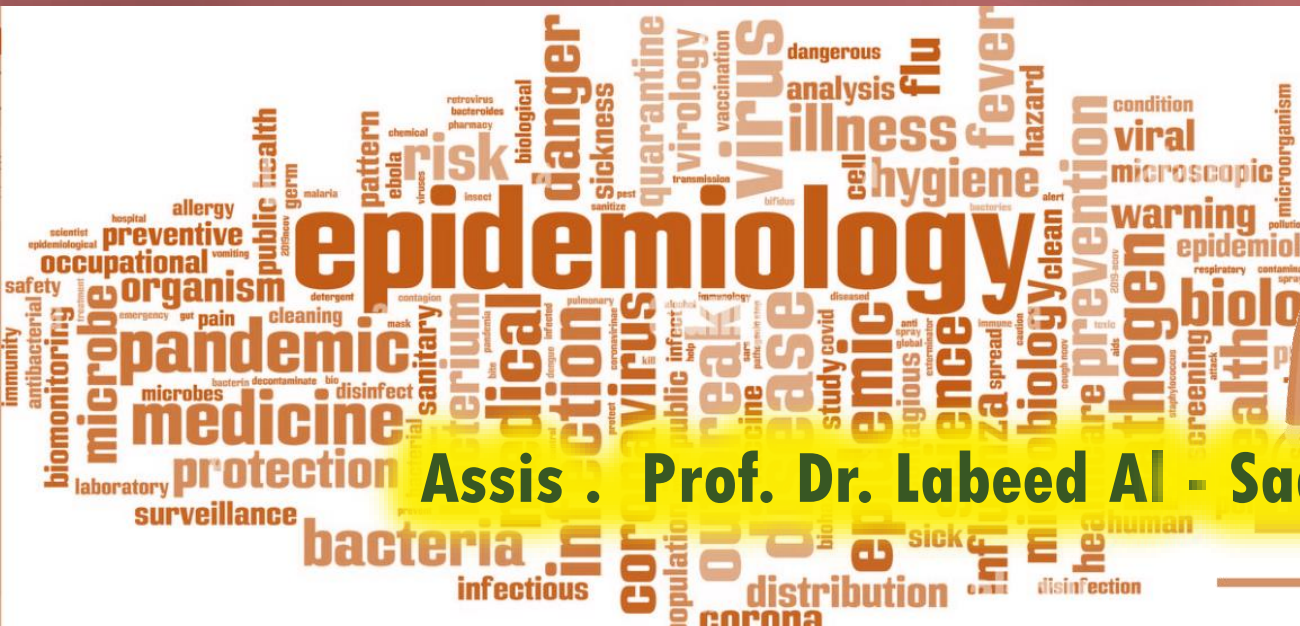
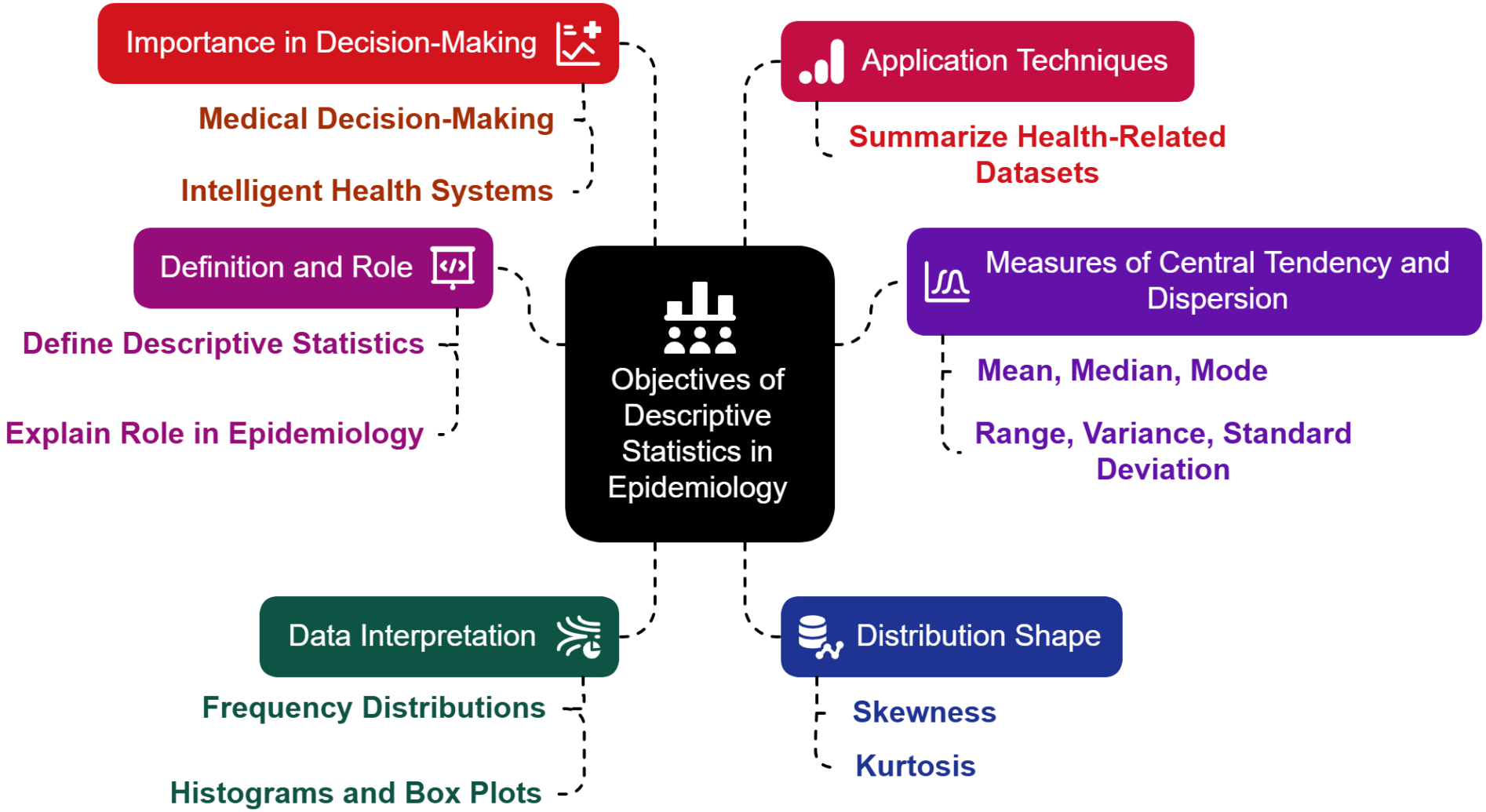
Epidemiology

Descriptive Statistics



Assis . Prof. Dr. Labeed Al - Saad

The objectives

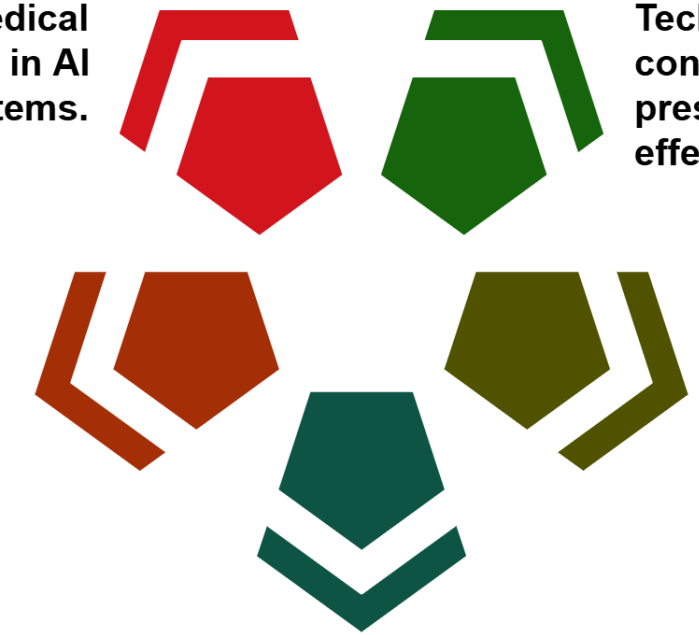


Introduction

What is Descriptive Statistic?
Why it is important?

AI Data Preprocessing
Preparing medical data for analysis in AI systems.

Summary Methods
Techniques used to condense and present data effectively.



Public Health Priorities
Identifying and prioritizing health issues for intervention.

Foundation for Analysis
Provides a base for more complex statistical evaluations.

Disease Pattern Understanding
Analyzing data to identify and understand disease trends.

Key Measures in Descriptive Statistics- Measures of Central Tendency

Measures of central tendency

Mean

Average of all data points,
Sensitive to outliers,
suitable for symmetric distributions

$$\bar{X} = \frac{\sum xi}{n} = \frac{x1 + x2 + x3 + \cdots + xn}{n}$$

$$\bar{X} = \frac{\sum fixi}{\sum fi}$$

Mode

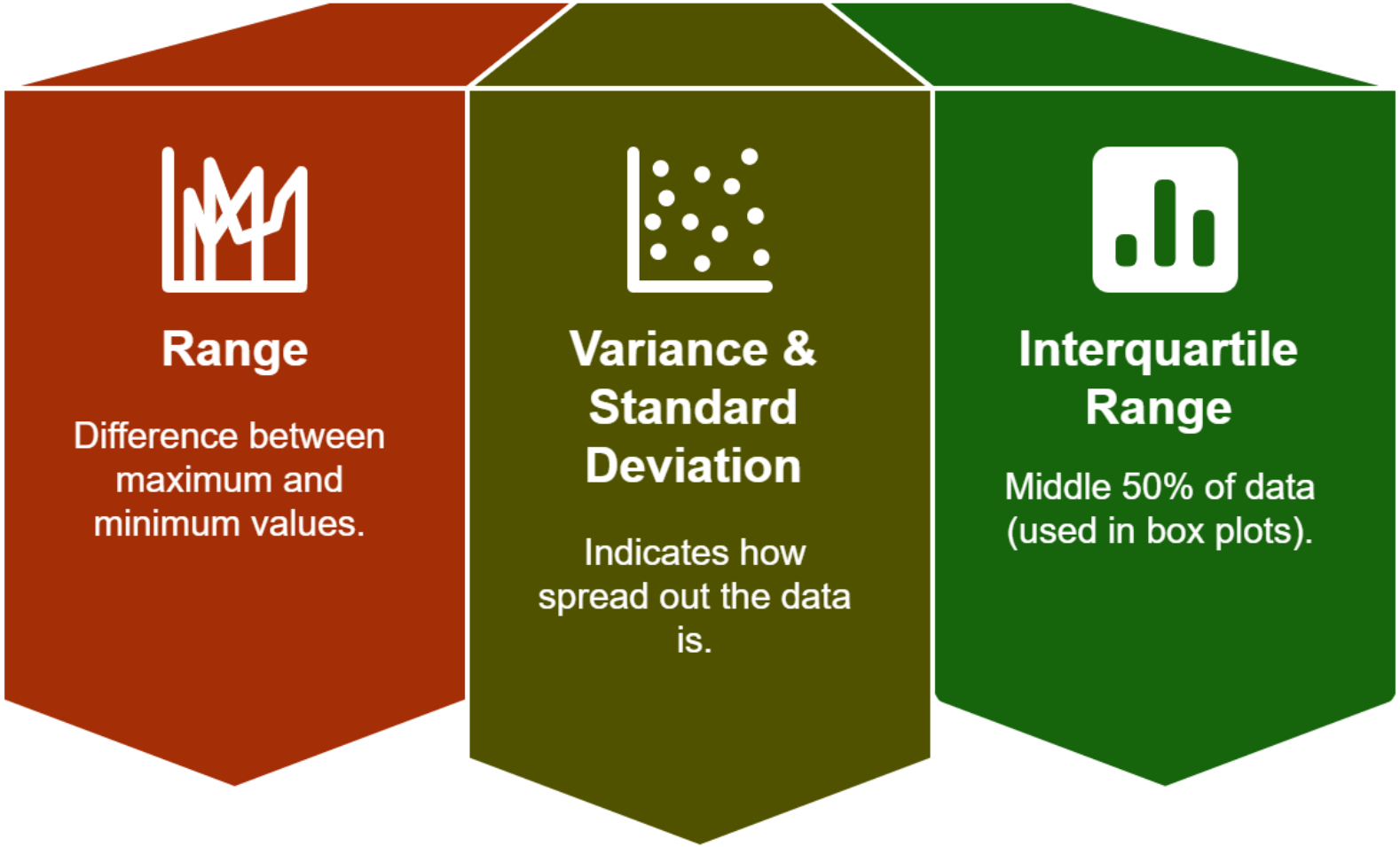
Middle value in an ordered dataset,
Best for categorical data, highlights most common value

Median

Most frequently occurring value,
Robust to outliers, ideal for skewed distributions

Example: Analyzing the average age of COVID-19 patients in a hospital dataset.

Key Measures in Descriptive Statistics- Measures of Dispersion (Variability)



$$R = Max. - Min.$$

$$S^2 = \frac{\sum X_i^2 - \frac{\sum (X_i)^2}{n}}{n-1}$$

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

$$Sd = \sqrt{\frac{\sum X_i^2 - \frac{\sum (X_i)^2}{n}}{n-1}}$$

$$Sd = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

$$IQR = Q3 - Q1$$

Application: Comparing blood pressure variability between two patient groups.

Data Visualization in Epidemiology

Visualization Techniques

Utilizing visual tools to enhance data comprehension.

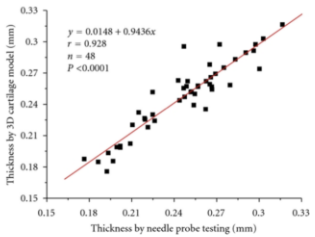
Data Summarization

Condensing data for easier understanding and analysis.

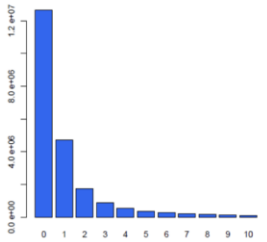
Meaningful Presentation

Presenting data in an informative and engaging manner.

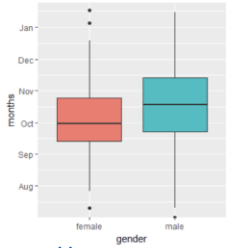
"The variables are related"



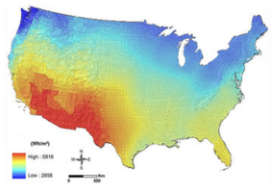
"The distribution is skewed"



"The groups differ in Y"



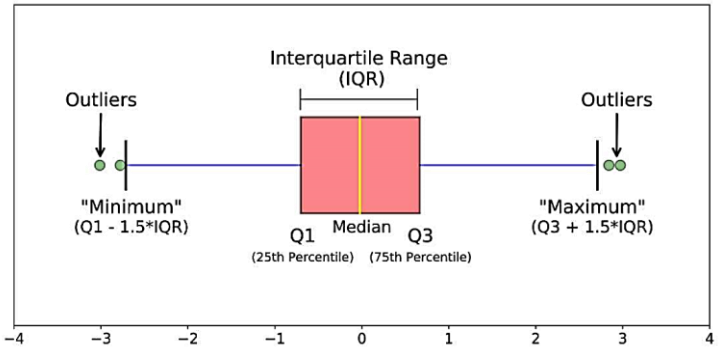
"Affects the Southwest"



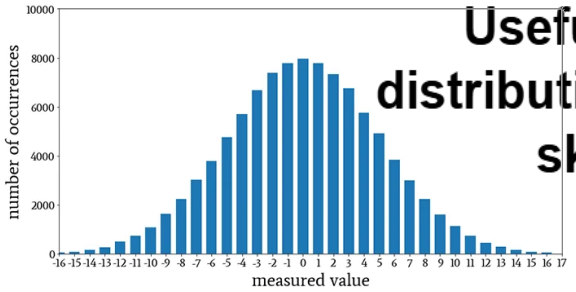
https://bookdown.org/jbrophy115/bookdown-clinepi/vis.html

Data Visualization in Epidemiology

Data visualization methods



Frequency Distributions & Histograms

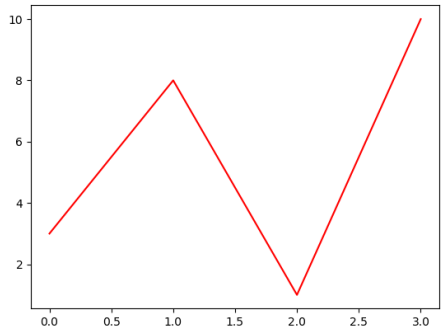


Useful for identifying distribution patterns and skewness in data.

Box Plots

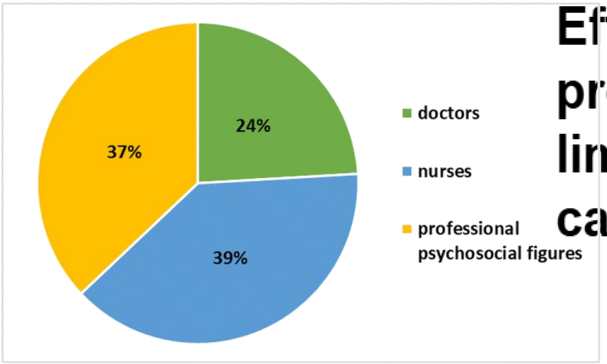
Ideal for displaying median, quartiles, and outliers, especially for group comparisons.

Line Plots



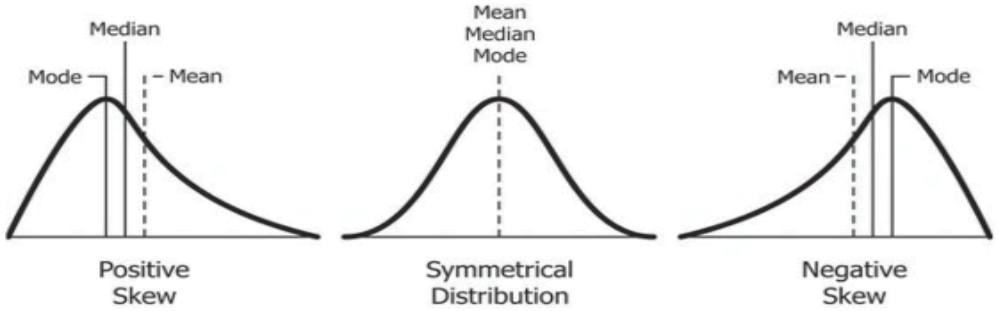
Best for trend analysis over time, such as tracking disease incidence.

Pie Charts



Effective for showing proportional data, though limited by the number of categories.

Shape of the Distribution - Skewness



$$Skewness = \frac{\sum_{i=1}^{n}(Xi - \bar{X})^3}{n \times \sigma^3}$$

Skewness

Positive Skew
- Mean > Median
- Right-tailed
- (> 0.5)

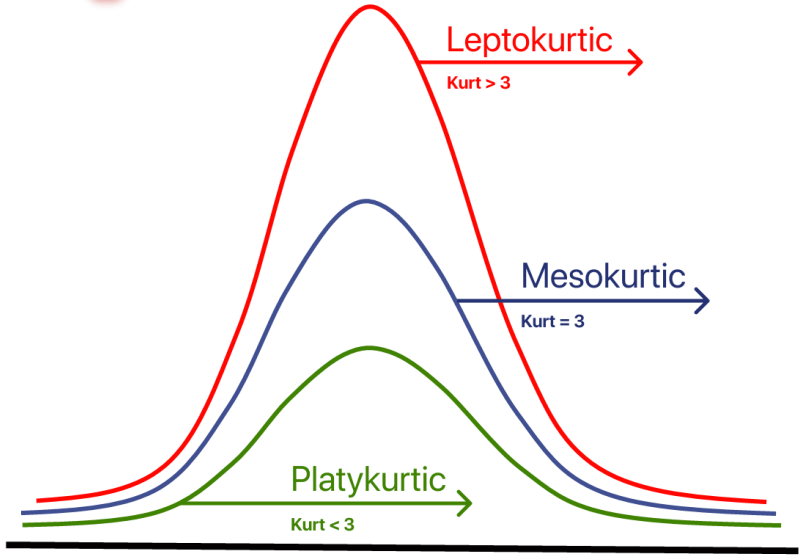
Zero Skew
- Normal Distribution
- (-0.5 – 0.5)

Negative Skew
- Mean < Median
- Left-tailed
- (< -0.5)

Importance in Epidemiology
- Identifying Disease Patterns
- Statistical Test Choice

Example: Hospital length of stay data is often right-skewed (most patients stay briefly, but a few stay much longer).

Shape of the Distribution - Kurtosis



Understanding Kurtosis in Epidemiology

- Types of Kurtosis
  - Leptokurtic
  - Platykurtic
  - Mesokurtic
- Kurtosis
- Importance in Epidemiology
  - Detecting Outliers
  - Risk Assessment Models

Leptokurtic

Indicates heavy tails and extreme values, useful for detecting outliers.

Platykurtic

Suggests lighter tails and flatter peaks, indicating less extreme values.

Mesokurtic

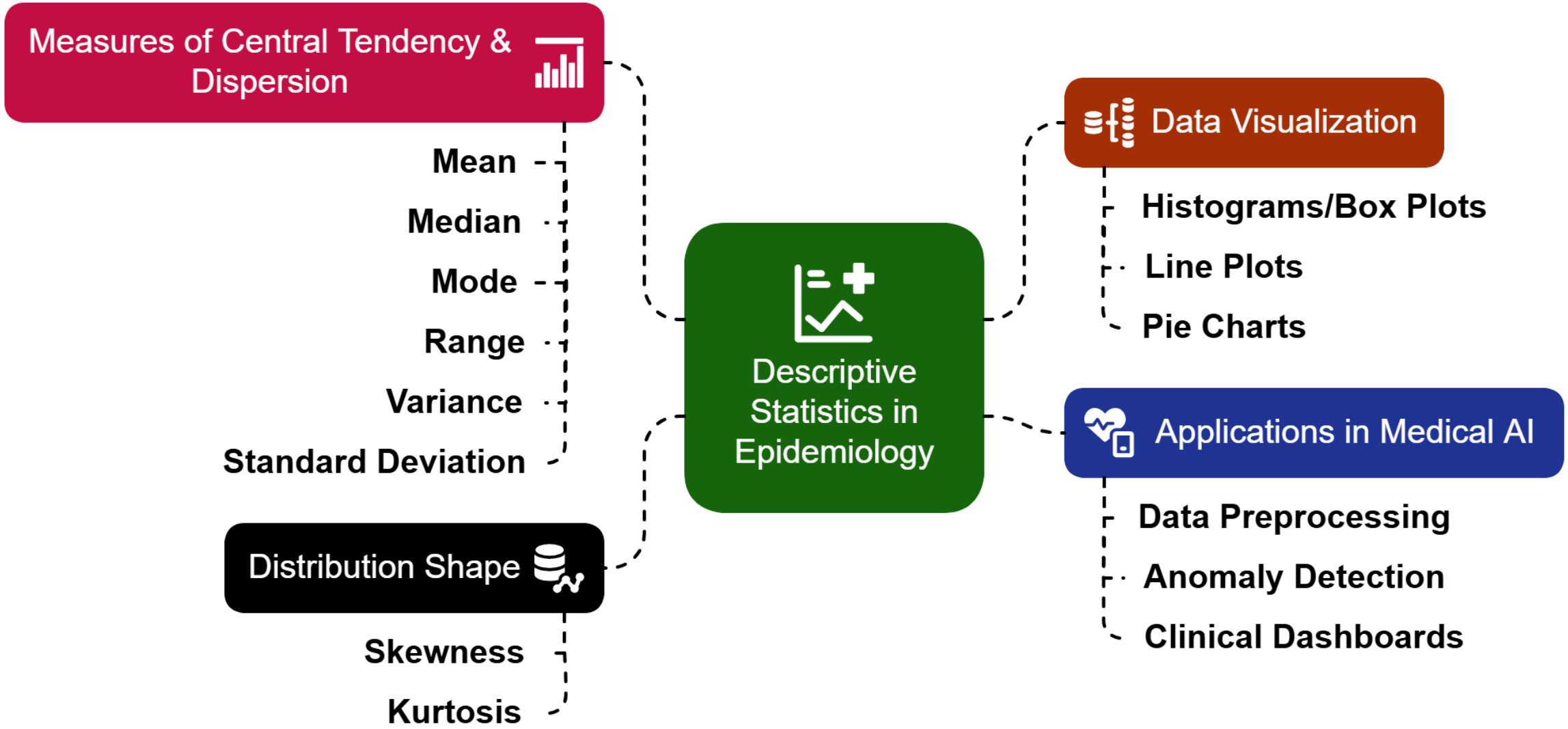
Represents normal distribution characteristics, serving as a baseline.

$$Kurtosis = \frac{\sum_{i=1}^{n}(Xi - \bar{X})^4}{n \times \sigma^4}$$

Example: Blood glucose levels in diabetic patients may exhibit high kurtosis if extreme highs/lows are common.

Summary

Descriptive Statistics in Epidemiology



Comprehensive Questions

Section 1: Fundamentals of Descriptive Statistics

1. **Define descriptive statistics and explain its importance in epidemiology.** How does it differ from inferential statistics?
2. **Compare and contrast** the three measures of central tendency (mean, median, mode). When would you use each in medical data analysis?
3. **Why is the median often preferred over the mean** when analyzing skewed health data (e.g., hospital stay durations)? Provide an example.

Section 2: Measures of Dispersion & Distribution Shape

1. **Explain the concept of standard deviation** in epidemiological studies. How does it help assess disease variability in a population?
2. **A dataset on patient blood pressure has a high variance.** What does this imply about the population's health status?
3. **Differentiate between positive and negative skewness** in disease incidence data. Provide real-world examples of each.
4. **How does kurtosis influence the interpretation of medical data?** Discuss its relevance in detecting rare health events.

Comprehensive Questions

Section 3: Data Visualization Techniques

1. **When would you choose a histogram over a box plot?** Compare their utility in visualizing patient age distributions.
2. **Critique the use of pie charts in epidemiology.** When are they effective, and when should they be avoided?
3. **A line plot shows a sudden spike in flu cases every December.** What epidemiological insights can you derive from this pattern?
4. **Design a data visualization strategy** to compare the prevalence of diabetes across three age groups. Justify your choice of graph(s).

Section 4: Applications in AI & Medical Systems

1. **How can skewness and kurtosis affect the performance** of an AI model predicting heart disease risk?
2. **Describe a scenario where descriptive statistics fail** to capture critical insights in a health dataset. What complementary methods would you use?
3. **Explain how box plots can aid in outlier detection** in electronic health records (EHRs). Why is this important for machine learning?

Comprehensive Questions

Section 3: Data Visualization Techniques

1. **When would you choose a histogram over a box plot?** Compare their utility in visualizing patient age distributions.
2. **Critique the use of pie charts in epidemiology.** When are they effective, and when should they be avoided?
3. **A line plot shows a sudden spike in flu cases every December.** What epidemiological insights can you derive from this pattern?
4. **Design a data visualization strategy** to compare the prevalence of diabetes across three age groups. Justify your choice of graph(s).

Section 4: Applications in AI & Medical Systems

1. **How can skewness and kurtosis affect the performance** of an AI model predicting heart disease risk?
2. **Describe a scenario where descriptive statistics fail** to capture critical insights in a health dataset. What complementary methods would you use?
3. **Explain how box plots can aid in outlier detection** in electronic health records (EHRs). Why is this important for machine learning?


THANK YOU!