# RANDOM VARIABLE –

* Variables whose values depend on the outcome of a random phenomenon.
* Random Phenomenon is something that occurs in real life.
  + Tips to a waiter
  + Total bill amount
  + Measurement of a species of flower
* Could be Numerical or Textual
* Could be discrete or continuous
* Discrete – Finite number of values or countable number of values. Categorical in nature
* Continuous – Infinite number of variables, uncountable .
* In Pandas the columns can be considered random variables.

# SAMPLE VS POPULATION

**Population** – Entire Data or the whole of the data for a particular random variable.

**Sample** – It is a subset selected from the entire population.

**Two branches of statistics evaluate the sample or the population.**

* Descriptive Statistics – It is a statistics of learning the characteristics of the sample drawn from the population. A quantitative summary of the features at hand.
* Inferential Statistics – Statistics of making inferences about the population based on the sample. Like Machine Learning.

**Measure of Descriptive Statistics**

* **Measures of Central tendency:** 
  + Mean – Arithmetic average.
    - Has a shortcoming that it gets affected by extreme values or outliers.
* Median – Middle value of the data arranged in order.
  + Not affected by extreme values or outliers.
* Mode – The most repeated values.
  + Mostly Used in discrete data set.
  + Can also be used in continuous but rarely.
  + Modes could be more than one value.
* Quartiles – Set of three data points that divide the data set into 4 equal parts.
* Median is the 2nd Quartile.
* The first or last quartile have 25% data more or less that the middle quartile respectively.
* Quantiles – Higher granularity of Quartiles. Divides the data set into equal distribution parts.
  + Eg – 1% of division of data would be equal to 100 quantiles.
* **Measures of Dispersion**
  + Range – Maximum minus Minimum of the data
    - Can be affected by extreme values or outliers.
* Interquartile range – Distance between the third quartile and first quartile.
  + Not affected by the outliers.
  + Variance – Most commonly used descriptive statistics. Average of the squared differences from the mean.
  + Standard Deviation – Squared root of the variance. Most used measured of dispersion.
* Central tendencies and Dispersions are used when we are trying to analyse a single variable.
* When trying to analyse multivariate or bivariate, correlation and covariance is used.
* Indicate the relationship between two or more variables. Whether they grow together or have an opposite relationship.
* **VIOLIN PLOT** 
  + **KDE Estimator –** Probability distribution function of the data that is currently present to us.
    - It is an estimation not the exact data.
    - Gives a good idea of the data spread and the theoretical distribution for the population.
* **BOX PLOT -** Gives, min and max range. The three quartiles and the outliers.
* **DISTPLOTS –** Gives the real data in sample in the form of histogram and the projection of the probability distribution function in the KDE. Can add central tendencies.
* x,y = kp.get\_lines()[0].get\_data() – Gets the object of the visualisation
* plt.fill\_between(x, y, 0, where=(condition), alpha) – Specify a y region between which you want to fill or shade one area under a given plot.
  + Where is a fill condition.

# MEASURES THAT TELL US ABOUT THE SHAPE OF THE DATA

* SKEW
* Kurtosis
* **SKEW :** Tells us about the asymmetry of the data. Could be positive or negative direction.
* Is given by (mean – mode) / std
* If the tail is towards right side, they are positively skewed data
* If the tail is towards left side, they are negatively skewed data
* SKEW Metric:
  + Df.col.skew() to check for skew
* Skewness can affect the outcomes when using machine learning on a certain dataset.
* Skew correction is to be applied before feeding the data into machine learning models.
* SKEW Correctio n:
  + Log Transform
* **KURTOSIS :**  Tells us about the tailedness or the peakedness of data
  + Tailedness means longer tails or more outliers and data
    - Normal Distribution of data – Mesokuritc
    - If there are more tails in the data , which could mean more peak in data – leptokurtic ( has positive values)
    - Data has very short tails or is very flat – platykurtic ( has negative values)
  + Kurtosis corrections are not commonly known in machine learning.
  + Knowing the shape of the data is probably an important action.

# PROBABILITY DISTRIBUTION

Gives us details about the populations from where the samples are being drawn.

* Classes of Probability distribution
  + Continuous
    - Uniform Distributions
    - Normal/Gaussian Distribution
  + Discrete
    - Poisson Distribution
    - Binomial Distribution
* Probability Distribution – It is a **statistical function** that describes **all possible values** and **probabilities/likelihoods** that a random variable takes (can be estimated from sample. )
* **Continuous Probability Distributions** : They are called probability density functions
  + They are produced from a continuous random variable.
  + The probability of getting a value in a range is given by the area under the curve.
    - Eg : Uniform distribution, Gaussian/ Normal distribution, Chi Squared, Beta distribution, Gamma distribution, Student’s distribution, Pareto distribution.
  + **Uniform Distribution :**  Has continuous distribution where each value in the event has a equal probability of occurrence.
    - It has a discrete variant
    - It takes continuous equivalent parameters, low and high values, where the distribution lies.
    - Example – Throw of dice, Choosing something from a deck of cards.
    - Histogram is an ideal plot to visualize this kind of data.
    - Distplot is the most desirable form of visualising continuous data.
  + **Normal/Gaussian distribution (Bell Curve) :** The parameters of this distribution are mean and standard deviation.
    - Examples - Most Natural Phenomena like i) The heights, weights or IQ of a certain section of population. ii) The marks a class gets in a certain . iii) The rainfall or temperature over a period of time. iv) The error in any calculation.
    - A common variant of the normal distribution is a standard normal distribution where mean is 0 and std is 1.
* **Discrete Probability Distributions :**  They are generated from discrete random variables.
  + They have a property distribution function called probability mass function.
  + The magnitude of the probability mass function at a point indicates the probability of a getting that value.
    - Eg: Bionomial, Beroulli, Hypergeometric, Logarithmic, Poisson
  + **Bionomial Distribution :** A discrete probability distribution of the number of successes in a sequence of n Bernoulli trials.
    - Has two perimeters.
      * n : number of trials
      * p : probability of success of any one trial.
    - Bernoulli or binomial trials are a random experiment that can produce only two possible outcomes. (Success or failure.)
      * Probability of success is p.
    - Examples – All those areas which get two possible outcomes success or failures, true or false.
      * Number of patients responding or not responding to a line of treatment.
      * Number of defective products in a factory assembly line.
      * Correct predictions by binary classification machine learning models.
    - n increases, peak gets smaller
  + **Poisson Distribution -** Gives the probability of a given number of events occurring in a fixed interval of time or space.
    - The events have to be independent.
    - The rate at which the events occur should be constant.
    - Parameter – Mean rate of events occurrences – lambda
    - Examples – Any place where event occurs at a constant rate.
      * Calls a call center gets for a particular period of time like an hour.
      * The number of mutations that a DNA strand takes per unit time.
      * How many bankruptcies are filed per month.
      * The page hits per minute on a website.
    - Lambda increases peak becomes flatter, tails become longer and lesser skewed.
* Approaches to generate probability distribution for visualising or working with them?
  + Take the formal definition of the probability and draw the distribution from it.
  + Draw random samples and see how the sample approximates the probability distribution as it grows.
  + Use np.random function to random samples out.

**Text

Description automatically generated**