Essential Statistics for Data Science

Topics:
Inferential Statistics, Descriptive Vs Inferential Statistics.

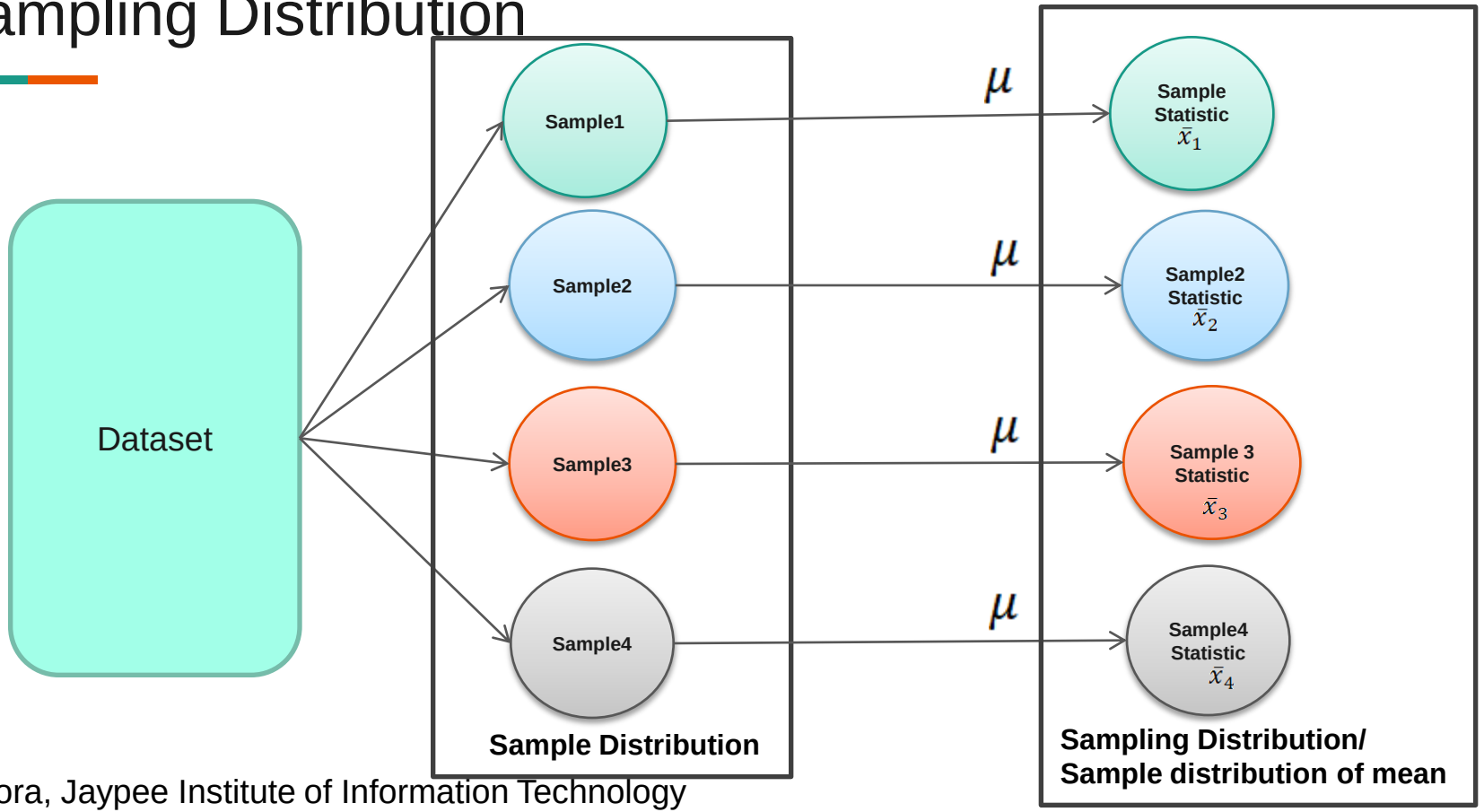
Sampling Distribution, Central Limit Theorem

Dr. Anuja Arora, Jaypee Institute of Information Technology

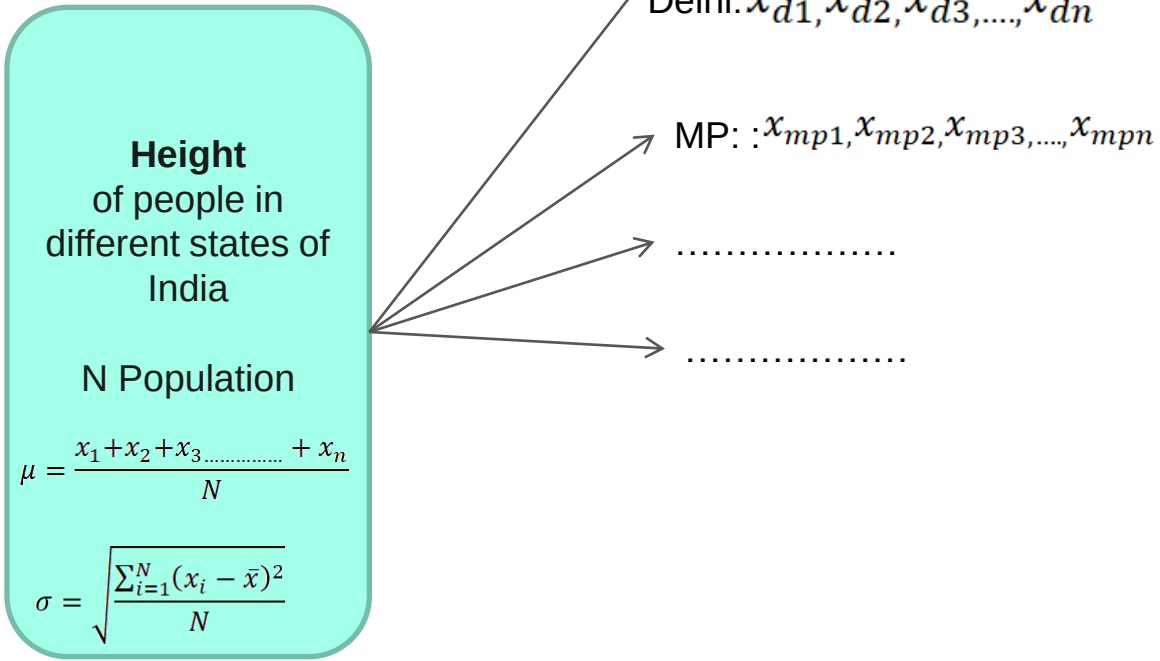
Deterministic Vs Inferential Statistics

Deterministic Statistics	Inferential Statistics
Measure for definitive measurement	Note the margin of error of research performed
Data Summary: Bar Graph, Histogram, pie chart, etc.	**Data Summary:** Takes sample data and makes inferences about the sample population. i.e used to make inference or draw a conclusion of population.
Measures of Central Tendency: Mean, Median, Mode **Measures of dispersion:** Range, Variance, standard deviation	Uses probability to determine how confident we can be that the conclusion we make is correct Normal (Gaussian distribution), Binomial distribution, etc.

Sampling Distribution



Sampling Distribution



Sample Distribution

$\overline{x_d}$ σ_d

$\overline{x_{mp}}$ σ_{mp}

Sampling distribution

$$mean(\bar{x}) \approx \mu$$

$$SD(\bar{x}) < \sigma$$

Central Limit Theorem (CLT)

- Three mean values:

$$\mu \rightarrow original\ Population\ Mean$$

$$\overline{x_{mp}}, \overline{x_d} \rightarrow Sample\ Mean\ varies\ for\ each\ sample$$

$$\bar{x} \rightarrow mean\ of\ Sample\ means$$

- **CLT Part-I:** Sample is taken from population, So mean of Sample will approximately be same as of original population. Sample size (large/ small) does not matter. Even mean of samples will also be approximately same as shown below

$$\bar{x} \approx \mu \approx \overline{x_{mp}}, \overline{x_d}$$

Central Limit Theorem (CLT)

- **CLT Part-II:** The standard deviation of the sample means decreases as the sample size increases i.e

$$SD(\bar{x}) < \sigma$$

Its formula looks like

$$SD(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

this is known as standard error as well

↑ n increase ↓ Standard error decrease

- **CLT Part-III:** If the population follows a normal distribution, then the sampling distribution is also normal distribution. The distribution of the sample means approaches a normal distribution, under certain conditions,

$$\bar{x} \sim N(\mu_{\bar{x}}, \sigma_{\bar{x}} / \sqrt{n})$$

Online simulation Link: http://www.ltcconline.net/greenl/java/Statistics/clt/cltsimulation.html

CLT Theorem:

Distribution of sample statistics is nearly normal, centred at the population mean, and with a standard deviation equal to the population standard deviation divided by square root of sample size.

$$\bar{x} \sim N(\mu_x, \sigma_x / \sqrt{n})$$