series and its lagged version at a specific lag. The lag represents the time shift between the observations. A positive correlation coefficient indicates a positive relationship between the time series and its lagged values, while a negative correlation coefficient indicates a negative relationship.

The autocorrelation plot is commonly used in time series analysis to identify the presence of any significant autocorrelation in the data. Autocorrelation can be indicative of underlying patterns, such as trends or seasonality, which are important to consider when analyzing and forecasting time series data.

The autocorrelation plot typically includes confidence bands to indicate the significance of the correlations. If a correlation point falls outside the confidence bands, it suggests that the correlation is statistically significant.

The interpretation of an autocorrelation plot involves examining the correlation coefficients at different lags. If the autocorrelation coefficients decay quickly to zero as the lag increases, it indicates that there is little or no autocorrelation in the data. On the other hand, if the autocorrelation coefficients remain high or significant at certain lags, it suggests the presence of autocorrelation and potential patterns in the data.

Autocorrelation plots can be created using various statistical software packages, such as Python's statsmodels library or R's acf function. These tools calculate the autocorrelation coefficients and generate the plot based on the provided time series data.

Overall, an autocorrelation plot is a helpful visualization tool to explore and understand the autocorrelation structure of a time series, enabling analysts to make informed decisions about modeling, forecasting, and understanding the underlying patterns in the data.

Stationary Time Series:

Formal definition of time series:

A time series is a sequence of data points or observations collected or recorded at successive and equally spaced time intervals. It represents the values of a variable or a set of variables over a specific time period. Each data point in a time series is associated with a particular time stamp, indicating when the observation was made.

Formally, a time series can be defined as a set of ordered pairs $\{(t_1, x_1), (t_2, x_2), ..., (t_n, x_n)\}$, where:

$t_1, t_2, ..., t_n$ represent the time stamps at which the observations are made, and they are typically arranged in a chronological order.

$x_1, x_2, ..., x_n$ are the corresponding values or measurements of the variable(s) of interest at each time stamp.

The time series can be univariate, meaning it involves only a single variable, or multivariate, involving multiple variables observed simultaneously at each time point. Time series data is commonly encountered in various fields such as finance, economics, weather forecasting, stock market analysis, and many others.

The analysis and modeling of time series data involve studying its statistical properties, identifying patterns, detecting trends or seasonality, and making predictions or forecasts based on historical observations.

6. **Sample Mean and Standard Error:**

In time series analysis, the sample mean and standard error are measures used to describe and analyze the central tendency and variability of a dataset. However, it's important to note that these measures are typically applied to cross-sectional data within a time series, rather than the entire time series itself.

Sample Mean:

The sample mean, also known as the average, is calculated by summing all the values in the dataset and dividing the sum by the total number of observations. It provides an estimate of the central tendency of the data.

Standard Error:

The standard error measures the variability or dispersion of the sample mean. It quantifies the uncertainty associated with estimating the population mean based on a finite sample. In the context of time series analysis, the standard error can be computed using various methods depending on the assumptions and characteristics of the data

One common approach to estimating the standard error in time series analysis is the "standard error of the mean" formula, which assumes independent and identically distributed (iid.) observations. This formula calculates the standard error as the standard deviation of the data divided by the square root of the number of observations:

Standard Error = Standard Deviation / √(Number of Observations)

It's important to note that time series data often exhibits autocorrelation, meaning that observations are not independent. In such cases, standard errors computed using the above formula may be biased or inconsistent. To account for autocorrelation, specialized methods such as autoregressive integrated moving average (ARIMA) models or generalized autoregressive conditional heteroskedasticity (GARCH) models are often used to estimate standard errors in time series analysis.

Additionally, when dealing with time series data, other measures like autocorrelation, partial autocorrelation, and the estimation of more complex models may be necessary to capture the dynamics and patterns present in the data.

7. Stationary processes:

In time series analysis, a stationary process refers to a stochastic process whose statistical properties do not change over time. It is an important concept because many time series analysis techniques and models assume or require stationarity for accurate analysis and prediction.

A stationary process has the following characteristics:

Constant mean: The mean of the process remains constant over time. In other words, the process is not affected by trends or systematic changes in the average value.

Constant variance: The variance of the process remains constant over time. It implies that the variability of the process does not change systematically with time.

Constant autocovariance: The autocovariance between any two observations of the process only depends on the time lag between them and not on the specific points in time. This property is also referred to as covariance stationarity.

No seasonality or periodic patterns: Stationary processes do not exhibit seasonal or periodic patterns that repeat over fixed intervals.

It's important to note that there are different types of stationarity:

Strict stationarity: A process is strictly stationary if the joint distribution of any set of observations is invariant under time shifts. It implies that all statistical properties, including moments and correlations, remain constant over time.

Weak stationarity: A process is weakly stationary (or second-order stationary) if the mean, variance, and autocovariance structure are constant over time, but the joint distribution of observations may not be invariant under time shifts. Weak stationarity is often the practical assumption used in time series analysis.

Why is stationarity important in time series analysis? Stationarity allows us to make certain assumptions and simplifications when analyzing time series data. Many statistical techniques and models, such as autoregressive integrated moving average (ARIMA) models, assume or require stationarity to estimate model parameters, make forecasts, and perform hypothesis testing. Stationarity also simplifies the interpretation of statistical properties, such as autocorrelation and partial autocorrelation functions.

If a time series is found to be non-stationary, it can be transformed to achieve stationarity. Common techniques for achieving stationarity include taking differences (differencing) to remove trends or applying transformations, such as logarithmic or Box-Cox transformations, to stabilize the variance.

In summary, stationarity is a fundamental concept in time series analysis. It allows for the application of various techniques and models that assume constant statistical properties over time, enabling accurate analysis, forecasting, and interpretation of time series data.

Types of Stationarity in time series.

In time series analysis, stationarity refers to a property of a time series where the statistical properties such as the mean, variance, and autocovariance do not change over time. Stationary time series are often easier to analyze and model than non-stationary time series. There are different types of stationarity that can be considered when studying time series data. Here are three common types:

Strict Stationarity: A time series is said to be strictly stationary if the joint distribution of any set of time points in the series is the same as the joint distribution of that same set of time points shifted by any fixed amount. In other words, the statistical properties of the time series do not change regardless of where we are in time. This implies that the mean, variance, and covariance structure remain constant over time.

Weak Stationarity (also known as Covariance Stationarity): Weak stationarity is a less restrictive form of stationarity compared to strict stationarity. A time series is considered weakly stationary if it’s mean, variance, and autocovariance structure remain constant over time. The mean is constant and does not depend on time, the variance is constant and does not depend on time, and the autocovariance between any two time points only depends on the time lag between them. This means that the statistical properties of the time series do not change on average over time, but there may be small fluctuations around this average behavior.

Trend Stationarity: Trend stationarity refers to a type of stationarity where the mean of the time series is constant over time, but the variance and autocovariance may vary. In other words, the time series may exhibit a trend or a systematic change in its mean value, but the statistical properties around this trend remain constant. Trend stationarity allows for deterministic trends, such as linear or polynomial trends, but does not allow for stochastic trends where the mean is influenced by random shocks or drifts.

It's worth noting that identifying the type of stationarity in a time series is important for choosing appropriate modeling techniques and interpreting the results accurately.

8. Statistical inference for time series:

Statistical inference for time series data involves making inferences and drawing conclusions about the underlying processes and parameters based on observed time series data. It allows us to analyze and interpret patterns, trends, and relationships within the data.

Here are some common methods used in statistical inference for time series:

Autocorrelation and Partial Autocorrelation: Autocorrelation measures the correlation between a time series and its lagged values. It helps identify the presence of temporal dependencies in the data. Partial autocorrelation measures the correlation between a time series and its lagged values, excluding the intermediate lags. These measures are used to determine the appropriate order of autoregressive (AR) and moving average (MA) components in autoregressive integrated moving average (ARIMA) models.

ARIMA Modeling: Autoregressive Integrated Moving Average (ARIMA) models are widely used for time series forecasting and inference. ARIMA models assume that the time series can be represented as a combination of autoregressive (AR), differencing (I), and moving average (MA) components. The parameters of the ARIMA model can be estimated using methods such as maximum likelihood estimation (MLE) or least squares estimation (LSE).

Stationarity Testing: Stationarity is an important assumption in time series analysis. It implies that the statistical properties of the time series, such as mean, variance, and autocorrelation, do not change over time. Various statistical tests, such as the Augmented Dickey-Fuller (ADF) test or the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, can be used to test for stationarity.

Model Selection and Diagnostic Checking: When fitting a time series model, it is important to select the appropriate model order and assess the goodness-of-fit. Techniques like information criteria (e.g., AIC, BIC) can be used to compare different models and select the best one. Diagnostic checks, such as residual analysis and model validation, help assess the adequacy of the chosen model.

Bayesian Inference: Bayesian methods can be used for time series analysis to estimate the posterior distribution of the parameters given the observed data. Markov Chain Monte Carlo (MCMC) algorithms, such as Gibbs sampling or the Metropolis-Hastings algorithm, can be employed to draw samples from the posterior distribution.

Bootstrap Methods: Bootstrap resampling techniques can be used to estimate the sampling distribution of a statistic or forecast accuracy measures for time series data. It involves randomly resampling the observed data with replacement to generate multiple bootstrap samples, from which the desired inference can be obtained.

These are just a few examples of statistical inference methods for time series analysis. The choice of method depends on the specific characteristics of the data and the research objectives. It is important to consider the assumptions and limitations of each method and select the most appropriate approach for the particular time series being analyzed.