There are three basic approaches to generating forecasts: regression-based methods,

And heuristic smoothing methods. and general time series models.

# Difference between Discrete Data and Continuous Data.

## Definition of discrete data.

Discrete data refers to a form of quantitative information characterized by countable figures and non-fractional values. Typically, discrete data is presented in the form of whole numbers that convey precise quantities. A common way to conceptualize discrete data is to preface it with "the number of," for instance, the number of patrons in a shop. This kind of data generally encapsulates distinct occurrences that are already in the past. In analysing discrete data, you can examine precise numbers, such as the quantity of products sold on a particular date or the duration of time an employee has worked in a given week.

## Definition of continuous data.

Continuous data is a quantitative data category that captures measurements that can be highly precise, extending to numerous decimal places as needed. It represents values that can be measured on a scale and can fall between any two amounts within a range. This data type is prevalent in sectors that demand exactness, such as healthcare, production, and research and development. Continuous data is dynamic, presenting the opportunity for organizations to scrutinize their processes and forecast upcoming patterns. An instance of its application could be a company monitoring the duration required by a team to fulfil assignments, providing insights into productivity and efficiency.

Top of Form

Bottom of Form

# Some common formulas and calculations used in Descriptive Statistics

## Sum of Squares

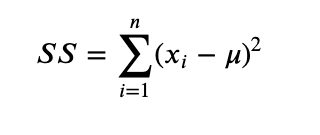
The sum of squares refers to the aggregated total of each data point's deviation from the mean, squared.

This calculation is a fundamental part of various statistical analyses, serving to quantify the variance within a dataset.

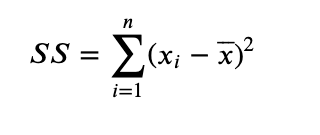
Top of Form

Bottom of Form

For a Population



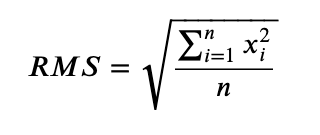
For a Sample



## 

## Root Mean Square

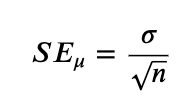
The root mean square (RMS) is a statistical measure that calculates the magnitude of a set of numbers. It is found by taking the square root of the average of the squares of the values in the set. This metric is especially useful in contexts where both positive and negative values in the dataset are treated equally, and it tends to give a higher value than the average due to the squaring of the values.



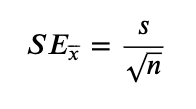
## Standard Error of the Mean

The standard error of the mean (SEM) is derived by dividing the standard deviation of the dataset by the square root of the number of observations (*n*). This metric indicates how much the sample mean is expected to vary from the true population mean.

For a Population



For a Sample



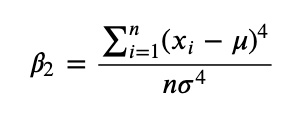
## Kurtosis

Kurtosis measures the "tailedness" of a distribution, indicating how outlier-prone a dataset is. High kurtosis suggests more extreme outliers than a normal distribution, while low kurtosis indicates fewer extreme outliers. This helps assess the extremity and concentration of tail data compared to a normal bell curve.

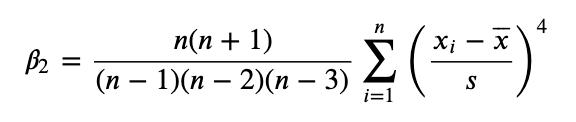
Top of Form

Bottom of Form

For a Population



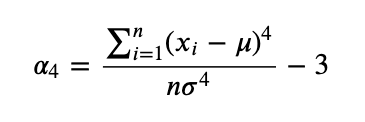
For a Sample



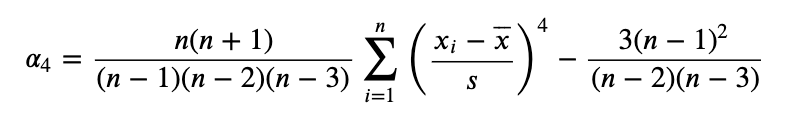
Kurtosis Excess

Excess kurtosis gauges the peak height of a distribution's tails, focusing on the concentration of outliers rather than their extremity. A distribution with high excess kurtosis indicates a significant presence of outlier data, pointing to more frequent extreme deviations from the mean compared to a normal distribution.

For a Population



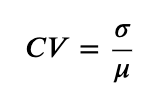
For a Sample #(This is just Kurtosis in MS Excel and Google Sheets)



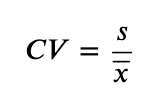
## Coefficient of Variation

The coefficient of variation (CV) measures the relative spread of data points around the mean, expressed as a ratio of the standard deviation to the mean. It's a useful statistic for comparing the degree of variability from one data series to another, even if the means are drastically different. The CV is calculated by dividing the standard deviation by the mean. This measure is particularly helpful in assessing the risk or variability in different contexts, such as finance and scientific research, where understanding relative dispersion is crucial.

For a Population



For a Sample



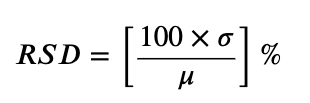
## Relative Standard Deviation

The relative standard deviation (RSD) quantifies the variation in a data set relative to its mean, presented as a percentage. It's computed by multiplying the standard deviation by 100 and then dividing by the mean. This statistic is valuable for comparing the variability of datasets with different units or means, providing a normalized measure of dispersion.

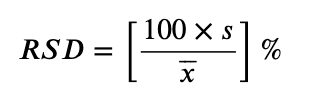
Top of Form

Bottom of Form

For a Population



For a Sample



## Frequency

Frequency measures how often each value appears in a dataset, essential for determining the mode, the value that occurs most frequently. This statistical concept helps in understanding the distribution and concentration of data points.

# Trend and Seasonality: In time series analysis, it's also important to consider trend and seasonality.

# These factors can significantly affect your central tendency measures.

# For example, a steadily increasing trend could make the mean over the entire series less representative of any specific point in time.

# Outliers: Especially with mean calculations, consider the impact of outliers.

# If your time series data includes extreme values, they could skew the mean.

# As can be seen in the described statistics, column Arrivals Air got close to 0

Kernel Density Estimate (KDE) Explained

#KDE is a non-parametric way to estimate the probability density function (PDF) of a random variable.

#It's smooth and not limited to a specific distribution shape (like normal or binomial distributions).

#KDE works by placing a kernel (a smooth, bell-shaped curve) on each data point and then summing all these kernels to produce a smooth estimate of the data's density function.

#The final curve is a weighted sum of those kernels and gives a smoothed estimate of the dataset's density.

#The bandwidth (bw) parameter controls the width of the kernels and thus the smoothness of the density estimate:

#a larger bandwidth leads to a smoother curve,

#while a smaller bandwidth leads to a curve that closely follows the data.

#Handling Multimodal\* Distributions

#For datasets with more than one mode,

#the KDE will show peaks at each mode, depending on the chosen bandwidth.

#A well-chosen bandwidth can reveal the multimodal nature of the data.

#The KDE does not calculate a normal distribution;

#rather, it estimates the data's density based on the existing data points.

#If the data is multimodal,

#the KDE will reflect those modes in its estimate.