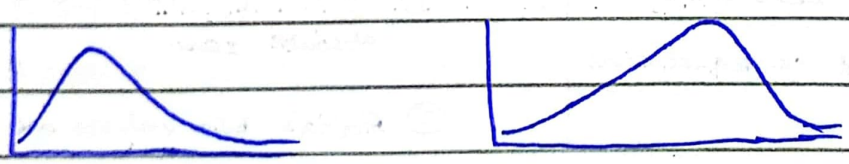
Lecture (4 & 5)

understanding statistical description in Data Science.

→ How data is spread accross different values
→ Gain insights from data
→ Identify data patterns
→ central value

Data Distribution:

① Symmetric distribution: (Both left and right parts are equal)
② Asymmetric distribution or skewed:
It can be positive and negative skewed
→ positive skew: Data spread at start, curve goes down slowly. Tail on right side
→ Negative skew: curve gradually increases, then abruptly closes. Tail on left side



Positive skew Negative skewed

→ Noise: Random unpredictable errors in the data
→ Bias (means data is leaning in one direction)
* Symmetric errors from incorrect assumptions made by model
→ Statistical description tells us about an outliers
→ Through statistical description we get to know how to clean our data.

Main types of statistical Description

→ Measure of central tendency

→ Measure of dispersion (shows how values vary from each other)

→ Measure of similarity

Measure of central tendency

➡ central tendency:

It's the value of feature/attribute that divide data into two halves

→ Typical value indicates the most common or average value in the distribution (central value)

➡ Measures of central tendency:

Mean (simple average) $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$ (con) Its sensitive to extreme values

Median (middle of sorted data) cons: computationaly cost

Mode (most frequent)

Trimmed Mean (It's modified mean where extreme values are removed before calculation) (max 2%.)

Midrange $\frac{(min + max)}{2}$ ∵ easy to calculate → use for brief estimate

→ If mean and median values are closer to each other then we can say its symmetrically distributed.

→ Mean, Median, mode also tells us wheather the data is symmetric or skewed

median = mean = mode (symmetric Distribution)

mode < median < mean (positive skew)

mean < median < mode (Negative Skew)