

Date

# FORMAL (I)

## LECTURE NO.1 & NO.2.

### STATISTICS :-

The analysis and interpretation of data, where the set of observations is called as a dataset / sample

Following observations can be made using statistics.

### (i) AVERAGE:-

The term average often refers to a measure central tendency that summarizes a set of data with a single representative value.

↓      ↓      ↗  
Mean   Median   Mode.  
(centroid).

### (ii) MEDIAN:-

Median is the middle value in a dataset when it is arranged in ascending or descending order.

(always first sort the data before finding the mean)  
If there are odd no. of values, the median is the middle one.

If there are even no. of values, then median is the average of the two middle values.

### MODE :-

The value that occurs the most frequently in a dataset is/or the most common data point (i.e mode).

Other measures of spread offered by statistics are

(i) Variance

Date (spread of data from the mean)

(ii) Std. deviation.

$\sqrt{\text{variance}}$

(iv) Range:- (measure of spread of the data).

Largest Value - Smallest Value (of a dataset).

Example:-

Consider the uniform size for many students

171, 162, 165, 153, 172, 180

Now Range  $\Rightarrow 180 - 153 = 27\text{cm}$ .

Now the value of Range (as high) describes that there is a significant difference between the height of the tallest and the smallest individual, therefore the uniform must be designed in multiple size to accommodate diff students.

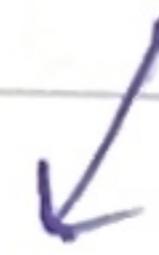
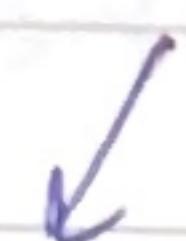
## DATA-TYPES :-

The datatype are divided into 2 primary.



Categorical

Numerical



Nominal

Ordinal

Discrete

Continuous

(that cannot be ranked ordered)  
i.e genders

Colors-

Marital status.

(that have a specific meaningful order)

- non fractional
- non decimal point

- fractional
- decimal point

- Education level e.g 1, 2, 3

e.g 0.123

- Customer

Satisfaction

100, -5, 65

0.487

0.961

level

- Military Ranks.

Date

Graphs types

(i) Histogram

(iv) Bar charts

(ii) Boxplot

(iii) Pie charts

Note:-

It is important to analyze / identify the type of data because different datatypes require different analysis Method.

## TWO SIDES OF STATISTICS



DESCRIPTIVE  
STATISTICS

o Organizes and Summarizes data using numbers and graphs.

e.g. i) avg. customer rating  
ii) total no. of clicks on a page

What is meant by organizing and summarizing data?

Organizing data: structuring and arranging to make it easy to work  
Summarizing data: reducing to key figures or visualiza

Organizing of data, techniques:

(i) Sorting data

(ii) Grouping data

(iii) Creating tables

Summarizing technique

(i) Median

(ii) Mean

(iii) Graph

## LECTURE NO. 3, NO. 4 & NO. 5

### MEAN:-

✓ It is the centroid of the data, by identifying the data set at that point, you can find the center of balance.

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

curly bracket ↓  
 represent input of  
 a data set . . . . .  
 total no. of  
 values.

E.g. dataset  $S1 = \{1, 2, 3, 4, 5, 6, 7, 12\} \Rightarrow 10$  data  
 considered an outlier.  
 may cause the data to be biased towards one end.

### ► SCALING OF MEAN :-

/ dividing.

✓ Multiplying mean with a const value is considered as scaling of mean.

✓ Scaling refers to multiplying / dividing all the values of a data set by a constant factor, it impacts not only the mean but all the other features of a dataset.

✓ It causes the graph to be stretched if  $\text{const} > 1$   
 and compressed if  $0 < \text{const} < 1$ .

$\Rightarrow$  relative distance mean ( $\{k \cdot x_i\}$ ) =  $k \cdot$  mean ( $\{x_i\}$ )  
b/w the points & from  
the mean remains proportional.  $k = \text{const.}$   $\Rightarrow$  the spread is changed  
accordingly.

### ► TRANSLATING THE MEAN

Translating refers to shift an entire dataset by adding or subtracting const value from each point in a dataset.

- \* This process moves the mean without affecting the spread or shape of the data.

## Why is there the need to scale a dataset?

Scaling ensures that different feature on a similar range or scale are maintained which are then used for :-

### (i) Machine Learning algorithms;

where same scale values will allow just distinguish b/w larger and smaller values allowing us to know which features shall dominate.

### (ii) Normalization for specific context:-

Interpretation and analysis of data from different sources in the same scale.

Why is there the need to translate a dataset?

(i) Mean centering:-

Translating the data so that the mean is at zero makes it easier to interpret results in many statistical analyses, particularly in regression etc.

- \* Many ML models perform better when mean is centered around zero.

(ii)

### IMPORTANT CHARACTERISTIC OF MEAN:-

(Why mean is used in stats especially for regression & hypothesis testing...)

- (i) The signed distances from the mean sum to 0.

$$\sum_{i=1}^N (x_i - \text{mean}\{x_i\}) = 0.$$

→ (a)

→ smallest total of squared distances from all numbers in our dataset

- (ii) The mean minimizes the sum of the squared distance from any real value.

$$\text{argmin}_\mu \sum_{i=1}^n (x_i - \mu)^2 = \text{mean}\{x_i\}$$

Date

Analyzing (a) from back why is there a need to square the distance from mean?

- (i) Squaring highlights larger difference more significantly.
- (ii) ensures all deviations contribute positively.
- (iii) simplifies calculation in regression and optimization.
- (iv) used in variance / std dev. to assess data spread.

## STD DEVIATION & VARIANCE.

The measure of spread with respect to mean.

- » Scaling would scale std. deviation.
- » Translating doesn't change the std. deviation.

Less std. deviation = less spread

More std. deviation = more spread.

## STANDARD COORDINATES/ NORMALIZATION OF DATA.

Sometimes the data may be taken from different sources or in consideration of different features, in order to analyze, compare or work with such data, it is important to standardize it.

mean of std coordinate = 0

Date \_\_\_\_\_

⇒ One standardization technique is to calculate the z-score of each data point even if from different datasets so that the z-score becomes comparable.

z-score: tells about how far a data point is from its mean of its dataset in terms of std deviation.

$$\checkmark \text{ z-score} = \frac{x - \mu}{\sigma} \Rightarrow x = \text{datapoint}$$

(z-score beyond +2 & -2 is considered as outlier).  $\mu = \text{mean of dataset}$   $\sigma = \text{std deviation of dataset}$ .

if z-score = 0 (student is at mean)

$\checkmark$  z-score  $> 0$  / positive (student is  $\frac{e.g. 1.5}{0.5}$  above mean)

$\checkmark$  z-score  $< 0$  / negative (student is  $\frac{e.g. -0.5}{0.5}$  below mean)

After standardization; the std coordinates always have;

$$\text{mean} = 0$$

$$\text{std. deviation} \& \text{variance} = 1$$

$$z_i = \frac{x_i - \text{mean}}{\text{std}}$$

Date

## LECTURE 5 & 6.

### MEDIAN :-

$$\text{dataset} = \{1, 3, 5, 4, 2\}$$

$$\begin{matrix} \text{sorted} \\ \text{dataset} \end{matrix} = \{1, 2, 3, 4, 5\}$$

Now the middle value is 3, it means that 50% of the data is below and above or less than or greater than 5.

- Scaling the dataset would also scale the median ✓
- Translating the dataset would not translate the median. ✓

### PERCENTILE :-

$k^{\text{th}}$  percentile is the value relative to which  $k\%$  of the data items have smaller or equal numbers.

Percentage indicates the exact value while percentile is more about a range or interval of values equal or smaller to the selected  $\$k$ .

\* Median is roughly 50<sup>th</sup> percentile.

$$P = \frac{n \cdot k}{100} \rightarrow \text{required percentile. [Rank Method]}$$

No. of data point  $\rightarrow$  out of total percentile

Rank starts from 1.

Mean & std.

Date

Median & interquartile.

Q. From a dataset containing 20 students find the 75<sup>th</sup> percentile.

→ first sort the data and assign rank. & the dataset  
 $\frac{20 \times 75}{100} \Rightarrow 15^{\text{th}}$  (at the 15<sup>th</sup> value we have 75<sup>th</sup> percentile.)

\* Visualize from slides.

Q. From a dataset containing 20 students find the 25<sup>th</sup> percentile.

$\frac{21 \times 25}{100} \Rightarrow 5.25^{\text{th}}$

↓

Now this rank is not available in our table.

∴ to cater this we will use the interpolation method.

Percentile  $\Rightarrow X_5 + (X_6 - X_5) \times \frac{0.25}{\text{whats inpoints}}$  → multiply with what's inpoints  
 $\Rightarrow 68 + (72 - 68) \times 0.25 \rightarrow 69$