1 Arima Model

The ARIMA (AutoRegressive Integrated Moving Average) model is a powerful statistical method used for time series forecasting and analysis. It combines three key components: autoregression (AR), differencing (I), and moving average (MA). Each component plays a specific role in capturing the patterns and trends present in the time series data.

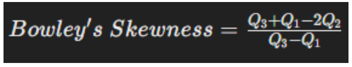
- **Autoregressive (AR) Component:** The autoregressive component captures the relationship between the current observation and its past values. It models the dependency of the current value on its own lagged values. The AR(p) component is represented by the formula:
- *Yt*=*c*+*ϕ*1*Yt*−1+*ϕ*2*Yt*−2+…+*ϕpYt*−*p*+*ϵtYt*=*c*+*ϕ*1*Yt*−1+*ϕ*2*Yt*−2+…+*ϕpYt*−*p*+*ϵt*
- **Integrated (I) Component:** The integrated component involves differencing the time series data to make it stationary. Stationarity means that the statistical properties of the data (such as mean and variance) remain constant over time. The differencing operator ∇∇ is applied to remove trends and seasonality, and the order of differencing is denoted by *dd*. The integrated component is represented by *Yt*′=*Yt*−*Yt*−1*Yt*′=*Yt*−*Yt*−1.
- **Moving Average (MA) Component:** The moving average component captures the relationship between the current observation and past forecast errors. It models the short-term fluctuations or noise in the data. The MA(q) component is represented by the formula:

The general form of an ARIMA model is ARIMA(p, d, q), where:

- *pp* is the order of the autoregressive component
- *dd* is the order of differencing
- *qq* is the order of the moving average component

2. Boweys Skewness

Bowley's skewness is a measure of skewness or asymmetry in a dataset. Unlike other measures of skewness like Pearson's skewness coefficient, Bowley's skewness uses quartiles instead of the mean and standard deviation, making it less sensitive to extreme outliers.
The formula for Bowley's skewness is given by:



$$Bowley's\ Skewness = \frac{Q_3+Q_1-2Q_2}{Q_3-Q_1}$$

Where:

- *Q*1*Q*1 is the first quartile (25th percentile)
- *Q*2*Q*2 is the second quartile or median (50th percentile)
- *Q*3*Q*3 is the third quartile (75th percentile)

Bowley's skewness formula essentially measures how far the median (*Q*2*Q*2) is from the average of the lower quartile (*Q*1*Q*1) and upper quartile (*Q*3*Q*3), relative to the interquartile range (*Q*3−*Q*1*Q*3−*Q*1). If the distribution is symmetric, Bowley's skewness will be close to zero; positive values indicate right skewness (longer tail on the right), while negative values indicate left skewness (longer tail on the left).

3. Multivariate Data Analysis

Multivariate data analysis (MDA) deals with datasets that involve multiple variables or features. The goal of MDA is to explore relationships, patterns, and dependencies among these variables. This field encompasses various statistical techniques and methods for analyzing complex datasets.

Techniques include:

1. **Principal Component Analysis (PCA):** Reduces dimensions while preserving data variance.
   a. Formula: *PC*1=*a*11*X*1+*a*12*X*2+…+*a*1*pXpPC*1=*a*11*X*1+*a*12*X*2+…+*a*1*pXp*
2. **Factor Analysis:** Identifies underlying factors explaining data patterns.

   a. Formula: Variables = Factors + Error

3. **Cluster Analysis:** Groups similar data points into clusters based on features.
   a. Formula: Distance metrics + Clustering algorithms
4. **Discriminant Analysis:** Predicts group membership based on predictor variables.
   a. Formula: Linear discriminant functions

MDA aims to uncover patterns, reduce complexity, and facilitate insights from multidimensional datasets.

4. Normal Distribution

The normal distribution, also known as the Gaussian distribution, is characterized by a symmetric, bell-shaped curve. Its formula is *f*(*x*)=12*πσ*⋅*e*−(*x*−*μ*)22*σ*2*f*(*x*)=2*πσ*1⋅*e*−2*σ*2(*x*−*μ*)2, where *μμ* is the mean and *σ*2*σ*2 is the variance.

Key points:

- Symmetric around the mean *μμ*.
- Follows the 68-95-99.7 rule: about 68% of data falls within 1 standard deviation of the mean, 95% within 2 standard deviations, and 99.7% within 3 standard deviations.
- Central Limit Theorem states that the distribution of sample means from any population approaches a normal distribution as sample size increases.

The normal distribution is fundamental in probability theory, statistics, and many scientific fields due to its mathematical properties and practical applications**.**

Certainly, here are some additional key points about the normal distribution:

- Standard Normal Distribution: A special case with mean *μ*=0*μ*=0 and standard deviation *σ*=1*σ*=1, often denoted as *ZZ*. It simplifies calculations and is used to calculate z-scores.
- Z-Score (Standard Score): A measure of how many standard deviations an individual data point is from the mean in a standard normal distribution. It's calculated as *Z*=*x*−*μσZ*=*σx*−*μ*, where *xx* is the data point, *μμ* is the mean, and *σσ* is the standard deviation.
- Area Under the Curve: The area under the normal curve represents probabilities. The total area under the curve is 1 (100%). Specific areas correspond to probabilities of events occurring within certain ranges of values.

5. Outlier

Outlier detection, also known as outlier identification or anomaly detection, is a process used in data analysis to identify observations that significantly deviate from the rest of the dataset. These outliers can represent rare events, measurement errors, or genuine anomalies that warrant further investigation.

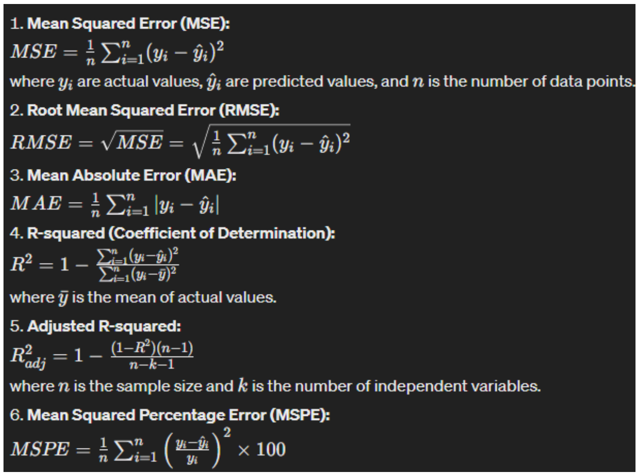
Methods include:

1. **Statistical Methods:** Z-score, modified z-score, and interquartile range (IQR).
2. **Distance-Based Methods:** Euclidean distance and Mahalanobis distance.
3. **Density-Based Methods:** DBSCAN and Local Outlier Factor (LOF).
4. **Model-Based Methods:** Regression and clustering models.
5. **Machine Learning Approaches:** Isolation Forest and One-Class SVM.

Considerations include data characteristics, threshold selection, and interpretation based on domain knowledge. Outlier detection is important for data quality assessment, anomaly detection, and model reliability.

6. performance evaluation

Regression performance evaluation involves assessing the accuracy and reliability of regression models in predicting outcomes. Several metrics are commonly used for regression performance evaluation:



1. **Mean Squared Error (MSE):**

$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

where y_i are actual values, $\hat{y}_i$ are predicted values, and n is the number of data points.

2. **Root Mean Squared Error (RMSE):**

$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

3. **Mean Absolute Error (MAE):**

$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$

4. **R-squared (Coefficient of Determination):**

$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$

where $\bar{y}$ is the mean of actual values.

5. **Adjusted R-squared:**

$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$

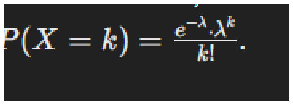
where n is the sample size and k is the number of independent variables.

6. **Mean Squared Percentage Error (MSPE):**

$MSPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\hat{y}_i}{y_i}\right)^2 \times 100$

7. poison distribution

The Poisson distribution is a discrete probability distribution used to model the number of events occurring in a fixed interval of time or space, assuming events occur independently at a constant average rate λλ. It is named after French mathematician Siméon Denis Poisson. Where:



$P(X = k) = \frac{e^{-\lambda}.\lambda^k}{k!}.$

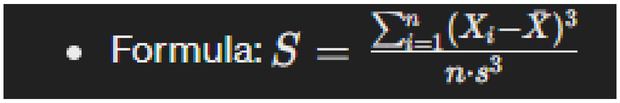
- *k* is the number of events or occurrences.
- *e* is Euler's number (approximately 2.71828).
- λ is the average rate of occurrence.

Certainly, here are some additional key points about the Poisson distribution:

- Interpretation: The Poisson distribution answers questions like "How many times will an event occur in a given time period, given its average rate of occurrence?"
- Assumptions: It assumes events occur independently and at a constant average rate throughout the interval.
- Parameter λλ: Represents the average rate of occurrence of events per unit of time or space.
- Mean and Variance: Both the mean and variance of the Poisson distribution are equal to λλ, making it a parameter for the distribution.
- Applications: Widely used in real-world applications such as modeling the number of arrivals at a service center, the number of accidents in a given period, or the number of customers in a queue.

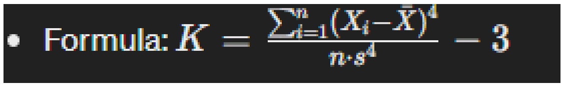
8. skewness andkurtosis

- Skewness: It indicates the direction and degree of asymmetry in a distribution. Positive skewness (right-skewed) means the tail is longer on the right side, and negative skewness (left-skewed) means the tail is longer on the left side. A skewness value close to 0 indicates symmetric distribution.
  - Positive skew: Mean > Median > Mode
  - Negative skew: Mode > Median > Mean



- Formula: $S = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^3}{n \cdot s^3}$

- Kurtosis: It measures the peakedness or flatness of a distribution's central peak relative to a normal distribution. Kurtosis values higher than 3 indicate heavy tails (leptokurtic), while values lower than 3 indicate light tails (platykurtic).
  - Leptokurtic (high kurtosis): More extreme values, sharper peak
  - Platykurtic (low kurtosis): Less extreme values, flatter peak



- Formula: $K = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^4}{n \cdot s^4} - 3$

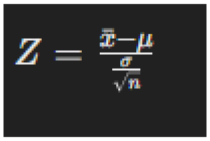
9. Univariate Data Analysis

Univariate data analysis focuses on analyzing and summarizing data from a single variable. It involves exploring the distribution, central tendency, dispersion, and shape of the data. Here's a theoretical explanation and some key formulas used in univariate data analysis:

- **Descriptive Statistics:** Summarize and describe the data using measures like mean, median, mode, variance, standard deviation, range, and quartiles.
- **Visualization Techniques:** Use histograms, box plots, and line graphs to visually represent the distribution and characteristics of the data.
- **Outliers Detection:** Identify and handle outliers, which are data points significantly different from the rest of the dataset and can impact statistical analyses.
- **Normality Testing:** Assess if the data follows a normal distribution using statistical tests like the Shapiro-Wilk test or graphical methods like Q-Q plots.
- **Skewness and Kurtosis Analysis:** Evaluate the skewness (asymmetry) and kurtosis (peakedness) of the data distribution, providing insights into its shape.
- **Probability Distributions:** Understand and apply different probability distributions (e.g., normal, binomial, Poisson) based on the characteristics of the data and the phenomenon being studied.
- **Data Cleaning and Preprocessing:** Handle missing values, transform data if needed (e.g., log transformation for skewed data), and ensure data quality before further analysis.

10. Ztest

The Z-test is a statistical hypothesis test used to determine whether the mean of a sample is significantly different from a known population mean when the population standard deviation is known. It is commonly used when dealing with large sample sizes and normally distributed data.



$$Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

- Purpose: To determine if a sample mean is significantly different from a known population mean when the population standard deviation is known.
- Assumptions: Random sampling, large sample size ($n \geq 30 n \geq 30$), and known population standard deviation ($\sigma\sigma$).
- Critical Values: Compare the calculated Z-test statistic to critical Z-values from a Z-table or use statistical software.
- Decision Rule: If $|Z||Z|$ exceeds the critical value or if the p-value is less than the chosen significance level (e.g., 0.05), reject the null hypothesis.
- Applications: Used in quality control, medical research, and business to make inferences about population means based on sample data.