Essential Statistics for Data Science

Topics:
About Data and Data Science

Data Characteristics: Measure of Central tendency

Measure of Data Dispersion

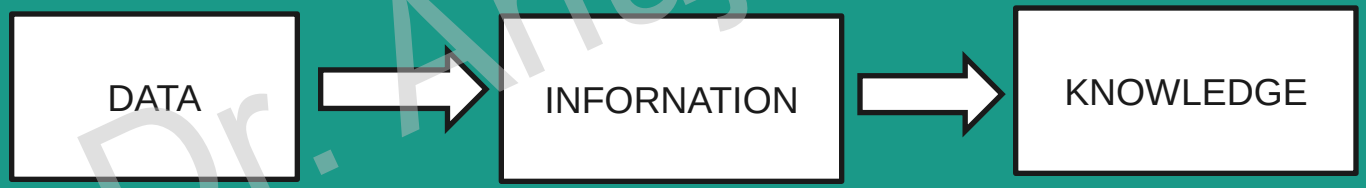
Dr. Anuja Arora, Jaypee Institute of Information Technology

Data Scientist Role

- **Collect large amount** of unruly data and **transforming** it into a more usable format.
- Solving business related problems using **data driven techniques**.
- Working with a variety of **programming language** such as SPSS, R, Python, etc.
- Having a solid grasp of **statistics**, including statistical tests and distribution.
- Staying on top of **data learning techniques** such as machine learning, deep learning, text analytics.
- **Communicating and collaborating** with both IT and Business.
- Extract out **pattern and order** in data as well as **spotting trends** that can help a business's bottom line.

Data Science

To extract out insights, knowledge, and information from data is known as ‘Data Science’



Descriptive Data Summarization

Data Characteristics (to understand distribution of data)

- Measure of Central tendency (mean, median, mode, mid range)
- Measure of Data dispersion Quartiles, interquartile range (IQR), Variance, Standard deviation

Measure of Central Tendency

- **Mean:** Average of Data

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - Weighted arithmetic mean (consider n sample size of i records)

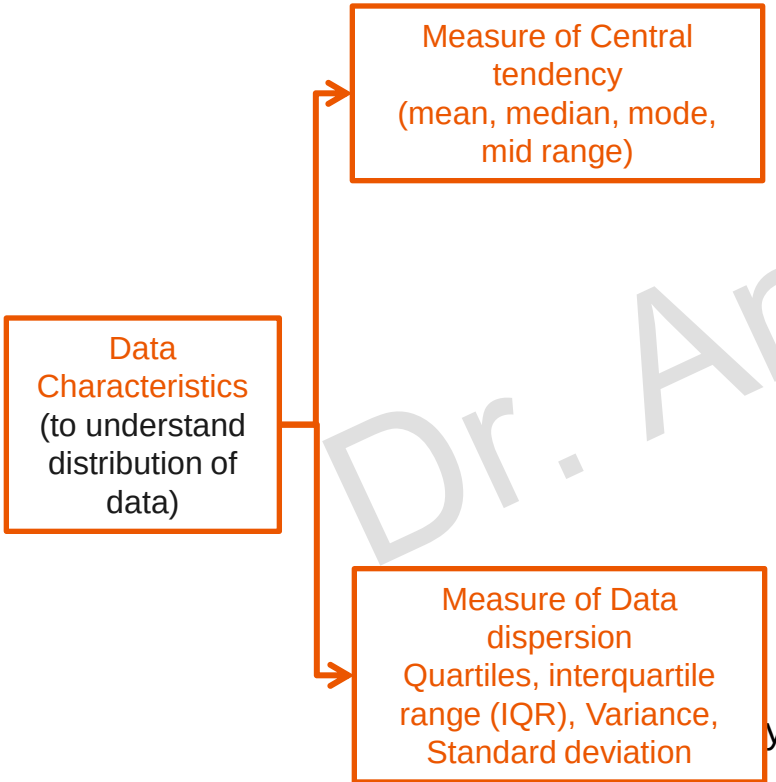
$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

  - Trimmed mean: chopping extreme values
- **Median:**
  - case1: Odd Numbers, Median= Middle Value
  - case2: Even Number, Median = Average of Middle 2 values
  - case3: Estimated by interpolation (for grouped data).

$$median = L_1 + (\frac{n/2 - (\sum freq)l}{freq_{median}})width$$

- **Mode:**
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula: $mean - mode = 3 \times (mean - median)$

Descriptive Data Summarization



Measure of Data Dispersion

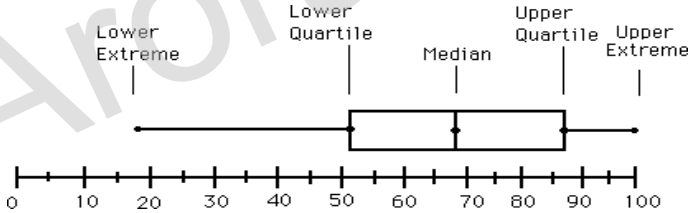
- **Quartiles, outliers and boxplots**
  - **Quartiles**: Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
  - **Inter-quartile range**: IQR = $Q_3 - Q_1$
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually.
  - **Outlier**: A value which is far away from the normal group of data. Usually, a value higher/lower than 1.5 x IQR
  - Usually, **High Range** = Q3 +1.5*IQR

    Low Range = Q1- 1.5*IQR

Box Plot Analysis

- **Five-number summary** of a distribution

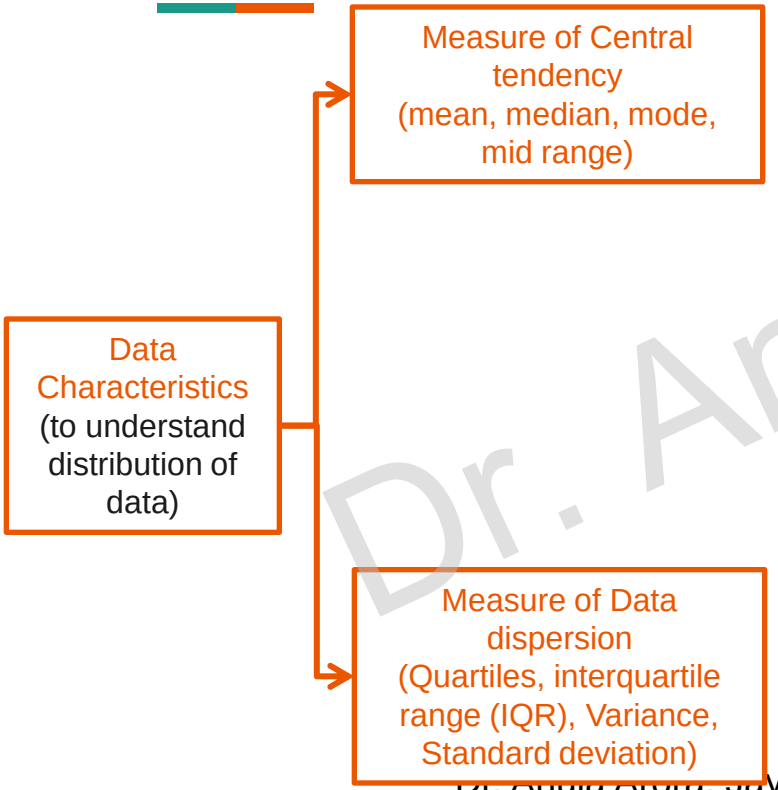
  Minimum, Q1, Median, Q3, Maximum



- **Boxplot**

  Data is represented with a box
  - The **ends of the box** are at the first and third quartiles, i.e., **the height of the box** is IQR
  - The **median is marked by a line** within the box
  - **Whiskers:** two lines outside the box extended to **Minimum and Maximum**
  - **Outliers**: points beyond a specified outlier threshold, plotted individually

Descriptive Data Summarization



- **Variance:**

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

- **Standard deviation** *s (or σ)* is the square root of variance s^2 *(or* σ^2*)*

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