Business Analytics

Glossary of Statistical Terms, Predictive Analytics and Data Visualization

Gerry Skews

Revision 2.0 Jan 2025

Contents

Glossary of Key Statistical Terms & Techniques

Mean (Average):

The sum of all values divided by the number of values. It gives a central point of the data distribution.

Median:

The middle value in a dataset when the numbers are arranged in order. It helps represent the centre of skewed distributions.

Mode:

The value that appears most frequently in a dataset. It is useful for categorical data.

Data Mining:

Data mining is the process of discovering patterns, trends, and useful information from large datasets using statistical, mathematical, and machine learning techniques.

Root Cause Analysis:

Root Cause Analysis (RCA) is a systematic process used to identify the underlying causes of problems or failures to prevent recurrence.

Drill Down Statistics:

A Drill Down Statistical Process is an analytical method that breaks down complex data into finer, more detailed levels to uncover specific patterns or insights.

Standard Deviation:

A measure of the amount of variation or dispersion in a dataset. A low standard deviation means the data points tend to be close to the mean, while a high standard deviation means they are spread out.

Variance:

The square of the standard deviation, it measures how much the data points differ from the mean.

Cross Tabulation:

Cross tabulation is a method used to analyse the relationship between two or more categorical variables by organising the data into a table (called a cross-tab or contingency table) that displays the frequency distribution of variables, allowing for comparison and identifying patterns or correlations between them.

Correlation:

A statistical measure that describes the degree to which two variables move in relation to each other. It ranges from -1 to +1, where +1 means perfect positive correlation and -1 means perfect negative correlation.

Regression Analysis:

A statistical method to determine the relationship between a dependent variable and one or more independent variables. Common forms include linear and logistic regression.

Multiple Regression:

An extension of regression analysis where two or more independent variables are used to predict the dependent variable.

ANOVA (Analysis of Variance):

A method used to compare the means of three or more samples to see if at least one is significantly different from the others.

Chi-Square Test:

A statistical test used to determine if a significant relationship exists between two categorical variables.

T-Test:

A test used to determine if there is a significant difference between the means of two groups, commonly used in comparing test scores, profits, etc.

Z-Test:

Similar to a t-test but used when the sample size is large, and the population variance is known.

P-Value:

A probability score that helps to determine the significance of your results in hypothesis testing. A low p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis.

Hypothesis Testing:

A method for testing a claim or hypothesis about a parameter in a population, using sample data.

Confidence Interval:

A range of values, derived from the sample data, that is likely to contain the true value of an unknown population parameter.

Sampling:

The process of selecting a subset of individuals from a population to estimate characteristics of the whole population.

Outliers:

Data points that are significantly different from other observations in the dataset. These can affect the results of an analysis.

Time Series Analysis:

A method used for analyzing data points collected or recorded at specific intervals over time to forecast future trends.

Bayesian Statistics:

A method of statistical inference that uses Bayes' theorem to update the probability for a hypothesis as more evidence becomes available.

Cluster Analysis:

A method used to group a set of objects in such a way that objects in the same group (or cluster) are more similar to each other than to those in other groups.

Factor Analysis:

A technique used to reduce the number of variables by identifying underlying factors that explain the pattern of correlations within the dataset.

Principal Component Analysis (PCA):

A dimensionality-reduction technique used to reduce the complexity of datasets by transforming them into a set of uncorrelated variables called principal components.

Monte Carlo Simulation

A computational algorithm that uses repeated random sampling to simulate and understand the behaviour of complex systems or processes.

Logistic Regression:

A regression model used for binary outcomes, often used in classification problems.

Kaplan-Meier Estimator:

A non-parametric statistic used to estimate the survival function from lifetime data, often used in medical research.

MANOVA (Multivariate Analysis of Variance):

An extension of ANOVA that allows for comparing more than one dependent variable across different groups.

Survival Analysis:

A branch of statistics that analyzes time-to-event data, such as the time until a product fails or the time until a patient relapses.

R-Squared (R^2):

A statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

Cross-Validation:

A technique for assessing how a predictive model performs by partitioning the data into subsets, training the model on one subset, and validating it on another.

Bootstrap Method:

A resampling technique used to estimate statistics on a population by sampling a dataset with replacement.

True Positive*:*

A true positive (TP) is an outcome in a data analysis or testing process, where a model **correctly identifies or predicts the presence of a specific condition** or event. In other words, it's when something that is actually true (e.g., a disease, a customer churn, or a fraud case) is accurately detected or classified as true by the system or model

False Negative:

A false negative (FN) is a "cock up" It is an outcome in a data analysis or testing process where a model or test **incorrectly fails to identify or predict the presence of a specific condition** or event. In other words, it's when something that is actually true (e.g., a disease, a customer churn, or a fraud case) is mistakenly classified or predicted as false by the system or model.

Glossary of Predictive Analytical Techniques

Definition of Predictive Analytical Techniques

Predictive analytical techniques involve the use of statistical methods, machine learning algorithms, and data modelling to analyze historical data and predict future outcomes or trends. These techniques focus on identifying patterns and relationships within datasets to forecast events, behaviours, or performance metrics.

Summary of Key Predictive Analytical Techniques

Linear Regression

Description: A statistical method used to model the relationship between a dependent variable and one or more independent variables.

Applications: Sales forecasting, price elasticity modelling.

Example: Predicting monthly sales based on advertising spend.

Logistic Regression

Description: Used for binary (Yes/No) classification problems to predict the probability of a categorical outcome.

Applications: Customer churn prediction, fraud detection.

Example: Predicting whether a customer will buy a product (yes/no).

Time Series Analysis

Description: Methods used to analyze data points collected or recorded at specific time intervals.

Applications: Demand forecasting, stock market analysis.

Example: Forecasting next quarter's revenue using past sales data.

Decision Trees

Description: A tree-like model of decisions and their possible consequences, including outcomes and resource costs.

Applications: Risk assessment, loan approval processes.

Example: Predicting customer eligibility for a credit card.

Random Forests

Description: An ensemble learning technique combining multiple decision trees for more accurate predictions.

Applications: Fraud detection, predictive maintenance.

Example: Predicting machine failure based on sensor data.

Support Vector Machines (SVM)

Description: A supervised learning method used for classification and regression by finding the best decision boundary between classes.

Applications: Text classification, image recognition.

Example: Classifying customer reviews as positive or negative.

Neural Networks

Description: Algorithms inspired by the human brain, capable of learning from large and complex datasets.

Applications: Demand forecasting, speech recognition, personalization.

Example: Predicting product demand in an e-commerce platform.

Clustering (e.g., K-Means)

Description: A technique for grouping similar data points into clusters without predefined labels.

Applications: Customer segmentation, market analysis.

Example: Identifying groups of customers with similar purchasing behaviours.

ARIMA (AutoRegressive Integrated Moving Average)

Description: A statistical model used for time series forecasting by capturing trends, seasonality, and noise in the data.

Applications: Financial forecasting, inventory management.

Example: Forecasting product demand for the holiday season.

Gradient Boosting Machines (GBM)

Description: An ensemble technique that builds predictive models iteratively to minimize errors.

Applications: Predictive marketing, credit scoring.

Example: Estimating the likelihood of a customer upgrading to a premium service.

Modern Applications of Predictive Analytics

Retail: Demand forecasting, inventory optimization, customer recommendation systems.

Healthcare: Predicting patient readmission rates, disease outbreak analysis.

Finance: Fraud detection, credit scoring, stock price prediction.

Manufacturing: Predictive maintenance, quality control.

Logistics: Route optimization, delivery time estimation.

Marketing: Campaign performance forecasting, customer lifetime value predi

Data Visualisation Techniques

We link to the Microsoft knowledge base in the following section, - other excellent products are available!

Pie Charts:

A pie chart is a circular graph divided into slices to illustrate numerical proportions, where each slice represents a category's contribution to the whole. Pie charts are commonly used to show percentage or proportional data.https://support.microsoft.com/search/results?query=Pie+Chart&isEnrichedQuery=false

Histograms:

A histogram is a graphical representation of the distribution of numerical data, often used to show the frequency of data points within specified ranges (or bins). It helps visualize how data is distributed over intervals, showing patterns like normal distribution or skewness.https://support.microsoft.com/search/results?query=Histogram&isEnrichedQuery=false

Data Dashboard:

A data dashboard is an interactive tool or display that consolidates and visualizes key performance indicators (KPIs), metrics, and data points in a single view, allowing users to monitor business processes and make data-driven decisions quickly and efficiently. It typically includes charts, graphs, and tables that update in real-time.https://support.microsoft.com/search/results?query=Dashboard&isEnrichedQuery=false

Radar Graph:

A radar graph (or spider chart) displays multivariate data on a two-dimensional chart, with each axis representing one variable. The data is plotted as points connected by lines, forming a polygon. It's commonly used to compare the performance of different categories across multiple variables.https://support.microsoft.com/search/results?query=Radar+Chart&isEnrichedQuery=false

Stock Graph:

A stock graph (or stock chart) visualises financial data, often showing a stock's price movement over time. These charts typically display data like opening, closing, high, and low prices in a single day, helping to analyse market trends and stock performance.https://support.microsoft.com/search/results?query=Stock+Chart&isEnrichedQuery=false

Surface Plot:

A surface plot is a three-dimensional chart that shows relationships between three continuous variables. It's often used to visualize how two independent variables affect a dependent variable, creating a 3D surface that helps identify peaks, valleys, and trends in data.

Cumulative Plot:

A cumulative plot, or cumulative frequency plot, is a graph that represents the cumulative sum or count of data points up to a certain point. It helps visualise the accumulation of data over time or across categories, showing growth trends or distributions.

Error Bars:

Error bars are graphical representations of the variability or uncertainty in data. They show the range of possible error or deviation from the measured value, often indicating confidence intervals, standard deviation, or standard error in a dataset.https://support.microsoft.com/search/results?query=Error+Bars&isEnrichedQuery=false

Regression Line:

A regression line is a straight line that best fits the data points on a scatter plot, showing the relationship between two variables. It is used in linear regression analysis to predict the value of a dependent variable based on the independent variable.

Polynomial Line:

A polynomial line is a curved line that represents a relationship between variables modelled by a polynomial equation. It can show more complex, non-linear trends in data, with the degree of the polynomial determining the curvature of the line.

Exponential Line:

An exponential line represents an exponential relationship between two variables, where one variable increases (or decreases) at a consistent rate relative to the other. The curve rises (or falls) steeply, often used to model growth or decay processes like population growth or radioactive decay.

Logarithmic Plot:

A logarithmic plot is a graph where one or both axes are scaled logarithmically, meaning the values increase by orders of magnitude rather than by equal increments. It is useful for visualising data that spans a wide range of values or for identifying multiplicative relationships.

Waterfall Charts:

Waterfall charts are a type of data visualisation that illustrates the cumulative effect of sequentially introduced positive or negative values. They are particularly useful for understanding how an initial value is affected by a series of intermediate values, ultimately leading to a final result. Waterfall charts are commonly used in financial

analysis to track revenues, expenses, or cash flows over time. for more info visit https://support.microsoft.com/search/results?query=Waterfall+Chart&isEnrichedQuery=false

Pareto Charts:

Pareto charts are a specialised type of bar chart that displays the relative frequency or impact of problems in descending order, combined with a cumulative line graph. Based on the Pareto principle (80/20 rule), these charts help identify the most significant factors contributing to an issue, allowing businesses to prioritise improvements effectively. They are often used in quality control and process improvement initiatives.https://support.microsoft.com/search/results?query=Pareto+Chart&isEnrichedQuery=false

Scatter Plots:

Scatter plots are graphical representations that display the relationship between two quantitative variables. Each point on the plot corresponds to an observation in the dataset, with one variable plotted on the x-axis and the other on the y-axis. Scatter plots help visualize correlations, trends, and patterns, making them valuable for analyzing data in fields such as business, science, and social research.https://support.microsoft.com/search/results?query=Scatter+Plot&isEnrichedQuery=false

Heat Maps:

Heat maps are a data visualisation technique that represents data values as colours in a two-dimensional space, allowing for easy identification of patterns and trends. The intensity of the colour indicates the magnitude of the data, making it effective for comparing different categories or time periods. Heat maps are commonly used in various fields, including marketing, operations, and finance, to analyse performance metrics, customer behavior, and more.https://support.microsoft.com/search/results?query=Heatmap&isEnrichedQuery=false

Pivot Tables:

Pivot tables are a data processing tool used in spreadsheet applications to summarise, analyse, and reorganise data. They allow users to dynamically arrange and manipulate large datasets by grouping data based on specific attributes, enabling the extraction of meaningful insights. Pivot tables are particularly useful for generating reports and performing complex calculations without altering the original dataset.https://support.microsoft.com/search/results?query=pivot+table&isEnrichedQuery=false

Tree Diagrams (e.g., Decision Trees):

Tree diagrams, including decision trees, are graphical representations used to illustrate decisions and their possible consequences, including chance event outcomes, resource

costs, and utility. In business, decision trees help visualise decision-making processes, making it easier to evaluate various options and their potential impacts. They are valuable for risk assessment and strategic planning, providing a structured approach to complex decision-making scenarios.https://support.microsoft.com/search/results?query=Tree+Diagram&isEnrichedQuery=false

Note, *Tree Diagrams are available in a number of software products, we link to Microsoft Visio as an example here.*

