**Chapter 2: PYTHON FUNDAMENTALS & STATISTICS**

**Topic – 1: Importance Of Python**

* **Interactivity** with library tools like **Plotly** & **Bokeh**.
* **Integration** with various tools.
* **Support for large datasets** with automatic memory management.

**Topic – 2: Pandas**

**Introduction**

* Built on top on **NumPy**.
* Provides two types of data structures to manipulate data, which are ***series*** & ***data frames***.

**Series**

* **One-dimensional** array containing axis labels called **indexes**.
* This label can hold **any** data type.
* **Series** is basically a **column** in **Excel** sheet.
* It can even be taken from a **SQL** database.

**Data Frame**

* **Two-dimensional** array of labels (**rows** & **columns**).
* It can also be taken from a **SQL** database.

**Topic – 3: NumPy**

**Introduction**

* **NumPy** stands for **Numerical Python**.
* It is used for **scientific computing** in Python.

**Types Of NumPy Arrays**

* **1D array –** Vectors
* **2D array –** Matrices

**Note!**

**🡪 Once a NumPy array is declared, we can’t change its size.**

**🡪 The data type of all elements in NumPy must be same.**

**Topic – 4: Introduction To Statistics**

**Types Of Statistics**

* Descriptive statistics
* Inferential statistics

**Descriptive Statistics**

* **Summarizes** a whole dataset.
* Measures **central tendency** (**mean**, **median**, **mode**).
* Also measures **dispersion** (**range**, **variance**, **standard deviation**).
* And also measures **shape** (***skewness***, ***kurtosis***).

**Inferential Statistics**

* It uses **sample data** from population to make **predictions** or draw **hypothesis**.

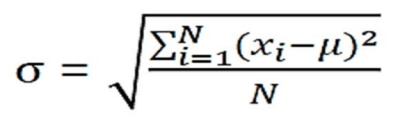
**Central Tendency & Dispersion**

* **Central tendency:** A **central value** that all the values in a dataset **tend** toward.
* It includes **mean**, **median** & **mode**.
* **Median:** **Middle value** in a sorted list of numbers.
* Use terms like ***'n'*** when defining these.

**Measure Of Dispersion**

* **Dispersion:** Tells how variable’s value varies.
* It includes **range**, **variance**, **standard deviation**.
* **Range:** Difference between **smallest** & **largest value** in a dataset.
* **Variance:** The **more** the **variance**, the **more** the **data are spread** out.

**Population Variance**

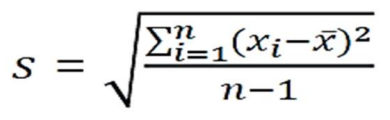
****

**xi = Value of ith item**

**N = Number of items in population**

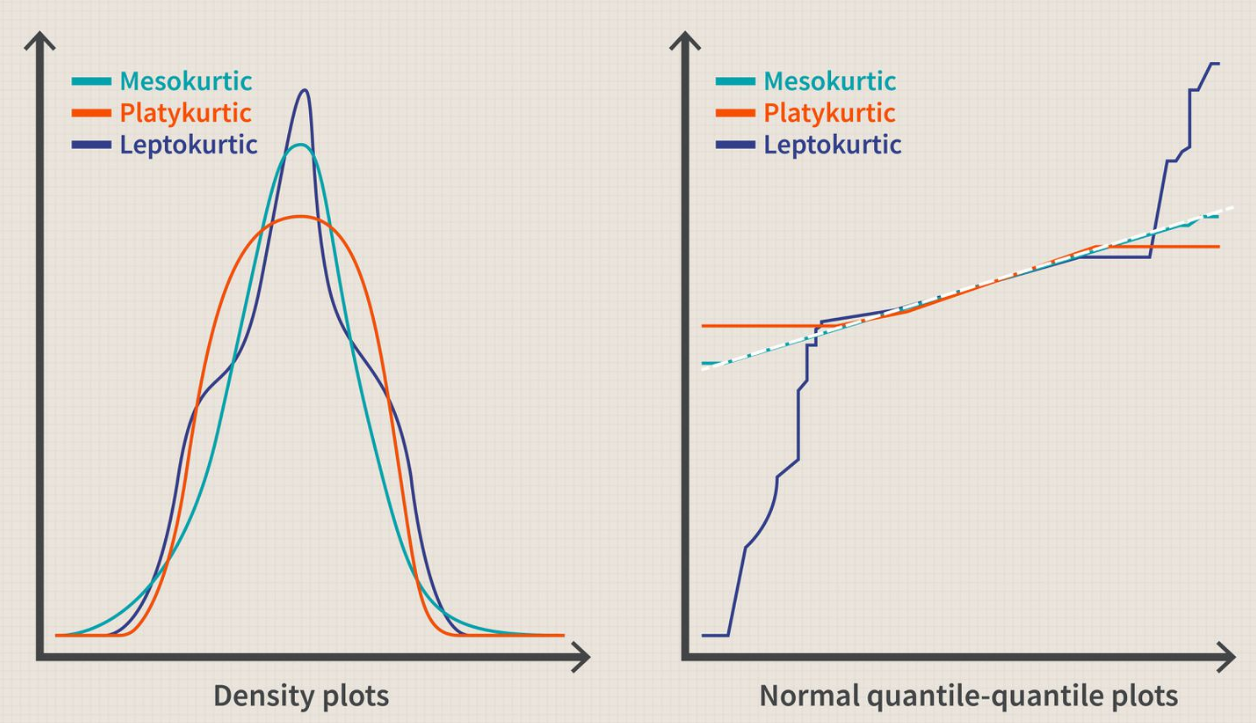
**u = Population mean**

**Sample Variance**



**Measure Of Shape**

* It described the **distribution/pattern** of data within a given dataset.
* **Skewness:** Tells us about **asymmetry** in data.
* **Positive skewness/ right-skewed:** Right side of the line is lifted more up, meaning the **mean** & **median** are more than the **mode**.
* **Kurtosis:** Tells us about **presence of outliers** as per the nature of tail.
* **Leptokurtic:** Heavy tails & sharp peaks.
* **Platykurtic:** Light tails & flatter peaks.
* **Mesokurtic:** Normal distribution.

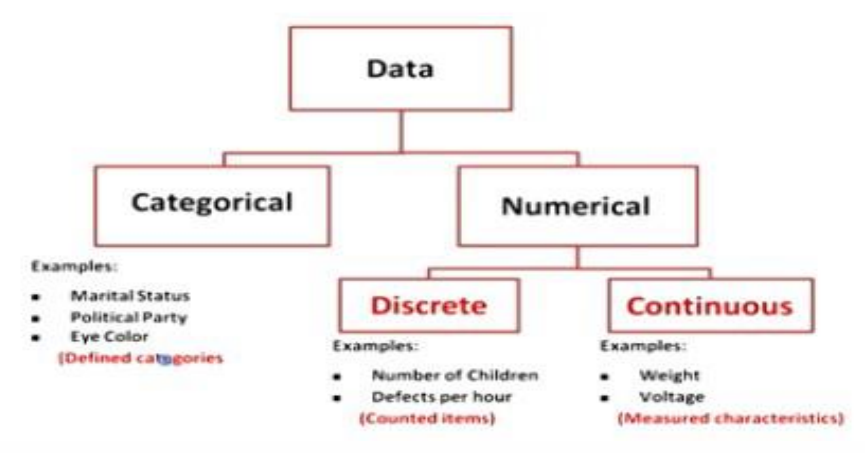


**Skewness v/s Kurtosis**

|  |  |
| --- | --- |
| **Skewness** | **Kurtosis** |
| **Measures asymmetry.** | **Measures tailedness.** |
| ***\*Positive skew definition\**** | ***\*Positive kurtosis definition\**** |
| **Zero skewness means the distribution is symmetric.** | **Zero kurtosis means the distribution is similar to normal distribution.** |
| **Tells the degree of asymmetry.** | **Tells the degree of outlier presence.** |
| **Used commonly at economy & finance related fields.** | **Used commonly at natural & formal sciences.** |
| **For example, income distribution in a country.** | **For example, particle physics & image processing etc.** |

**Topic – 5: Data Measurement**

**Types Of Variables**



**Levels Of Data Measurement**

* **Nominal data:** Categories without ranking.
* **Ordinal data:** Categories with ranking.
* **Interval data:** Measurement without any zero point.
* **Ratio data:** Measurement with true zero point.

**Topic – 6: Sampling**

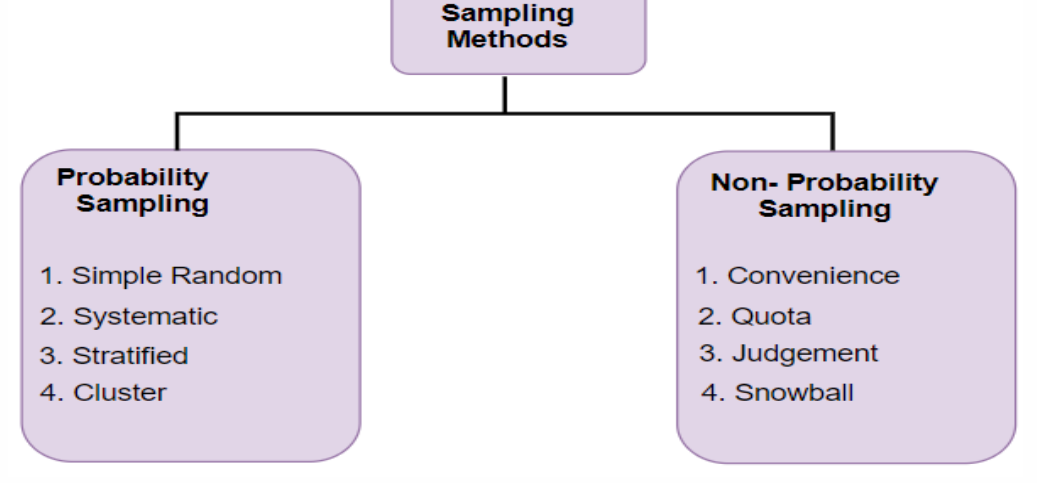
**Group Terms**

* **Population:** Set of entities or items on which the conclusion has to be drawn.

**Properties Of Sample**

* Must contain data of many **varieties**.
* But stay **unbiased** about their properties.
* Sample data must be chosen **randomly**.

**Sampling Methods**



**Probability Sampling**

* Here the probability of an object to be selected is **known**.
* **Simple random sampling:** Selection are made randomly, **by chance**.
* **Stratified random sampling:** Many sub-groups (***strata***) are created & each one is sampled for test.
* **Systematic sampling:** Sampling is done with certain **parameters** set for selecting objects/data.
* **Cluster sampling:** Same as **stratified random sampling** but the samples from each group are brought & tested **together**.

**Non-Probability Sampling**

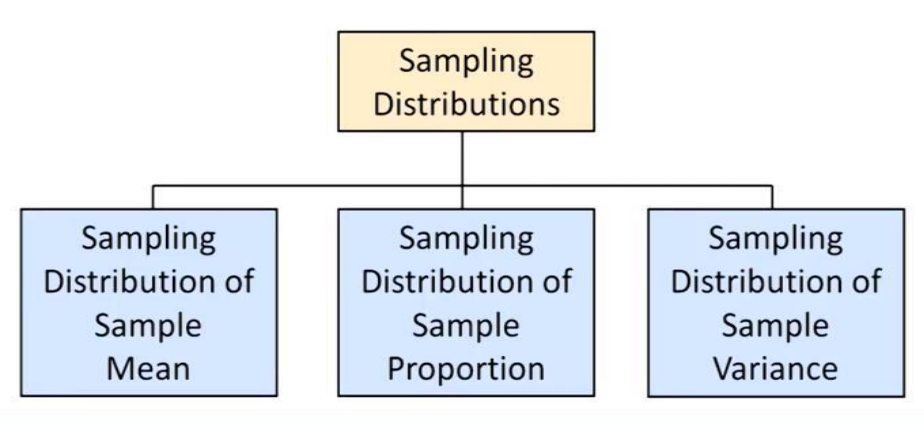
* Here the probability of an object to be selected is **unknown**.
* **Convenience sampling:** Most accessible sample is collected.
* **Purposive sampling:** Sampling is done with some biases.
* **Quota sampling:** Many sub-groups (strata) are created & samples from them are selected as per a bias/quota.
* **Snowball sampling:** At first, sample is selected & then some members from it refer to other members to be added.

**Topic – 7: Sampling Distribution**

**Introduction**

* ***Sampling distribution*** means **statistically distributing data**.
* There are many **factors** we can think of, affecting it.

**Types Of Sampling Distributions**



**Sampling Distribution Of Sample Mean**

* Statistical distribution of **means** of various data.
* ***Central limit theorem*** says that mean is quite good when taken in **large quantities**.

**Sampling Distribution Of Proportions**

* Statistical distribution of **successful portions** from sample, represented with **0s** & **1s**.
* Here too the **central limit theorem** says that mean is quite good when taken in **large quantities**.
* **Sampling proportion** is equal to the **population proportion** (I **don’t** know what the hell it means).

**Sampling Distribution Of Sample Variance**

* Central limit theorem says it tends to follow chi-square distribution when taken in large quantities.
* Sampling proportion is equal to the population proportion.

**Topic – 8: Confidence & Estimators**

**Confidence Interval Estimation**

* **Confidence interval:** Range of values which when taken again from another sample, falls in **same range**.
* **Confidence interval estimation:** Range of values in which the population mean likely lies.

**Confidence level = 1 – α**

**Point Estimator & Interval Estimate**

* **Point estimator:** A statistical value like **mean**, used as a **point** to **compare** other values statistically.
* **Interval estimate:** A range we get after adding & subtracting **possible error value**.

**Population mean interval estimation = Mean + Margin on error**