2024-1

Multivariate Data Analysis

Youjin Shin
(yj.shinn@catholic.ac.kr)

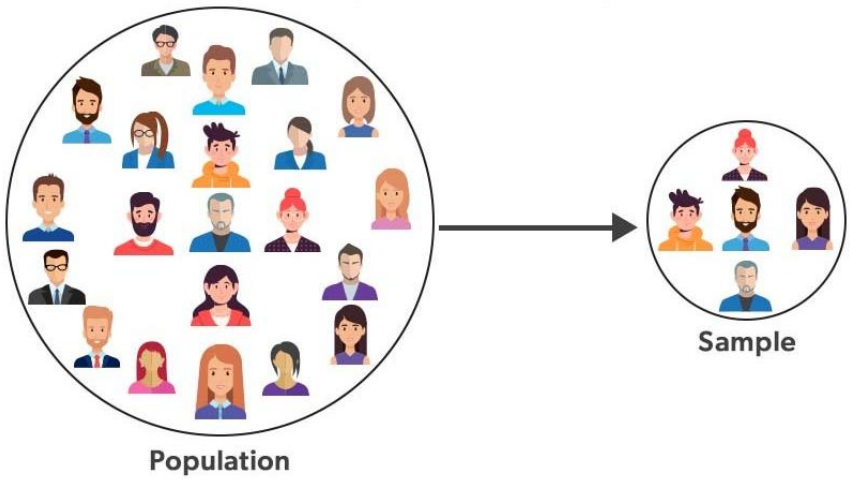
Population and Sample

◆ **Population:**

- Refers to the entire group that you are interested in studying. The population is the total set of observations that can be made. For example, if you're studying the average height of adult men in a country, the population would be all adult men in that country.

◆ **Sample:**

- A subset of the population, selected for the actual study. Sampling is used because it is often impractical, too expensive, or physically impossible to collect data from every member of the population.



Population and Sample

◆ Why Sampling is Necessary

- Feasibility: In many cases, it's not feasible to study the entire population due to its size. For example, it would be impractical to measure the height of every adult man in a country.

- Cost and Time Efficiency: Sampling allows researchers to obtain results faster and with less expense compared to studying the entire population.

- Data Quality: In some cases, focusing on a sample can allow for more accurate data collection methods, as resources can be concentrated on a smaller group.

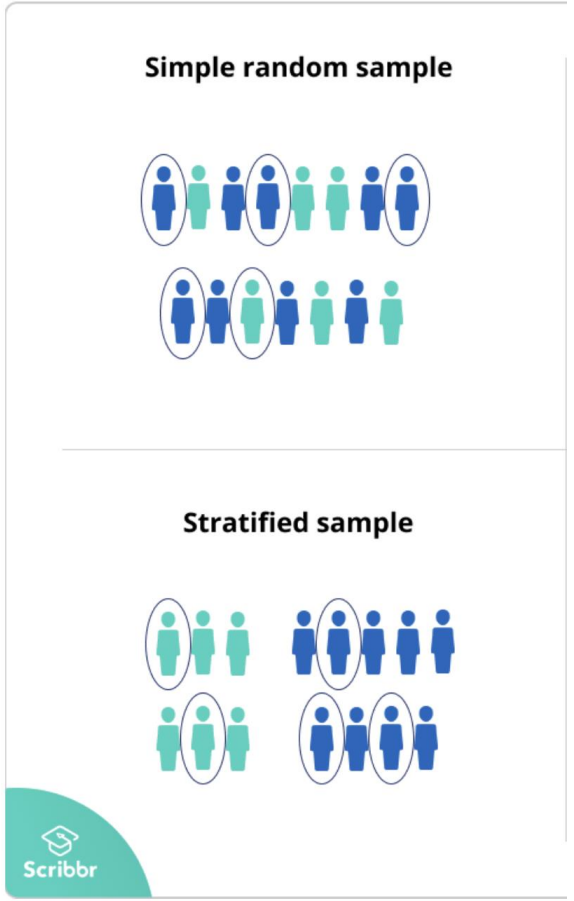
Types of Sampling

1. Random Sampling

- Definition: A sampling method where each member of the population has an equal chance of being included in the sample.
- Advantages: Simple and straightforward, ensuring that the sample is representative of the population.
- Disadvantages: Can be challenging to implement with very large populations.

2. Stratified Sampling

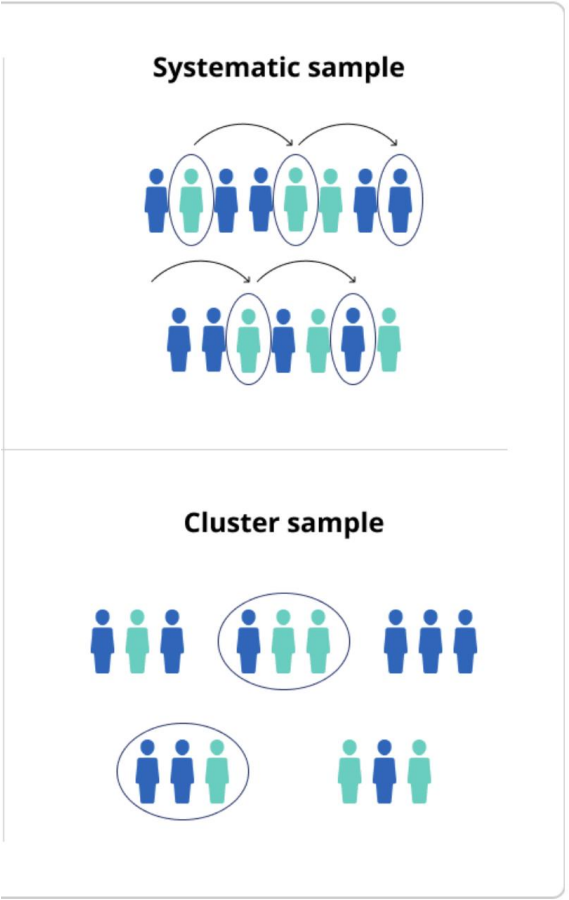
- Definition: The population is divided into smaller groups (strata) that share similar characteristics, and samples are drawn from each of these strata independently.
- Advantages: Ensures that specific subgroups are fairly represented in the sample.
- Disadvantages: Determining the criteria for stratification can be challenging.



Types of Sampling

3. Systematic Sampling

- Definition: Selects samples based on a fixed, periodic interval (e.g., every 10th member) after randomly choosing a starting point within the first interval.
- Advantages: Easy to implement and less complex than random sampling.
- Disadvantages: The periodic selection process can introduce bias if there's a pattern that aligns with the sampling interval.



4. Cluster Sampling

- Definition: Divides the population into clusters (groups) and randomly selects a number of these clusters; then all members within chosen clusters are sampled.
- Advantages: Cost-effective and time-saving for large populations.
- Disadvantages: May reduce the representativeness of the sample compared to stratified sampling.

Descriptive Statistics

◆ Key Components of Descriptive Statistics

1. Measures of Central Tendency: These are measures that describe the central position of a frequency distribution for a data set.
   - **Mean (Average)**: The sum of all measurements divided by the number of observations in the data set.
   - **Median**: The middle value when the data set is ordered from the least to the greatest. If there is an even number of observations, it is the average of the two middle numbers.
   - **Mode**: The most frequently occurring value(s) in a data set.

2. Measures of Spread (Variability): These indicators show how spread out the data points are.
   - **Range**: The difference between the highest and the lowest values in the data set.
   - **Interquartile Range (IQR):** The range of the middle 50% of the data points, calculated as the difference between the 75th percentile (Q3) and the 25th percentile (Q1).
   - **Variance**: The average of the squared differences from the Mean.
   - **Standard Deviation**: A measure of the amount of variation or dispersion in a set of values.

3. Measures of Shape: These describe the shape of the data distribution.
   - **Skewness**: A measure of the asymmetry of the probability distribution.
   - **Kurtosis**: A measure of the "tailedness" of the probability distribution.

Descriptive Statistics

◆Suppose we have a data set representing the ages of 10 individuals:

◆Ages: 23, 26, 20, 29, 24, 22, 27, 28, 22, 24

Mean (Average):

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}$$

$$\text{Mean} = \frac{23+26+20+29+24+22+27+28+22+24}{10}$$

$$\text{Mean} = \frac{245}{10} = 24.5$$

Median:

For the median, we first sort the data:

$$\text{Sorted Data} = [20, 22, 22, 23, 24, 24, 26, 27, 28, 29]$$

Since we have an even number of observations (10), the median is the average of the 5th and 6th values.

$$\text{Median} = \frac{24+24}{2} = 24$$

Mode:

The mode is the most frequently occurring value(s). In this dataset, 22 and 24 each appear twice, making them the modes.

Descriptive Statistics

◆Suppose we have a data set representing the ages of 10 individuals:

◆Ages: 23, 26, 20, 29, 24, 22, 27, 28, 22, 24

Variance:

$$\text{Variance} = \frac{\sum(x_i - \text{Mean})^2}{n}$$

$$\text{Variance} = \frac{\sum(x_i - 24.5)^2}{10} = 7.65$$

Standard Deviation:

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

$$\text{Standard Deviation} = \sqrt{7.65} \approx 2.77$$

The Interquartile Range (IQR)

◆ Just as the median is the point that divides a distribution in half, quartiles are the points that divide a distribution into quarters. We will use the symbols Q_1, Q_2 and Q_3 to represent the quartiles. Note that $Q_2 = M$, the median.

◆ **The Interquartile range**

➢ The interquartile range(IQR) is defined to be the spread of the middle 50% of data values, so that:

$$IQR = Q_3 - Q_1$$

◆ To calculate the IQR, it is necessary to first calculate the quartiles Q_1 and Q_3. In principle, this is straight forward as:

➢ Q_1 is the midpoint of the lower half of the data values

➢ Q_3 is the midpoint of the upper half of the data values

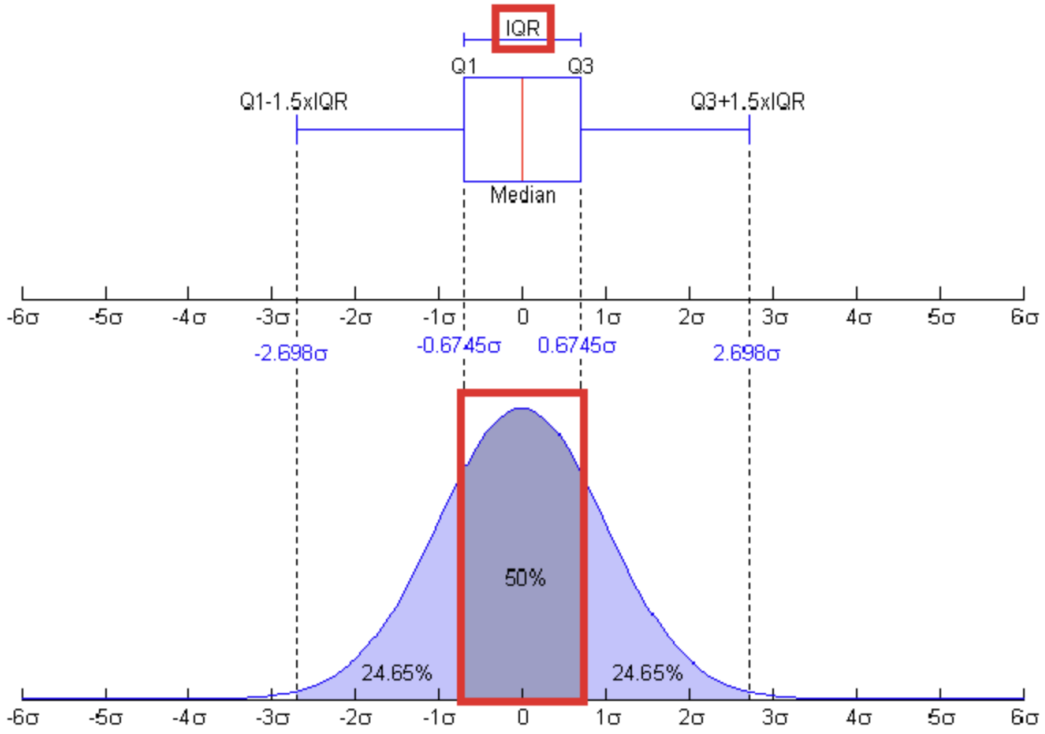
* if the data has been ordered, the computation of the quartiles is relatively straightforward.

The Interquartile Range (IQR)

◆**Suppose we have a data set representing the ages of 10 individuals:**

◆**Ages: 23, 26, 20, 29, 24, 22, 27, 28, 22, 24**

$$\text{Sorted Data} = [20, 22, 22, 23, 24, 24, 26, 27, 28, 29]$$



Range:

$$\text{Range} = \text{Maximum value} - \text{Minimum value}$$

$$\text{Range} = 29 - 20$$

Interquartile Range (IQR):

IQR is the difference between the 75th percentile (Q3) and the 25th percentile (Q1).

$Q1 = 22$

$Q3 = 27$

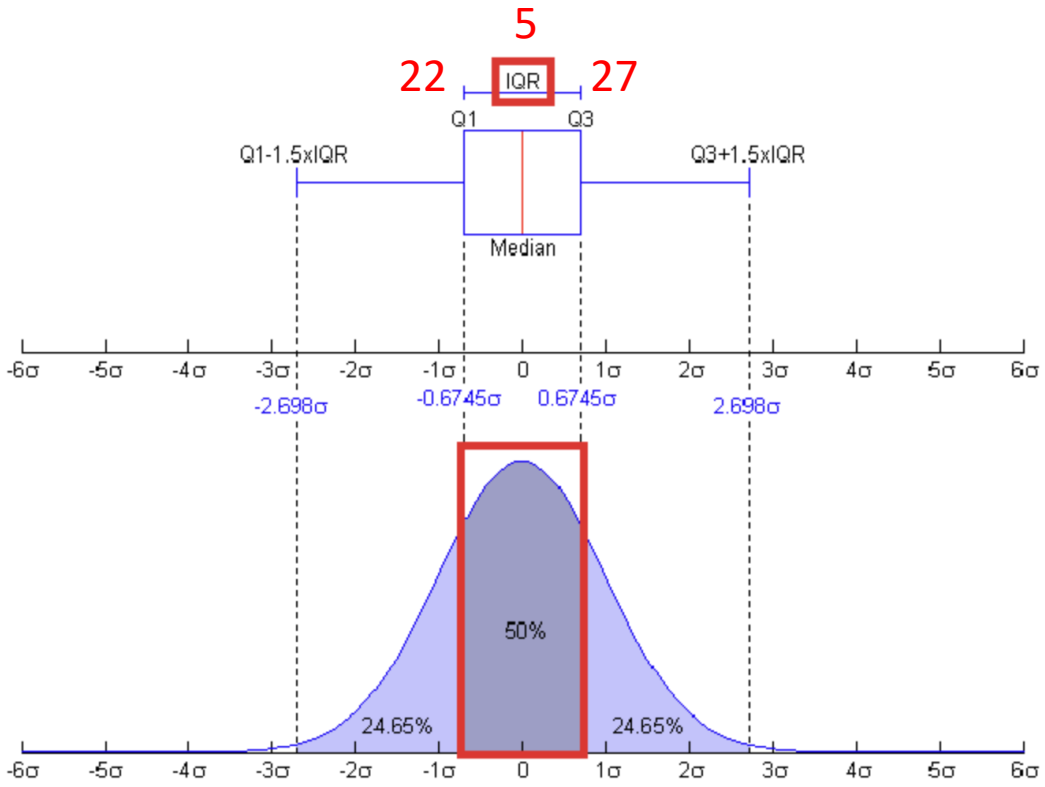
$\text{IQR} = Q3\text{-}Q1 = 5$

The Interquartile Range (IQR)

◆**Suppose we have a data set representing the ages of 10 individuals:**

◆**Ages: 23, 26, 20, 29, 24, 22, 27, 28, 22, 24**

$$\text{Sorted Data} = [20, 22, 22, 23, 24, 24, 26, 27, 28, 29]$$



Range:

$$\text{Range} = \text{Maximum value} - \text{Minimum value}$$

$$\text{Range} = 29 - 20$$

Interquartile Range (IQR):

IQR is the difference between the 75th percentile (Q3) and the 25th percentile (Q1).

$Q1 = 22$

$Q3 = 27$

$\text{IQR} = Q3\text{-}Q1 = 5$

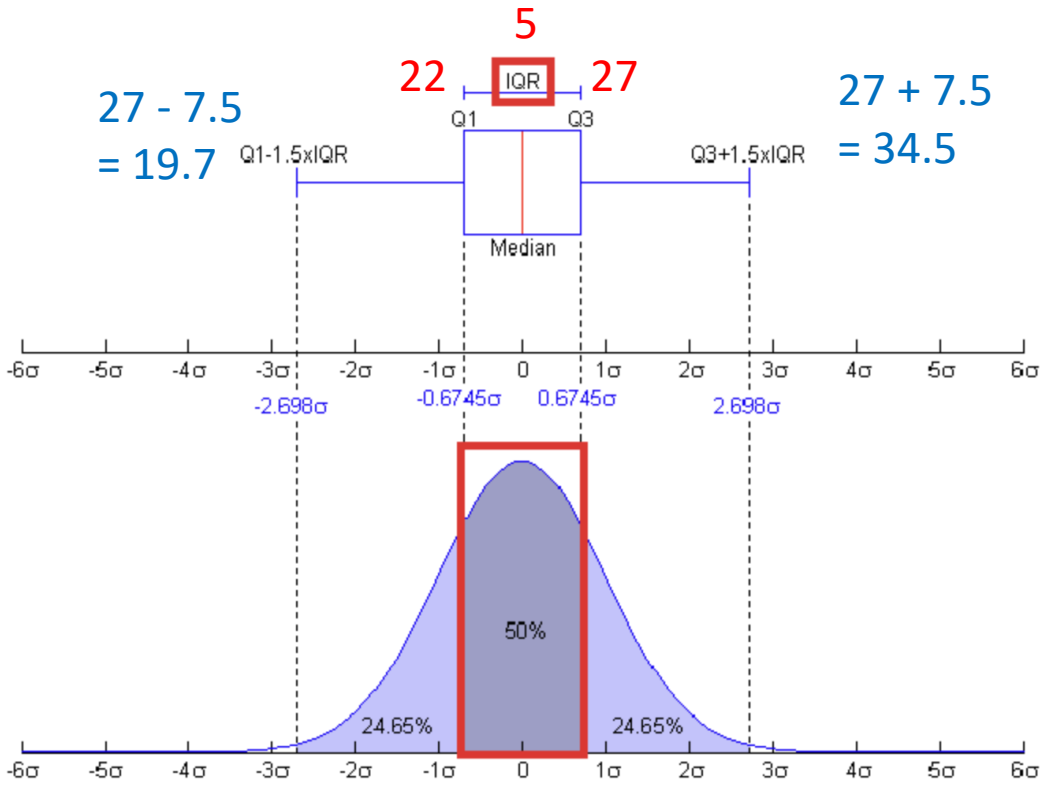
The Interquartile Range (IQR)

1.5 * IQR
= 1.5 * 5
= 7.5

◆**Suppose we have a data set representing the ages of 10 individuals:**

◆**Ages: 23, 26, 20, 29, 24, 22, 27, 28, 22, 24**

$$\text{Sorted Data} = [20, 22, 22, 23, 24, 24, 26, 27, 28, 29]$$



Range:

$$\text{Range} = \text{Maximum value} - \text{Minimum value}$$

$$\text{Range} = 29 - 20$$

Interquartile Range (IQR):

IQR is the difference between the 75th percentile (Q3) and the 25th percentile (Q1).

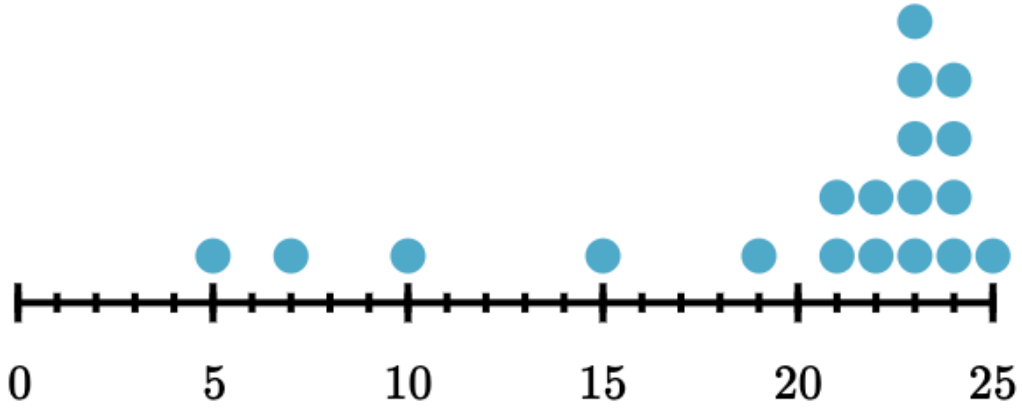
$Q1 = 22$

$Q3 = 27$

$\text{IQR} = Q3\text{-}Q1 = 5$

The Interquartile Range (IQR)

- **QUIZ: Suppose we have a data set representing the ages of 19 individuals:**
- **Ages (sorted):** $5, 7, 10, 15, 19, 21, 21, 22, 22, 23, 23, 23, 23, 23, 24, 24, 24, 24, 25$

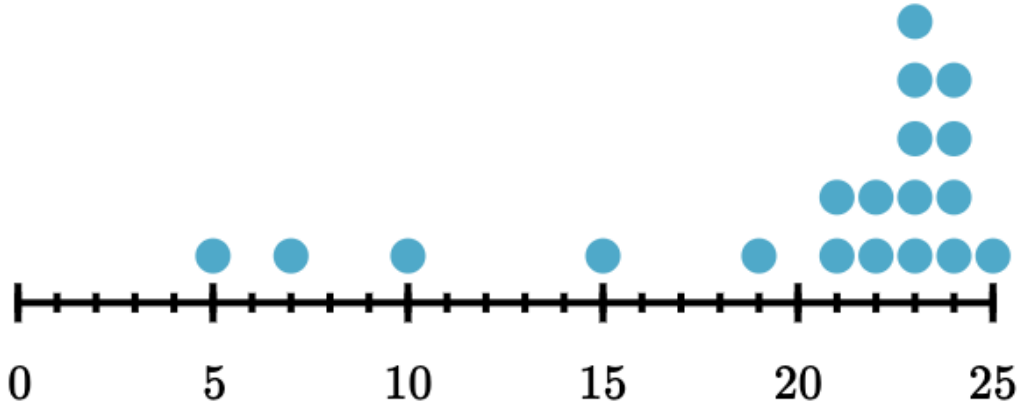


https://ko.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/identifying-outliers-iqr-rule

- Median =
- Q1 =
- Q3 =
- IQR =
- Lower Bound (coefficient = 1.5) =
- Upper Bound (coefficient = 1.5) =
- How many outliers below the lower bound =
- How many outliers over the upper bound =
- If the coefficient = 2.0, how many outliers are there? =

The Interquartile Range (IQR)

- **QUIZ: Suppose we have a data set representing the ages of 19 individuals:**
- **Ages (sorted):** $5, 7, 10, 15, 19, 21, 21, 22, 22, 23, 23, 23, 23, 23, 24, 24, 24, 24, 25$



https://ko.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/identifying-outliers-iqr-rule

- Median = 23
- Q1 = 19
- Q3 = 24
- IQR = 5
- Lower Bound (coefficient = 1.5) = 11.5
- Upper Bound (coefficient = 1.5) = 31.5
- How many outliers below the lower bound = 3
- How many outliers over the upper bound = 0
- If the coefficient = 2.0, how many outliers are there? = 2

Skewness

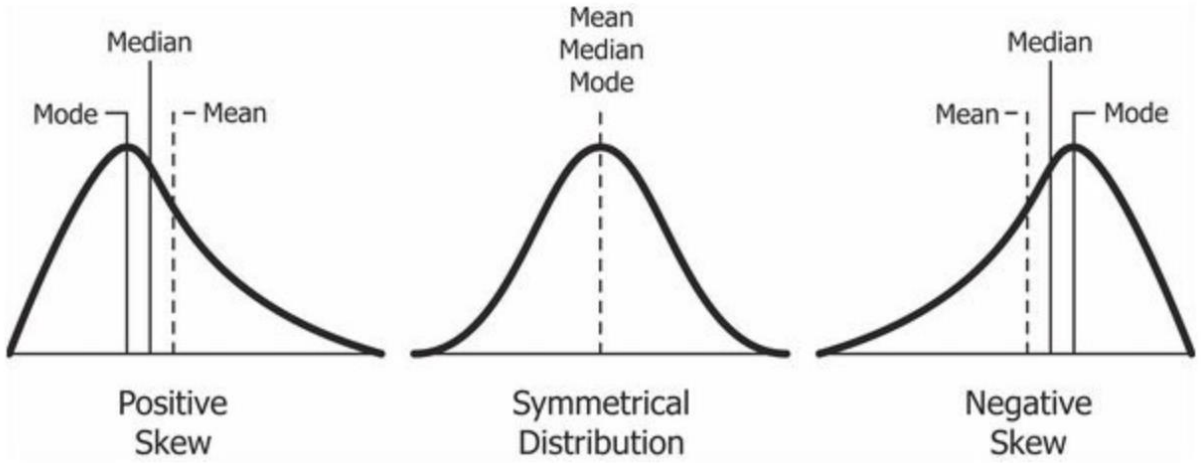
◆ Skewness

Skewness measures the asymmetry of the distribution around its mean. A distribution is:

- ➢ Symmetric (Skewness = 0): The left and right sides of the graph are mirror images of each other.
- ➢ Positively Skewed (Skewness > 0): The right tail is longer or fatter than the left tail. Most of the data is concentrated on the left. Mean and Median are greater than Mode.
- ➢ Negatively Skewed (Skewness < 0): The left tail is longer or fatter than the right tail. Most of the data is concentrated on the right. Mean and Median are smaller than Mode.

◆ Graphical Illustrations for Skewness:

- ➢ Symmetric Distribution: Bell-shaped, like a normal distribution.
- ➢ Positive Skewness: Tail extends to the right.
- ➢ Negative Skewness: Tail extends to the left.



Kurtosis

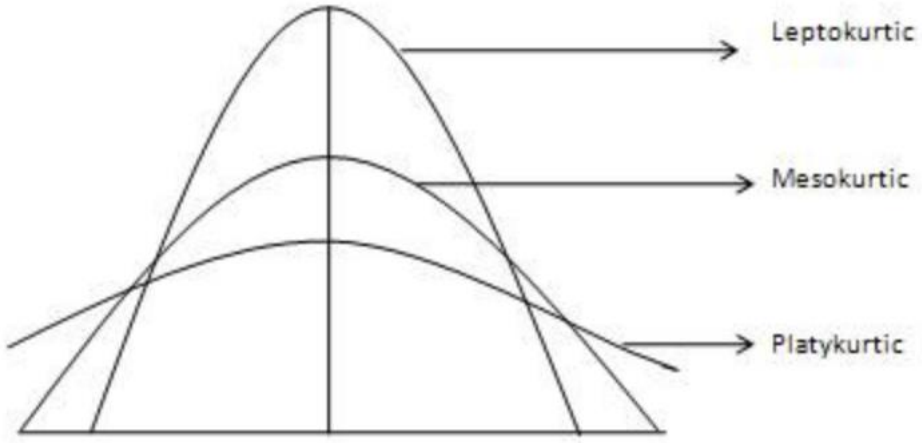
◆ Kurtosis

Kurtosis measures the "tailedness" of the distribution or the sharpness of the peak of a distribution.

- Mesokurtic (Kurtosis ≈ 3): The distribution has a similar kurtosis to a normal distribution, indicating a moderate peak and tails.
- Leptokurtic (Kurtosis > 3): The distribution has heavy tails and a sharp peak, indicating that data are more clustered around the mean and have frequent extreme values.
- Platykurtic (Kurtosis < 3): The distribution has light tails and a flat peak, indicating that data are spread out more evenly than in a normal distribution, with fewer extreme values.

◆ Graphical Illustrations for Kurtosis

- Mesokurtic: Similar to the normal distribution curve.
- Leptokurtic: More peaked than a normal distribution, with fatter tails.
- Platykurtic: Flatter peak and thinner tails compared to a normal distribution.



Data Scaling

- Scaling data is indeed a critical step in many data analysis and machine learning workflows, especially when dealing with features (or variables) that are measured on different scales. Without scaling, features with larger numerical ranges could dominate those with smaller ranges in the analysis, leading to biased results. Here's why scaling is essential:
  - Uniformity: Scaling ensures that all features have the same scale, making it easier to compare and combine them. This is particularly important in algorithms that measure distances between data points, such as k-nearest neighbors (KNN) and k-means clustering, where features on larger scales can disproportionately influence the outcome.
  - Improved Algorithm Performance: Many machine learning algorithms converge faster when features are on similar scales, particularly gradient descent-based algorithms used in neural networks and logistic regression. Scaling can lead to a smoother and quicker convergence because it ensures that all features contribute equally to the result.
  - Balanced Regularization: In regularization methods (like LASSO and Ridge regression), scaling is crucial because it applies penalties uniformly across all features. Without scaling, features with larger values might be penalized more than those with smaller values, irrespective of their contribution to the model's predictive power.

Data Scaling

- (EX) Who's better

	Steve	Tony
GPA	4.4	2.0
TOEIC	910	920

Data Scaling

◦ (EX) Who is better?

	Steve	Tony
GPA	4.4	2.0
TOEIC	910	920

GPA+TOEIC	**914.4**	**922.0**
Avg(GPA,TOEIC)	**457.2**	**461.0**

Tony is better

- Range of TOEIC : 0~990
- Rage of GPA : 0 ~ 4.5

Data Scaling

- There are different methods of scaling data, including:
  - **Min-max scaling** : a simple method to scale data where each feature is scaled to a specific range, typically 0 to 1. The formula for min-max scaling is:

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

where:

- X_{scaled} is the scaled value.
- X is the original value.
- X_{min} and X_{max} are the minimum and maximum values of the feature, respectively.

Data Scaling

- There are different methods of scaling data, including:
  - **Z-score standardization** : rescales data so that it has a mean of 0 and a standard deviation of 1. The formula for standardization is:

$$Z = \frac{X - \mu}{\sigma}$$

where:

- Z is the standardized value.
- X is the original value.
- μ is the mean of the feature.
- σ is the standard deviation of the feature.

Data Scaling

◆ Let's apply both scaling methods to a simple dataset to illustrate their effects. Consider the following dataset of ages: [20, 22, 25, 27, 29].

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

$$Z = \frac{X - \mu}{\sigma}$$

where:

- Z is the standardized value.
- X is the original value.
- μ is the mean of the feature.
- σ is the standard deviation of the feature.

➢ Min-max scaling

- Min = 20
- Max = 29
- (20-20) / 9 , (22-20) / 9, (25-20) / 9, (27-20) / 9, (29-20) / 9
- 0, 2/9, 5/9, 7/9, 1

  = 0. 0.22222222 0.55555556 0.77777778 1.

➢ Z-score standardization

- μ= (20+22+25+27+29) / 5 = 24.6
- σ= 3.2619012...
- (20-24.6) / 3.26 , (22-24.6) / 3.26, (25-24.6) / 3.26, (27-24.6) / 3.26, (29-24.6) / 3.26
- -1.41022048 -0.79708114 0.12262787 0.73576721 1.34890655

Data Scaling

- (ex) Who is better
  - Min-max scaling
  - (ex) GPA: min = 0 , max = 4.5 / TOEIC: min = 0, max = 990

	Steve		Tony	
		Scaled		Scaled
GPA	4.4		2.0	
TOEIC	910		920	

GPA+TOEIC	914.4		922.0	
Avg(GPA,TOEIC)	457.2		461.0	

Steve is better

Data Scaling

- (ex) Who is better
  - Min-max scaling
  - (ex) GPA: min = 0 , max = 4.5 / TOEIC: min = 0, max = 990

	Steve		Tony	
		Scaled		Scaled
GPA	4.4	0.9778	2.0	0.4444
TOEIC	910	0.9192	920	0.9293

GPA+TOEIC	914.4	**1.897**	922.0	**1.3737**
Avg(GPA,TOEIC)	457.2	**0.9485**	461.0	**0.6869**

Steve is way better!


Bye Bye