

Foundations of Statistics

Antonio Reyes Pérez (PhD)

Facultad de Lenguas y Letras

Universidad Autónoma de Querétaro

Data Processing Workshop

April 2025



Why do we need Statistics?

- To measure
- To represent
- To interpret
- To generalize



Two main approaches

- Descriptive Statistics
- Inferential Statistics



Descriptive Statistics

- It describes or summarizes features from a sample
- Focus is on quantitative data
- Based on measures of central tendency and measures of dispersion or variability



Inferential Statistics

- It learns from Descriptive Statistics
- Focus is on inferring properties to generalize patterns
- Based on Probability Theory

Measures of central tendency

- Mean
- Mode
- Median

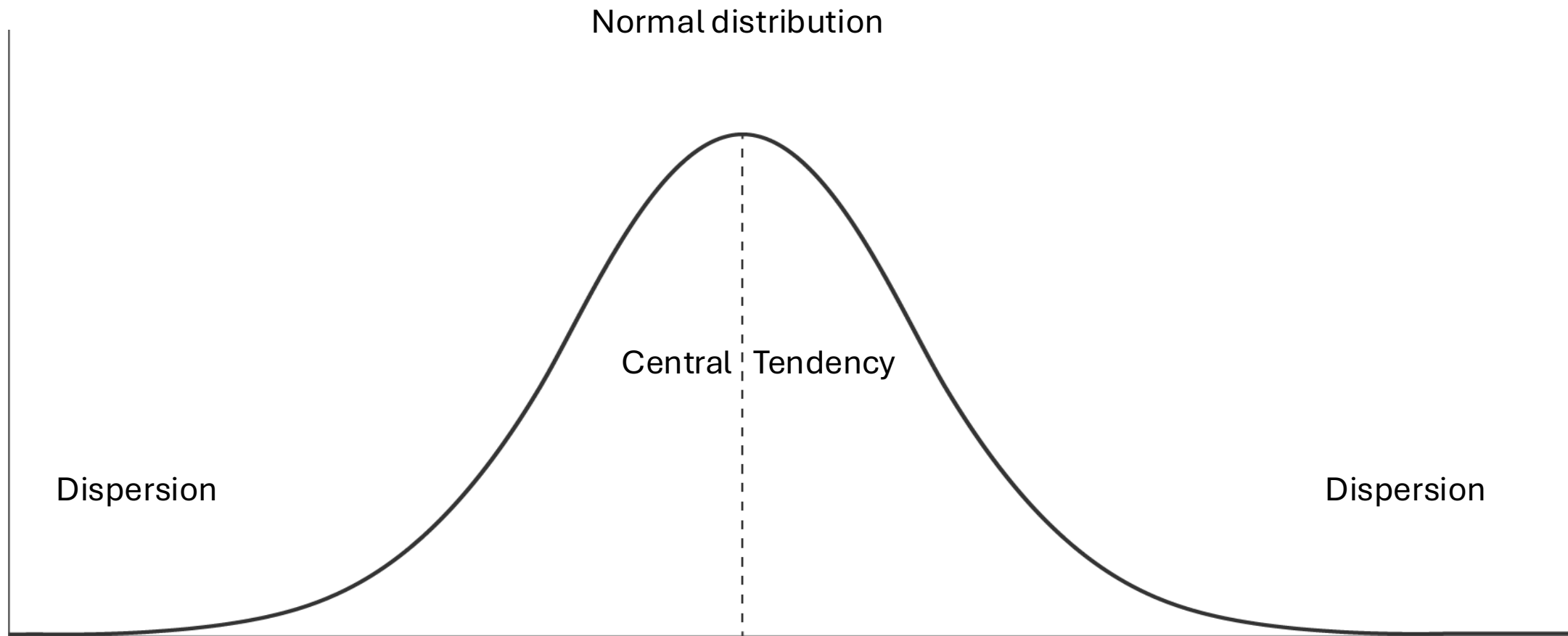
- *Frequency



Measures of dispersion

- Variance
- Standard deviation*
- How much a datum is deviating from the central tendency?
- *By obtaining the squared root, the values are comparable with the central tendency





A visual insight



Frequency

- Number of repetitions of N item
- Absolute frequency: number of times an item occurs
- Relative frequency: number of time an item occurs divided by the total number of items

Mean

- Average
- Sum of the numerical values of each observation, divided by the total number of observations
- Influenced by outliers (large or small)

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + \dots + x_n}{n}$$



$$\bar{X} = \frac{\sum X}{N}$$

Mode

- Value (item) that appears most often in a set of data
- Value that is most likely to be sampled
- Unimodal, bimodal, trimodal, multimodal



$$x = \operatorname{argmax}_{x_i} P(X = x_i)$$

Median

- Value between the higher half and the lower half
- The middle value (ordered values: smallest to greatest)
- Unlike the mean, it is not influenced by outliers

$n = \text{odd value}$

$$\text{med}(x) = x_{(n+1)/2}$$



$n = \text{even values}$

$$\text{med}(x) = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2}$$