MM 225 – AI and Data Science

Day 11: Descriptive Statistics - Numerical

Instructors: Hina Gokhale, MP Gururajan, N. Vishwanathan

22 AUGUST 2024

Outline

1. Why descriptive statistics?

2. Types of descriptive statistics
   a. Numerical Methods
   b. Graphical methods

Why Descriptive Statistics

- To summarise the data
- To get to know the data enough to describe it to someone

Different names:

a. Exploratory Data Analysis
b. Cross Examination of data

What is achieved?

a. Data insight
b. Data cleaning – from errors of various kind
c. Relationship among data

Numerical tools

Measure of Central Tendency

Measure of dispersion

Measure of skewness and kurtosis

Measures of Central Tendencies

Mean / Average

- Let $x_1, x_{2,} \ldots, x_n$ be n data points, then mean of the data is defined as

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

Mean provides the central value about which the data is spread out.

Drawing an equivalence from Physics Mean value of data is like centre of gravity of the matter

Measures of Central Tendencies

Median is the value which divides the data in two halves

- Let $x_1, x_{2,} \ \dots, x_n$ be n data points,
- Order the data values $\boldsymbol{x_{(1)} \leq x_{(2)} \leq \ \dots \leq \ x_{(n)}}$
- If the number of data points is odd, sample median is the value in the position (n+1)/2
- If the number of data points is even, sample median is the average of values in positions n/2 and (n+1)/2

Measures of Central Tendencies

Mode:

- Mode is a value in data that occurs with highest frequency
- It’s the most probable value of the data
- It is possible to have data that has more than one Mode value. Such a data is called multimodal.

Measures of Central Tendencies

Other statistics

- Percentiles
  - Order the data set in ascending order: $x_{(1)} \leq x_{(2)} \leq \ \ldots \leq \ x_{(n)}$
    - Then, P_1 is called 1st percentile if 1% of points lie below this value
    - Similarly, P_k is called kth percentile if k% of data points lie below this value, where $0 \leq k \leq 100$
- Quartiles
  - P_{25} is also called 1^{st} quartile Q_1
  - P_{75} is also called 3^{rd} quartile Q_3
  - P_{50} is median

Mean or Median?

Both the measures provide “middle” value of the data, so how do they compare?

- Median is robust against extreme values in the data,
- While Mean is affected by extreme value

Example: Let 8.0, 9.0, 10.0, 11.0, 12.0 be five data points.

- Mean = 10.0 and Median = 10.0
- Replace 12.0 by 18.0
  - Mean = 11.2, but the median = 10.0

Measure of Dispersion

Measures the spread of data

- Range: measures the total spread of the data
- Variance or Standard Deviation
  - Measures spread about mean / average value of the data
  - This measure is akin to second moment of matter in Physics

Interquartile Range

- Measures the spread about median value of the data

Measure of Dispersion

Range = M-m, where,

- M = max{x_1, x_2, …, x_n}
- m = min{x_1, x_2, …, x_n}

Variance

- $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)} = \frac{1}{(n-1)}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right]$
- Standard Deviation = S

Interquartile Range : $Q_3 - Q_1$

Skewness & Kurtosis

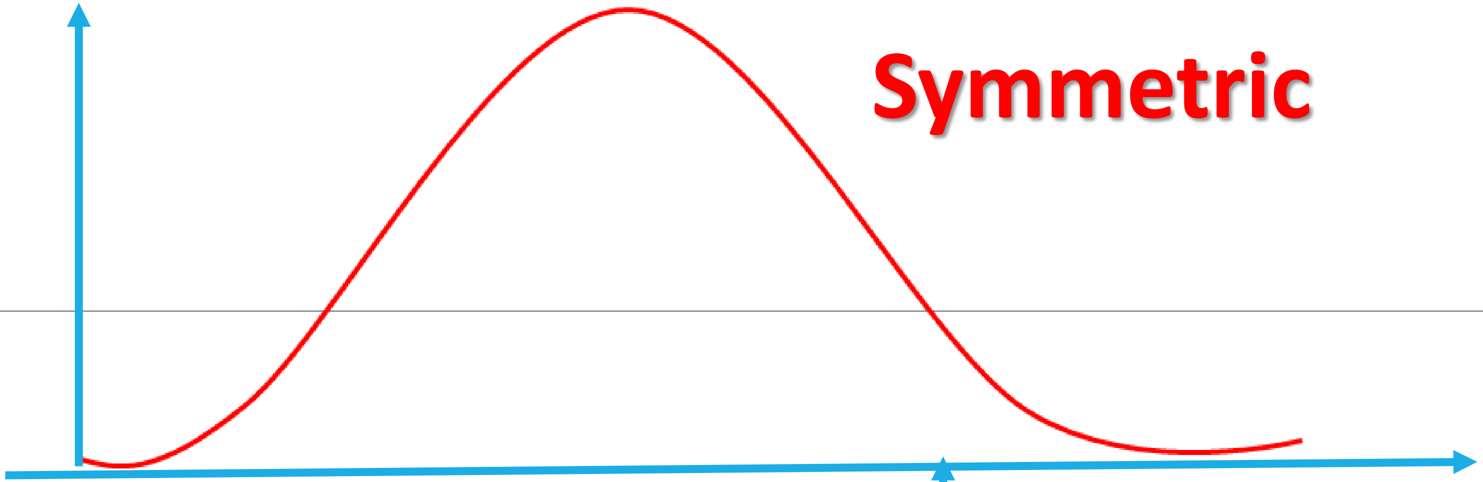
Let $x_1, x_{2,}\ \ldots, x_n$ be n data points,

- Then

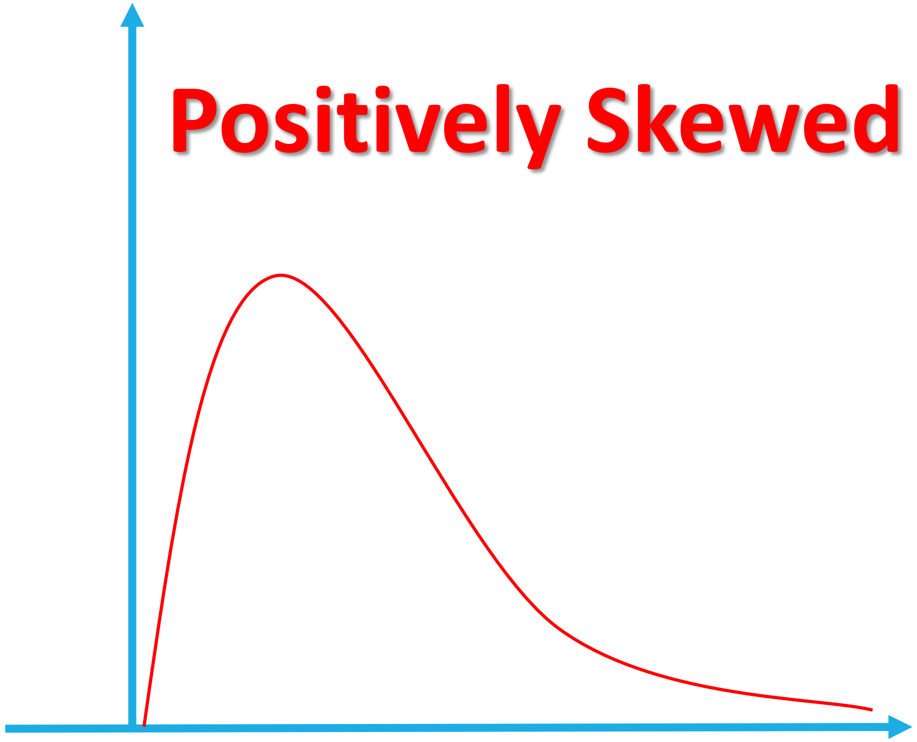
$$skewness = \frac{\sum(x_i - \bar{x})^3}{\left[\sum(x_i - \bar{x})^2\right]^{3/2}}$$

$$kurtosis = \frac{\sum(x_i - \bar{x})^4}{\left[\sum(x_i - \bar{x})^2\right]^2}$$

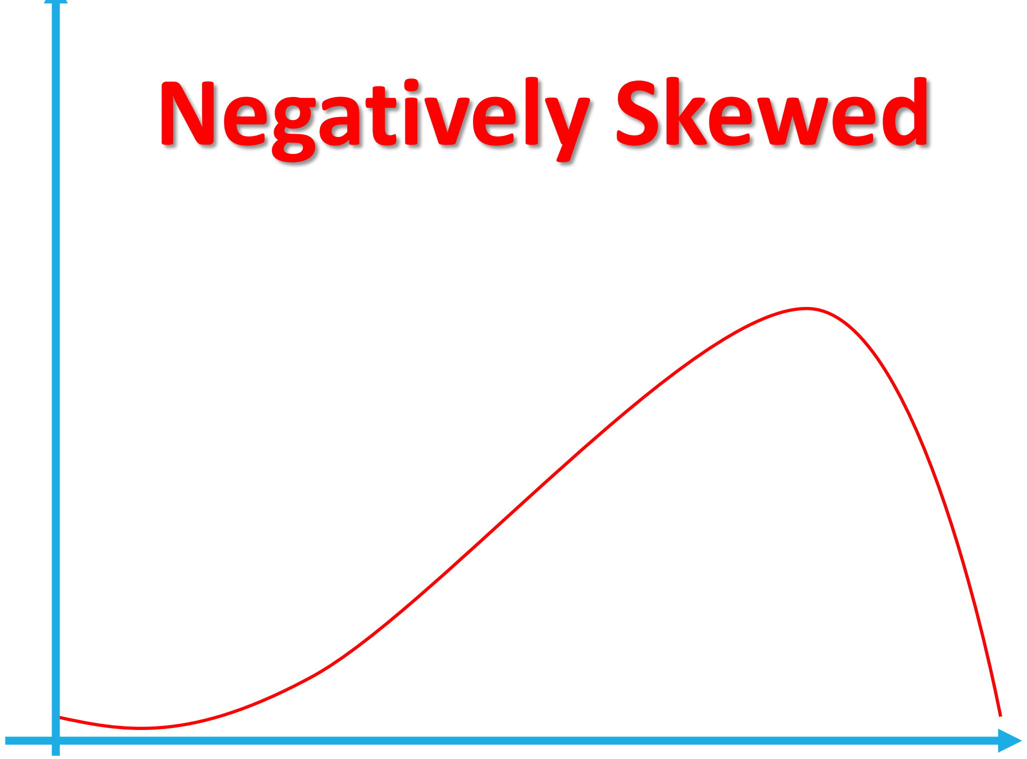
Skewness



Symmetric



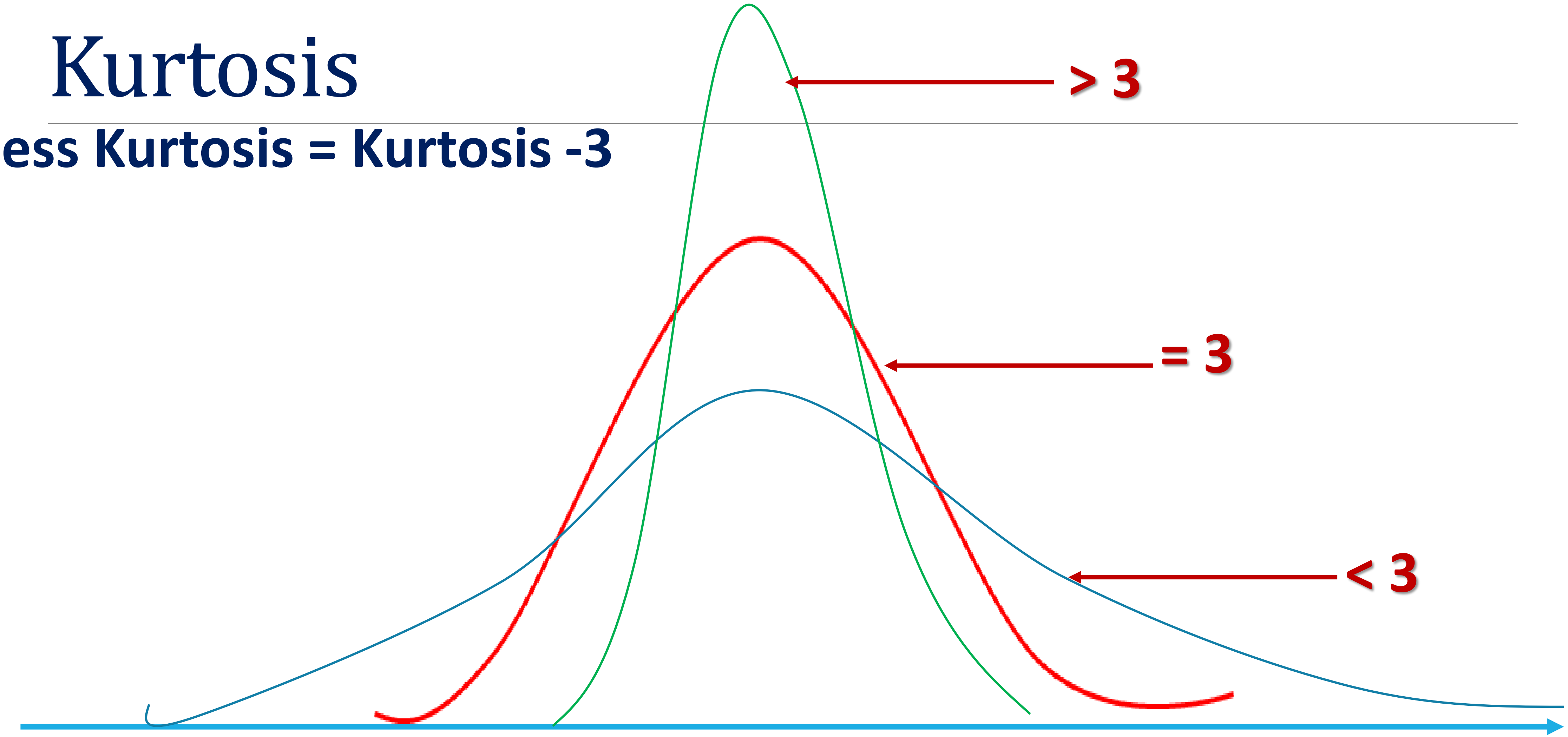
Positively Skewed



Negatively Skewed

Kurtosis

Excess Kurtosis = Kurtosis -3



Thank you…..