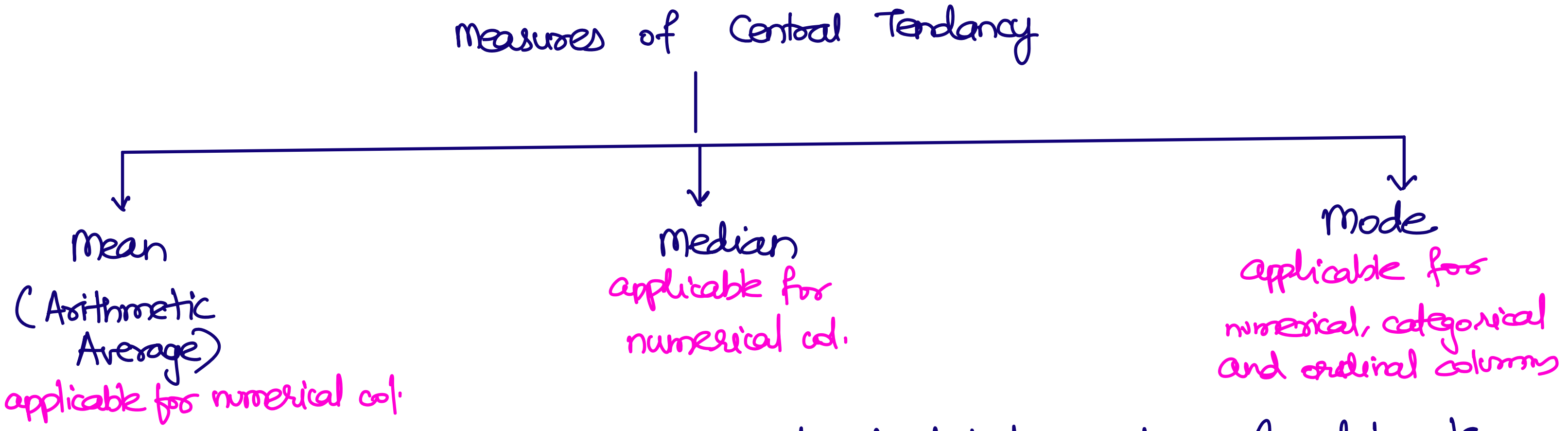
Applied Data Science

Trainer: Prashant Nair

Descriptive Statistics

Measures of Central Tendancy:

These are statistical tools used to summarize the data based on the central point of the data.



Measures of Central Tendancy

Mean (Arithmetic Average)

applicable for numerical col.

Median

applicable for numerical col.

Mode

applicable for numerical, categorical and ordinal columns

ⓐ Mean: The sum of all data points divided by number of data pts.

$$\text{mean} = \frac{\sum_{i=1}^{n} x_i}{n}$$

n.... no. of records

eg: [2, 3, 4, 5]

$$\text{mean} = \frac{2+3+4+5}{4} = 3.5$$

ⓑ Median: Its the middle value of the dataset when its ordered from smallest to largest.

Calc. median

① Order the data in ascending order.

② Identify the middle position

- if no. of records is odd, the median value is at $\left(\frac{n+1}{2}\right)^{th}$ position.

eg: [3, 1, 4, 1, 5]

→ [1, 1, 3, 4, 5]

Here,
the middle position will be $\frac{5+1}{2} = 3^{rd}$ position ⇒ 3//

- if no. of records is even, the median value is the avg of $\left(\frac{n}{2}\right)^{th}$ position value and $\left(\frac{n}{2}+1\right)^{th}$ position value.

eg: [3, 1, 4, 1]

→ [1, 1, 3, 4]

Here,
middle position $\left(\frac{4}{2}\right) = 2^{nd}$ & $\left(\frac{4}{2}+1\right) = 3^{rd}$

median = $\frac{1+3}{2} = 2$

③ Mode → most frequent data. / data with highest frequency (count)

Steps:

① Count the frequency of each unique value.

② Identify the value with highest frequency.

eg [1, 2, 2, 3, 3, 3, 4]

1 → 1   4 → 1

2 → 2

3 → 3

∴ mode = 3//

Note: if all unique data has same frequency, in that case, sort the data in ascending order and return the first value

eg: [1, 2, 3, 4] → 1

[1, 2, 2, 3, 3] →

1 → 1

2 → 2 }

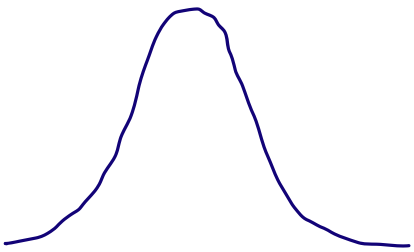
3 → 2 } → sort → 2

Data Distributions

This concept is applicable only for numerical columns

(a) Normal Distribution | Gaussian Distribution | Bell curve

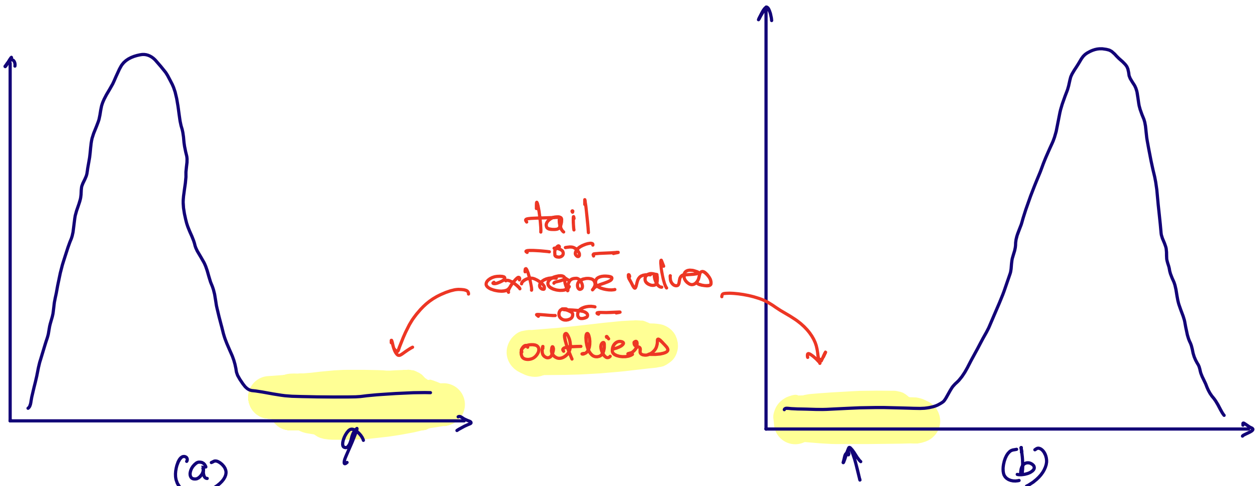
GENERALIZED DATA



if data column forms a bell curve, chances are data is generalized.

(the data may represent the population)

(b) Skewed distribution.



(a) Right Skewed

(b) Left skewed

Quartiles

Quartiles are the values that divide the dataset into 4 equal parts.

25%	50%	75%	100%
Q_1	Q_2	Q_3	Q_4
First Quartile	Second Quartile (median)	Third Quartile	maximum

Consider a dataset $D = \{12, 15, 7, 23, 10, 8, 14, 18, 5, 20\}$

To calc quartiles

Step 1: Arrange the data in ascending order

$\{5, 7, 8, 10, 12, 14, 15, 18, 20, 23\}$

1 2 3 4 5 6 7 8 9 10

Step 2: Identify the median

median = 13 ⇒ Q2

Step 3: Calc first quartile (Q1)

Q1 is the median of the lower half of the dataset

Q1 = 8 //

Step 4: Calc third quartile (Q3)

Q3 is the median of the upper half of the dataset

Q3 = 18 //

Step 5: Calc fourth quartile (Q4)

Q4 is the max value

Q4 = 23 //

Outliers

Outliers are those extreme values that affect the general tone of the domain.

To identify & remove an outlier, you can use Tukey's Method

—or—

1.5 IQR rule.

Tukey's Method | 1.5 $\underbrace{\text{IQR}}$ rule

↳ Inter Quartile Range

$IQR = Q3 - Q1$

Algo:

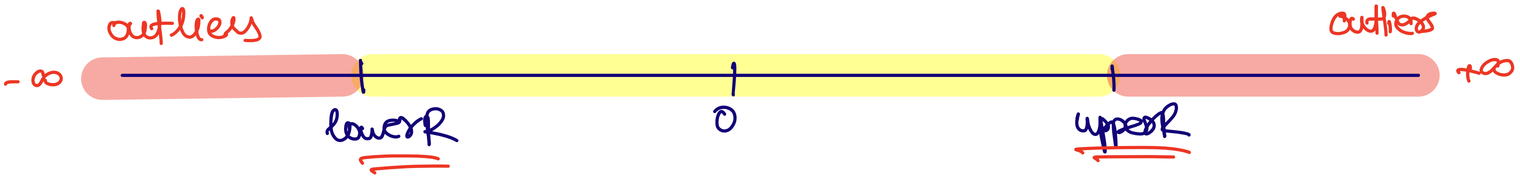
① Calc IQR

$IQR = Q3 - Q1$

② Calc the valid range of the given column.

$lowerRange = Q1 - (1.5 * IQR)$

$upperRange = Q3 + (1.5 * IQR)$


outliers
outliers
-∞
+∞
lowerR
0
upperR