Parul® University

Data Analytics and Data Visualization(303105314)

Unit 2 : Introduction to Python Fundamentals and Statistics

Arun Chauhan

Assistant Professor

Computer Science & Engineering

Parul® University

DIGITAL LEARNING CONTENT

Outline

- Introduction to Python
- Importance of Python
- Python Libraries for stats
- Introduction to stats
- Central tendency and Dispersion
- Types of variable
- Levels of Data measurement
- Sampling and Sampling Distribution
- Distribution of Sample Means, Population and Variance
- Confidence interval estimation

Parul®
University

DIGITAL LEARNING CONTENT

Introduction to Python

Python:

- …is a general purpose interpreted programming language.
- …is a language that supports multiple approaches to software design, principally structured and object-oriented programming.
- …provides automatic memory management and garbage collection
- …is extensible
- …is dynamically typed.

Importance of Python

- Rich Ecosystem of Libraries: Python boasts powerful libraries such as Matplotlib, Seaborn, Plotly, and Bokeh, which offer extensive functionality for creating various types of visualizations. These libraries provide high-quality plots with customizable features.

- Ease of Use: Python's syntax is clean and easy to understand, making it accessible for beginners and experts alike. With straightforward code, users can quickly generate complex visualizations without extensive programming knowledge.

Contd.

- Interactivity: Python libraries like Plotly and Bokeh allow for interactive visualizations, enabling users to explore data dynamically. Interactivity enhances the viewer's engagement and facilitates deeper insights into the data.

- Integration with Data Analysis Tools: Python seamlessly integrates with popular data analysis libraries such as Pandas and NumPy. This integration streamlines the process of data manipulation, analysis, and visualization, creating a cohesive workflow.

Contd.

- Customization: Python visualization libraries offer extensive customization options, allowing users to tailor visualizations to specific requirements. From adjusting colors and styles to incorporating annotations and labels, Python empowers users to create visually appealing and informative plots.

- Support for Large Datasets: Python's efficient memory management and scalability make it suitable for handling large datasets. Whether dealing with millions of data points or complex multidimensional arrays, Python can efficiently process and visualize vast amounts of data.

Contd.

- Reproducibility: Python promotes reproducible research by enabling users to document their visualization process through Jupyter Notebooks or scripts. This transparency ensures that others can replicate the analysis and verify the results.

- Community Support: Python has a vibrant and active community of users and developers who contribute to the development of visualization tools and provide support through forums, tutorials, and documentation. This rich ecosystem fosters collaboration and knowledge sharing.

Python Libraries For Stats

Pandas:

- Pandas or Python Pandas is a library of Python which is used for data analysis.
- The term Pandas is derived from “Panel data system”.
- Pandas is an open-source data manipulation and analysis library for Python.
- It provides easy-to-use data structures and functions for working with structured data, such as tables and time series.
- We can perform different stats function on data by using pandas library inbuilt function (mean, median, mode, variance etc)

Contd.

- It is built on top of the numpy library which means that a lot of the structures of numpy are used or replicated in Pandas.
- Pandas generally provide two data structures for manipulating data. They are series and data frames.

Pandas Series

- A Pandas Series is a one-dimensional labeled array capable of holding data of any type (integer, string, float, Python objects, etc.). The axis labels are collectively called indexes.

Contd.

- The Pandas Series is nothing but a column in an Excel sheet. Labels need not be unique but must be of a hashable type.
- The object supports both integer and label-based indexing and provides a host of methods for performing operations involving the index.
- Pandas Series is created by loading the datasets from existing storage (which can be a SQL database, a CSV file, or an Excel file).

- Pandas Series can be created from lists, dictionaries, scalar values, etc.

Contd.

Pandas Data Frame:

- Pandas DataFrame is a two-dimensional data structure with labeled axes (rows and columns).

Creating Data Frame:

- Pandas DataFrame is created by loading the datasets from existing storage (which can be a SQL database, a CSV file, or an Excel file).
- Pandas DataFrame can be created from lists, dictionaries, a list of dictionaries, etc.

Contd.

NumPy:

- NumPy, which stands for Numerical Python, is a fundamental package for scientific computing in Python.
- It provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays efficiently.
- NumPy is widely used in fields such as data science, machine learning, engineering, and scientific research.

Contd.

NumPy arrays come in two forms-

- 1-D array – also known as Vectors.
- Multidimentional arrays also known as Matrices.

- Once a NumPy array is created, you cannot change its size.

- NumPy array contain elements of homogenous type.

Introduction to stats

- Statistics is a branch of mathematics that deals with the collection, analysis, interpretation, presentation, and organization of data.
- It provides methods for making inferences and decisions in the presence of uncertainty or variability.

- It is mainly of two types: Descriptive and Inferential

Descriptive stats:

- Descriptive statistics involve methods for summarizing and describing the features of a dataset.

Contd.

- descriptive statistics include measures of central tendency (mean, median, mode), measures of dispersion (range, variance, standard deviation), and measures of shape (skewness, kurtosis).

Inferential statistics:

- Inferential statistics is a branch of statistics that involves making inferences or predictions about a population based on a sample of data taken from that population. It's used to draw conclusions, make predictions, or test hypotheses about a larger group (population) based on the characteristics observed in a smaller sample from that group.

Parul®
University

DIGITAL LEARNING CONTENT

Central tendency and Dispersion

Central tendency: Central tendency refers to the typical or central value of a data set. There are three main measures of central tendency:

Mean: The mean is the average of a set of numbers. It's calculated by adding up all the values in the data set and then dividing by the total number of values.

Contd.

The average of your dataset

The value obtained by dividing the sum of a set of quantities by the number of quantities in the set

Example: (22+18+30+19+37+33) = 159 ÷ 6 = 26.5

The mean is sensitive to extreme values

Mean= $$m = \frac{\text{sum of the terms}}{\text{number of terms}}$$

Contd.

Median: The median is the middle value in a sorted list of numbers. If there is an even number of values, the median is the average of the two middle numbers.

The median is not as sensitive to extreme values as the mean

Odd number of numbers, median = the middle number

Median of 2, 4, 7 = 4

$$\text{Median} = (n+1)/2^{th} \text{ observation}$$

Contd.

Even number of numbers, median = mean of the two middle numbers

Median = $[(n/2)^{th}$ obs.+ $((n/2)+1)^{th}$ obs.$]/2$

•Median of 2, 4, 7, 12 = (4+7) /2 = 5.5

Contd.

Mode: The mode is the value that appears most frequently in a data set.

Consider this dataset showing the retirement age of 11 people, in whole years:

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

- This table shows a simple frequency distribution of the retirement age data.

Contd.

Age	Frequency
54	3
55	1
56	1
57	2
58	2
60	2

The most commonly occurring value is 54, therefore the mode of this distribution is 54 years.

Contd.

Measures of dispersion: Measures of dispersion describe the spread or variability of data points within a data set. Common measures of dispersion include:

- Range
- Variance
- Standard Deviation

Contd.

Range: The range is the difference between the largest and smallest values in a data set.

Consider the following set of numbers representing the ages of a group of people:

{18,22,25,30,35,40,45}

To calculate the range:

Identify the smallest value: 18

Identify the largest value: 45

Subtract the smallest value from the largest value:

45−18=27

Contd.

Variance:

- Commonly used measure of dispersion whose computation depends on all the data.
- The larger the variance, the more the data are spread out from the mean and the more variability one can expect in the observations.
- The formula used for calculating the variance is different for populations and samples.

Contd.

The formula for the variance of a population is :

xi: value of the ith item

N: number of items in population

μ is the population mean.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Contd.

Essentially, the variance is the average of the squared deviations of the observations from the mean.

- A significant difference exists between the formulas for computing the variance of a population and that of a sample. The variance of a sample is calculated using the formula
- where n is the number of items in the sample and $\bar{x}$ is the sample mean

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

Contd.

	A	B	C	D
1	Observation	Cost per order	(xi - mean)	(xi - mean)^2
2	x1	$2,700.00	-$23,595.32	$556,739,085.74
3	x2	$19,250.00	-$7,045.32	$49,636,521.91
4	x3	$15,937.50	-$10,357.82	$107,284,417.52
5	x4	$18,150.00	-$8,145.32	$66,346,224.04
93	x92	$74,375.00	$48,079.68	$2,311,655,710.74
94	x93	$72,250.00	$45,954.68	$2,111,832,692.12
95	x94	$6,562.50	-$19,732.82	$389,384,151.56
96	Sum of cost/order	$2,471,760.00	Sum of squared deviations	$82,825,295,365.68
97	Number of observations	94		
98				
99	Mean cost/order	$26,295.32	Variance	890,594,573.82
100				
101			Excel VAR.S function	890,594,573.82

Contd.

Standard Deviation

- The standard deviation is the square root of the variance. For a population, the standard deviation is computed as

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

Contd.

Sample variance formula: $$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Population variance formula: $$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Sample standard deviation formula: $$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Population standard deviation formula: $$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Measure of Shape

- Measures of shape describe the distribution (or pattern) of the data within a dataset.

Skewness:

- **Definition**: Skewness quantifies the asymmetry of a distribution.
- **Types**:
  - **Positive Skewness (Right-Skewed)**: The tail on the right side is longer or fatter. The mean and median will be greater than the mode.

Contd.

Negative Skewness (Left-Skewed): The tail on the left side is longer or fatter. The mean and median will be less than the mode.

Definition: Kurtosis measures the "tailedness" of a distribution, indicating the presence of outliers. Means it describe the measure of tail to its peak

Types:

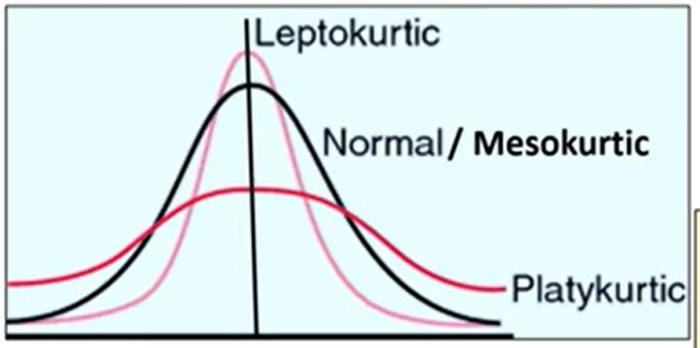
- **Leptokurtic**: They have heavy tails and a sharp peak.

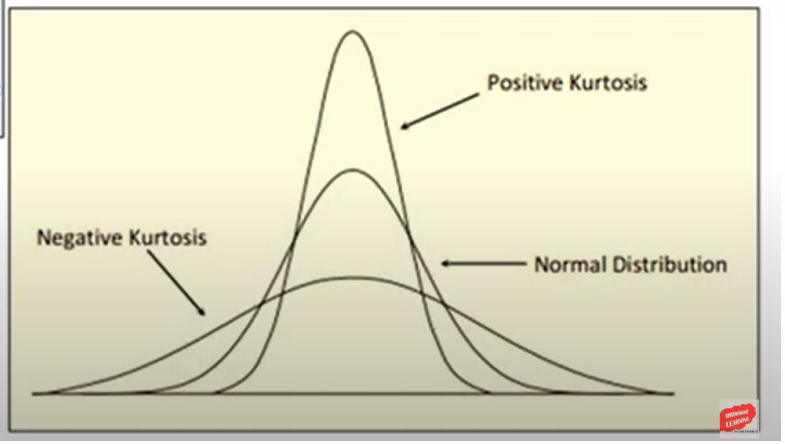
Contd.

- **Platykurtic**: They have light tails and a flatter peak.
- **Mesokurtic**: This is typical of a normal distribution.

Contd.

Kurtosis forms



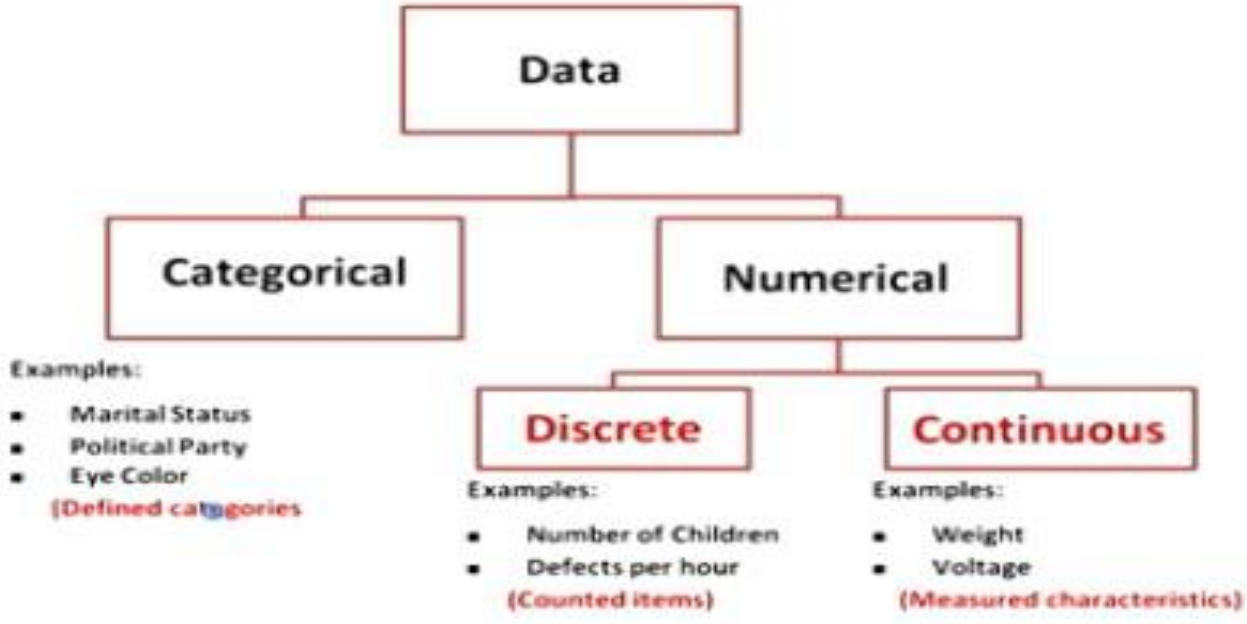


Contd.

Skewness	Kurtosis
Skewness measures the asymmetry of a probability distribution	Kurtosis measures the tailedness or peakedness of a probability distribution
Positive skew indicates a right-skewed distribution, with the tail extending to the right	Positive kurtosis indicates a distribution with heavier tails, often referred to as “leptokurtic”
Negative skew indicates a left-skewed distribution, with the tail extending to the left	Negative kurtosis indicates a distribution with lighter tails, often referred to as “platykurtic”
A skewness value of zero indicates a symmetric distribution	A kurtosis value of zero indicates a distribution similar to the normal distribution, often referred to as “mesokurtic”
Used to identify the direction and degree of asymmetry	Used to identify the presence of outliers or extreme values
Sensitive to changes in the tails of the distribution	Sensitive to changes in the center and shoulders of the distribution
Commonly used in fields such as economics, finance, and social sciences	Commonly used in statistics, engineering, and physical sciences
Examples: income distribution, stock returns	Examples: particle physics, image processing

Types of Variables

6.1 Types of Variables



Levels of Data measurement

- We classified as the categorical data and numerical data. There is another way of classification is, classifying into nominal data, ordinal data, interval data and ratio data.

Nominal:

- A nominal scale classifies data into distinct categories in which no ranking is implied.
- Example: Gender, Marital Status
- gender suppose you are conducting a questionnaire. Suppose you captured the gender male 0, female 1. This 0 1 represents just the gender. You cannot do any arithmetic operations with the help of the 0 & 1.

Contd.

Ordinal

- An ordinal scale classifies data into distinct categories in which ranking is implied.
- Example : Product satisfaction: Satisfied, Neutral, Unsatisfied
  Faculty rank: Professor, Associate Prof, Assistant Prof

Interval

- An interval scale is an ordered scale in which the difference between measurements is meaningful quantity but the measurements do not have a true zero point

Contd.

- For example, in the case of temperature, a reading of 0°C or 0°F doesn't mean there is no temperature; it simply represents an arbitrary point chosen as the freezing point of water. This lack of a true zero point prohibits operations like multiplication or division based on the scale. You can't say that 20°C is "twice as hot" as 10°C because the zero point is arbitrary and doesn't represent the complete absence of temperature.

Contd.

Ratio

- The ratio scale is the ordered scale in which the difference between the measurements is a meaningful quantity and the measurements have the true zero point.
- Weight, age, salary and the Kelvin temperature comes under ratio scale. Because 0 Kelvin that means the absence ofthe heat.
- So in the ratio scale, he can do all kinds of arithmetic operation.

Sampling and Sampling Distributions

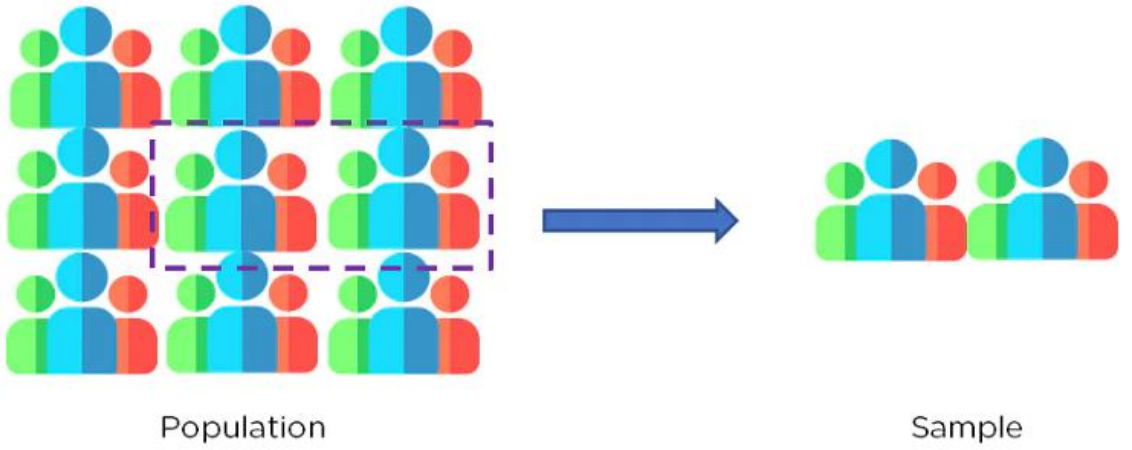
Population

- A population refers to the entire group of individuals, items, or events about which you want to draw conclusions or make inferences.

Sampling and Sampling Distributions

Sample

- A sample is a subset of the population selected for observation or data collection.
- It represents a smaller, manageable portion of the population from which data can be collected and analyzed.



Contd.

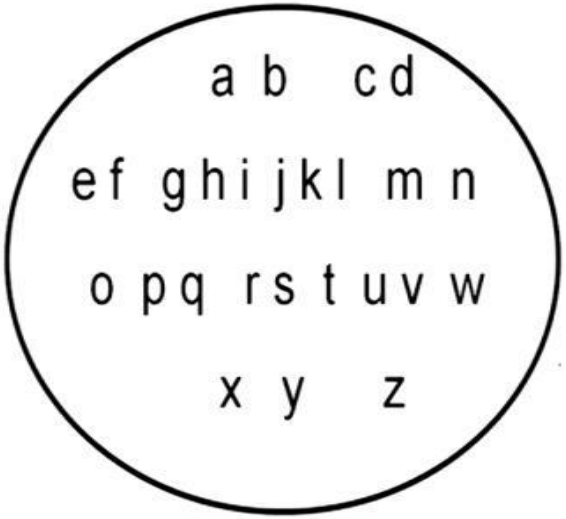
A sample should generally :

- Satisfy all different variations present in the population as well as a well-defined selection criterion.
- Be utterly unbiased on the properties of the objects being selected.
- Be random to choose the objects of study fairly.

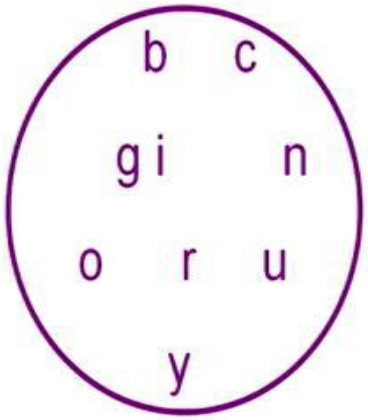
Contd.

Population vs. Sample

- **Population**



- **Sample**



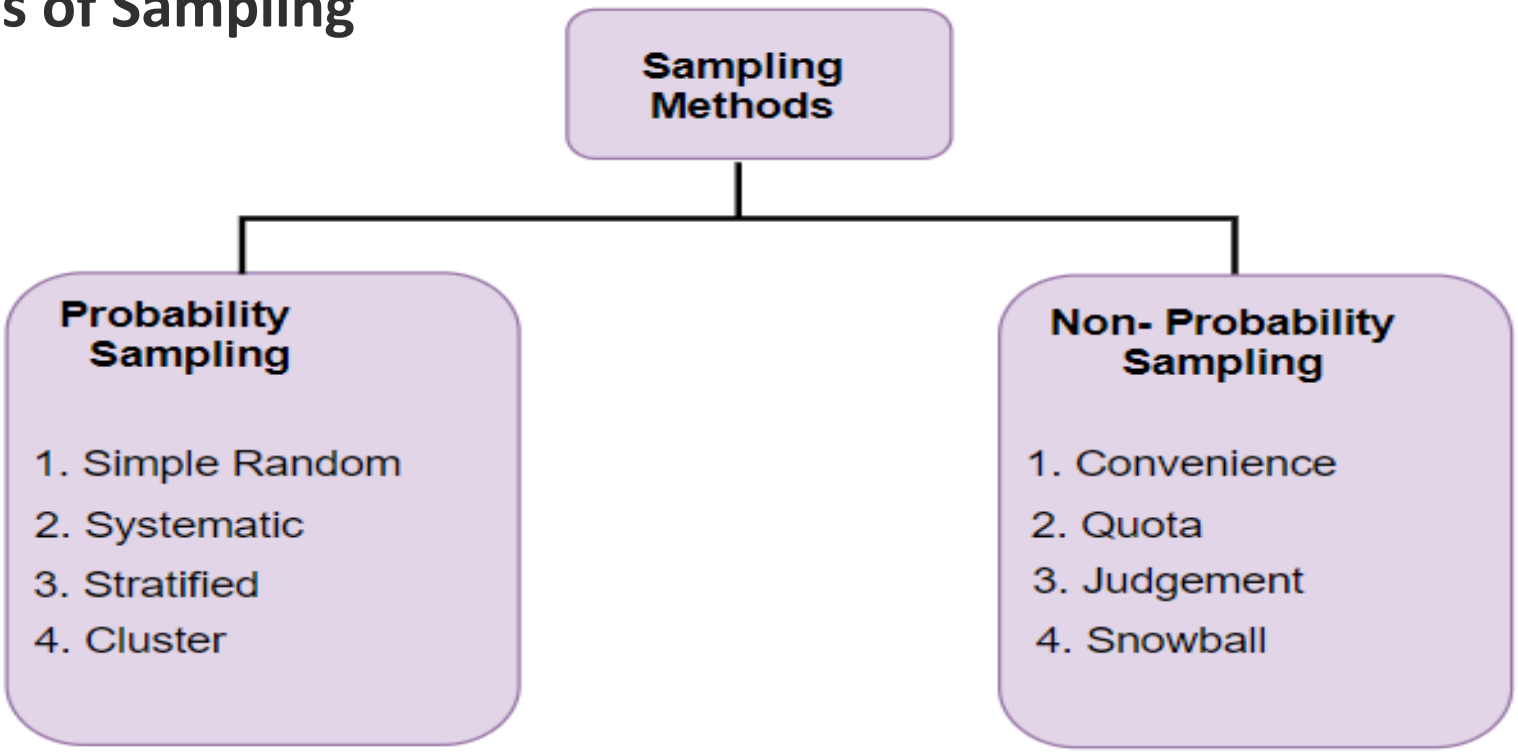
Contd.

Sampling

- In statistics, sampling is the selection of a subset or a statistical sample (termed sample for short) of individuals from within a statistical population to estimate characteristics of the whole population.
- The subset is meant to reflect the whole population and statisticians attempt to collect samples that are representative of the population.
- Sampling has lower costs and faster data collection compared to recording data from the entire population, and thus, it can provide insights in cases where it is infeasible to measure an entire population.

Contd.

Types of Sampling



Contd.

Probability Sampling

- In probability sampling, every member of the population has a known and non-zero chance of being selected.
- **Simple Random Sampling:** Each member of the population has an equal chance of being selected, and selections are made entirely by chance.
- **Stratified Random Sampling:** The population is divided into homogeneous subgroups (strata), and then simple random samples are taken from each stratum.
- **Systematic Sampling:** A sample is drawn by selecting every nth member from a list of the population, where n is calculated based on the population size and desired sample size.

Contd.

- **Cluster Sampling:** The population is divided into clusters, and then a random sample of clusters is selected. All members within the chosen clusters are included in the sample.

Non-probability Sampling

- In non-probability sampling, the probability of any particular member being selected is unknown or cannot be calculated.
- **Convenience Sampling:** Samples are selected based on their convenience or accessibility to the researcher.
- **Purposive Sampling:** Samples are selected based on the researcher's judgment or specific criteria, often chosen to fulfill a particular objective or purpose.

Contd.

- **Quota Sampling:** The population is divided into subgroups, and a specific number of individuals are sampled from each subgroup based on predetermined quotas.
- **Snowball Sampling:** Initial participants are selected, and then additional participants are recruited based on referrals from those initial participants.

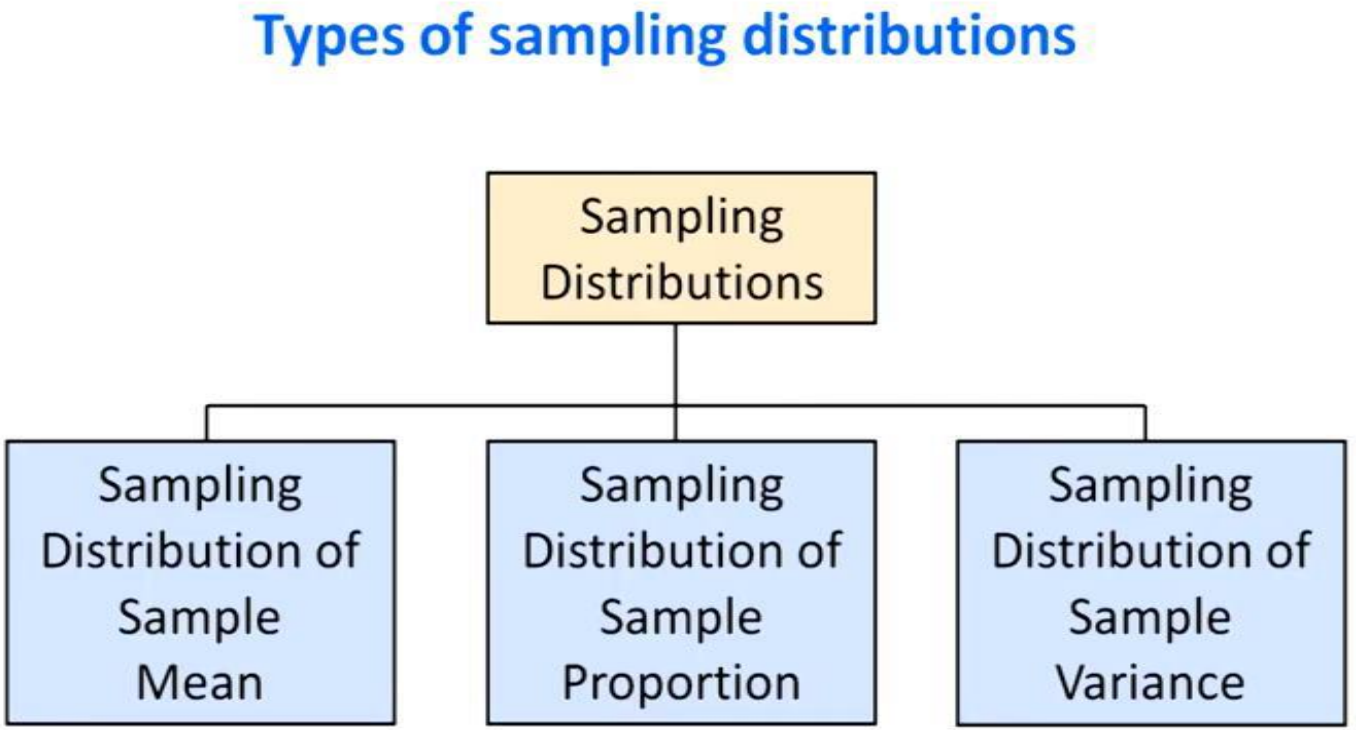
Contd.

Sampling Distribution

- sampling distribution is a distribution of all of the possible values of your statistic for a given size sample selected from the population.
- The sampling distribution depends on multiple factors – the statistic, sample size, sampling process, and the overall population. It is used to help calculate statistics such as means, ranges, variances, and standard deviations for the given sample.

Distribution of Sample Mean, Population and Variance

Types of Sampling Distributions



Contd.

1. Sampling distribution of sample mean

- This distribution represents the distribution of sample means obtained from all possible samples of a particular size taken from a population.
- It is often approximately normal, especially for large sample sizes, due to the Central Limit Theorem.
- The mean of the sampling distribution of sample means is equal to the population mean, and the standard deviation (or standard error) is equal to the population standard deviation divided by the square root of the sample size.

Contd.

2. Sampling distribution of proportion

- This distribution represents the distribution of sample proportions (e.g., the proportion of successes in a sample) obtained from all possible samples of a particular size taken from a population.
- It is approximately normal for large sample sizes, following the Central Limit Theorem, and is bounded between 0 and 1.
- The mean of the sampling distribution of sample proportions is equal to the population proportion, and the standard deviation is calculated using the population proportion and the sample size.

Contd.

3. Sampling distribution of sample variance

- This distribution represents the distribution of sample variances obtained from all possible samples of a particular size taken from a population.
- It tends to follow a chi-square distribution, especially for large sample sizes.
- The mean of the sampling distribution of sample variances is equal to the population variance, and the variance of the sampling distribution is related to the population variance and the sample size.

Confidence Interval Estimation

Confidence Interval

- A confidence interval is a range of values, calculated from sample data, that is likely to contain the true value of a population parameter with a certain level of confidence.

Confidence Interval Estimation

- Confidence interval estimation involves calculating a range of values within which a population parameter, such as a mean, is likely to lie.

Contd.

- A confidence interval is the mean of your estimate plus and minus the variation in that estimate. This is the range of values you expect your estimate to fall between if you redo your test, within a certain level of confidence.

- Confidence, in statistics, is another way to describe probability. For example, if you construct a confidence interval with a 95% confidence level, you are confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval.

Contd.

- Your desired confidence level is usually one minus the alpha (α) value you used in your statistical test:
- Confidence level = 1 − a
- So if you use an alpha value of 0.05 for statistical significance, then your confidence level would be 1 − 0.05 = 0.95, or 95%.

Contd.

Pont Estimator and confidence interval

- A point estimator is a statistic used to estimate an unknown parameter in a statistical model. It's called a "point" estimator because it provides a single, specific value as an estimate for the parameter of interest. For example, if you're interested in estimating the mean of a population, you might use the sample mean as a point estimator.

- A point estimator cannot be expected to provide the exact value of the population parameter.

Contd.

- An interval estimate can be computed by adding and subtracting a margin of error to the point estimate.
- Point Estimate +/□ Margin of Error
- The purpose of an interval estimate is to provide information about how close the point estimate is to the value of the parameter.
- The general form of an interval estimate of a population mean is

$$\bar{x} \pm \text{Margin of Error}$$

DIGITAL LEARNING CONTENT

Parul® University

www.paruluniversity.ac.in