DATA AND
ARTIFICIAL INTELLIGENCE

Statistics and Probability

Learning Objectives

By the end of this lesson, you will be able to:

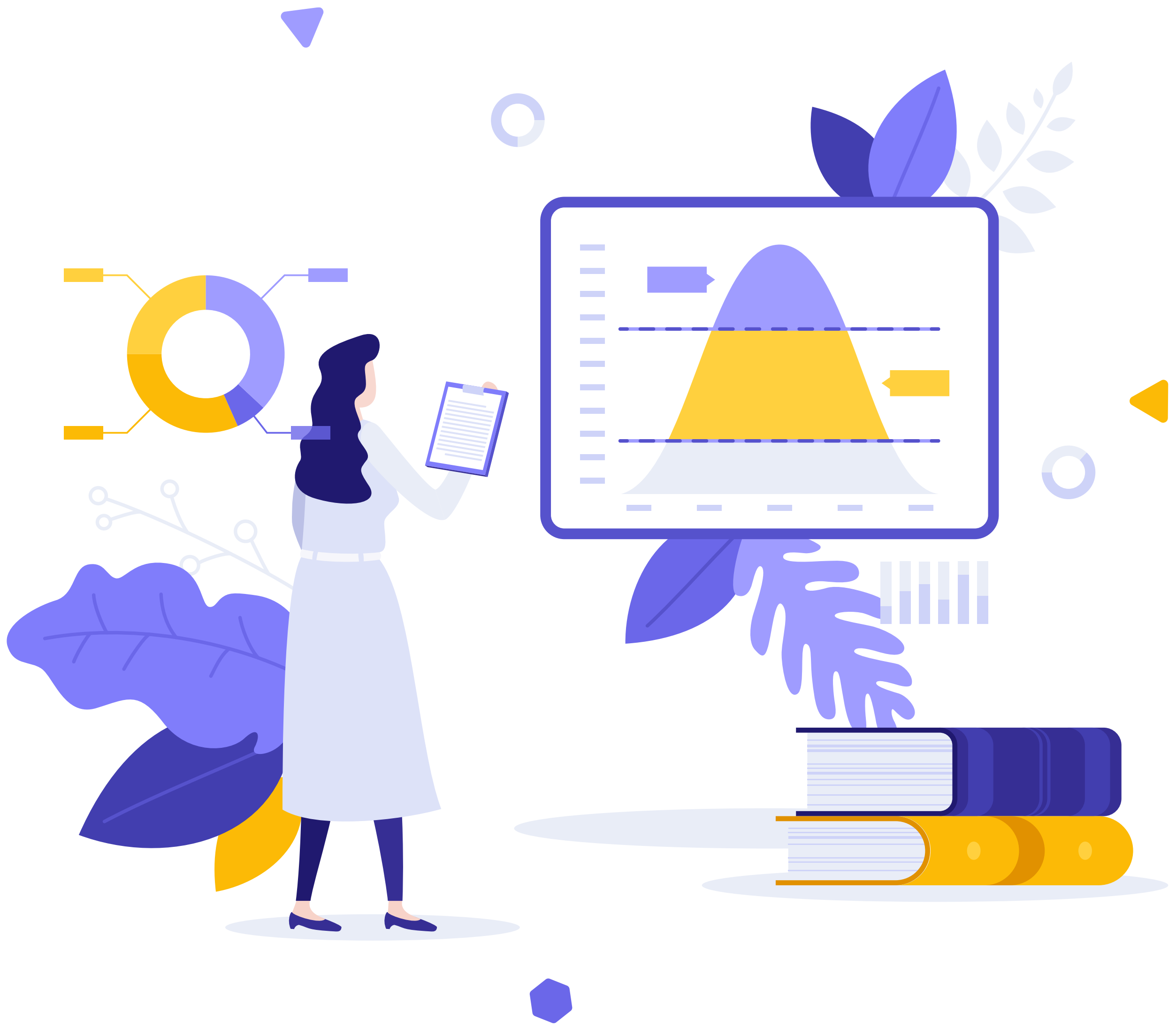
- Explain the concepts of statistics and probability
- Describe conditional probability
- Define the chain rule of probability
- Discuss the measure of variance
- Identify the types of gaussian distribution

DATA AND
ARTIFICIAL INTELLIGENCE

Basics of Statistics and Probability

simplilearn

Probability and Statistics



- Data science relies heavily on estimates and predictions.
- A significant portion of data science is made up of evaluations and forecasts.
- Statistical methods are used to make estimates for further analysis.

Probability and Statistics

- Probability theory is helpful for making predictions.
- Statistical methods are highly dependent on probability theory, and all probability and statistics are dependent on data.

Probability and Statistics

Data is information acquired for reference or research via observations, facts, and measurements.

Data is a set of facts structured in a form that computers can interpret, such as numbers, words, estimations, and views.

Importance of Data

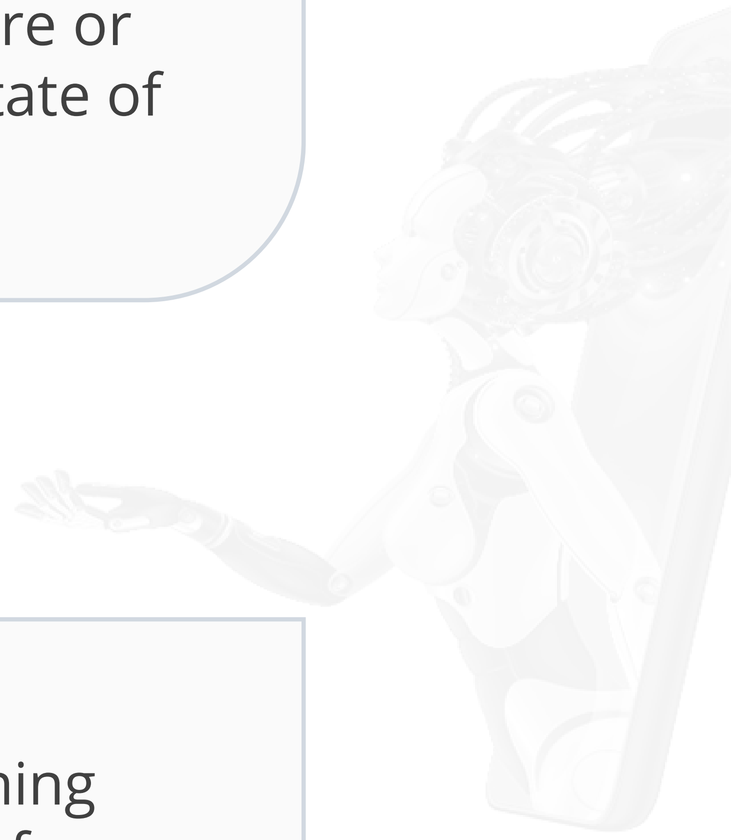
Data

Data aids in seeing more about the information by identifying possible connections between two features.

Based on prior information patterns, data may be utilized to anticipate the future or predict the current state of affairs.

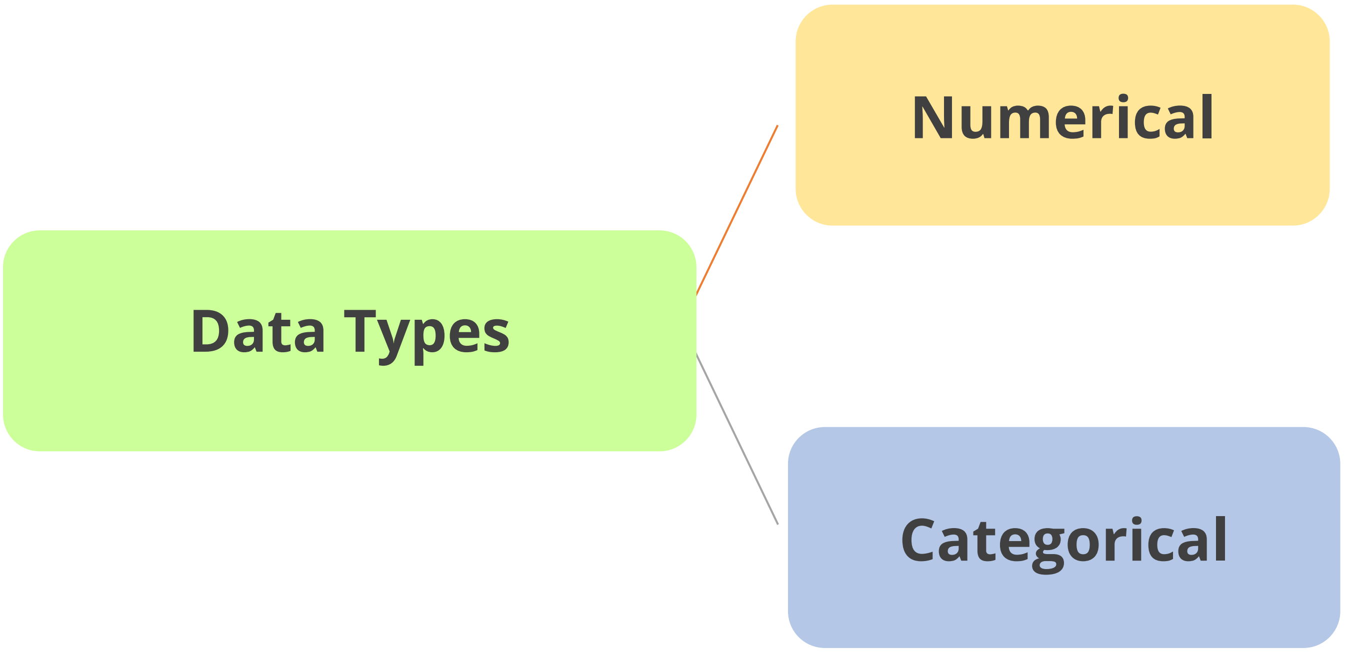
Data assists in the detection of distortion by uncovering hidden patterns.

Data aids in determining whether two pieces of information have any instances in common.



Types of Data

Data might be numerical, such as age, or categorical, such as gender.



Types of Data: Example

A person's bank data may be categorized into numerical and category data.

RowNumber	CustomerID	Surname	Geography	Gender	Age	Balance	HasCrCard	IsActiveMember
1	15634602	Hargrave	France	Female	42	0	1	1
2	15647311	Hill	Spain	Female	41	83807.86	0	1
3	15619304	Onio	France	Female	42	159660.8	1	0
4	15701354	Boni	France	Female	39	0	0	0
5	15737888	Mitchell	Spain	Female	43	125510.8	1	1
6	15574012	Chu	Spain	Male	44	113755.8	1	0
7	15592531	Bartlett	France	Male	50	0	1	1
8	15656148	Obinna	Germany	Female	29	115046.7	1	0
9	15792365	He	France	Male	44	142051.1	0	1
10	15592389	H?	France	Male	27	134603.9	1	1
11	15767821	Bearce	France	Male	31	102016.7	0	0
12	15737173	Andrews	Spain	Male	24	0	1	0
13	15632264	Kay	France	Female	34	0	1	0
14	15691483	Chin	France	Female	25	0	0	0
15	15600882	Scott	Spain	Female	35	0	1	1

Numerical

CustomerID, Age, Balance

Categorical

Geography, Gender, HasCrCard, IsActiveMember

Descriptive Statistics

A descriptive measurement is a summary measure that quantitatively portrays the most important features of a set of data, allowing for a better comprehension of the information.

Measurement levels of data

- **Qualitative**
  - Nominal
  - Ordinal
- **Quantitative**
  - Ordinal
  - Interval
  - Ratio

DATA AND
ARTIFICIAL INTELLIGENCE

Introduction to Descriptive Statistics

Descriptive Statistics

Nominal

- Data is categorized using names, labels, or qualities.
- Example: Brand name, zip code, and gender

Ordinal

- Data can be arranged in order or ranked and can be compared.
- Example: Grades, star reviews, position in a race, and date

Descriptive Statistics

Interval

- Data can be ordered as it is in a range of values and meaningful differences between the data points can be calculated.
- Example: Temperature in Celsius, and year of birth

Ratio

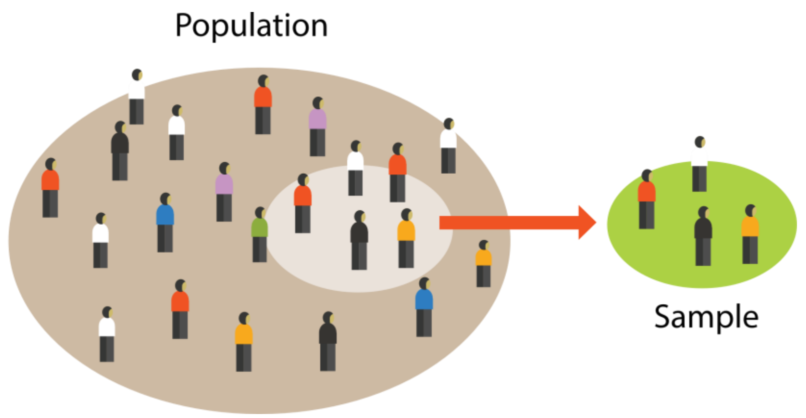
- Data at this level is similar to the interval level with the added property of inherent zero.
- Mathematical calculations can be performed on these data points.
- Example: Height, age, and weight

Population vs. Sample

Before analyzing the data, it's important to figure out if it's from a population or a sample.

Population

- Population is a collection of all available items (N), as well as each unit in our study.
- Population data is used when the data pool is very small and can give all the required information.



Sample

- Sample is a subset of the population (n) that contains only a few units of the population.
- Samples are collected randomly and represent the entire population.

Measures of Central Tendencies

Measure of Central Tendencies

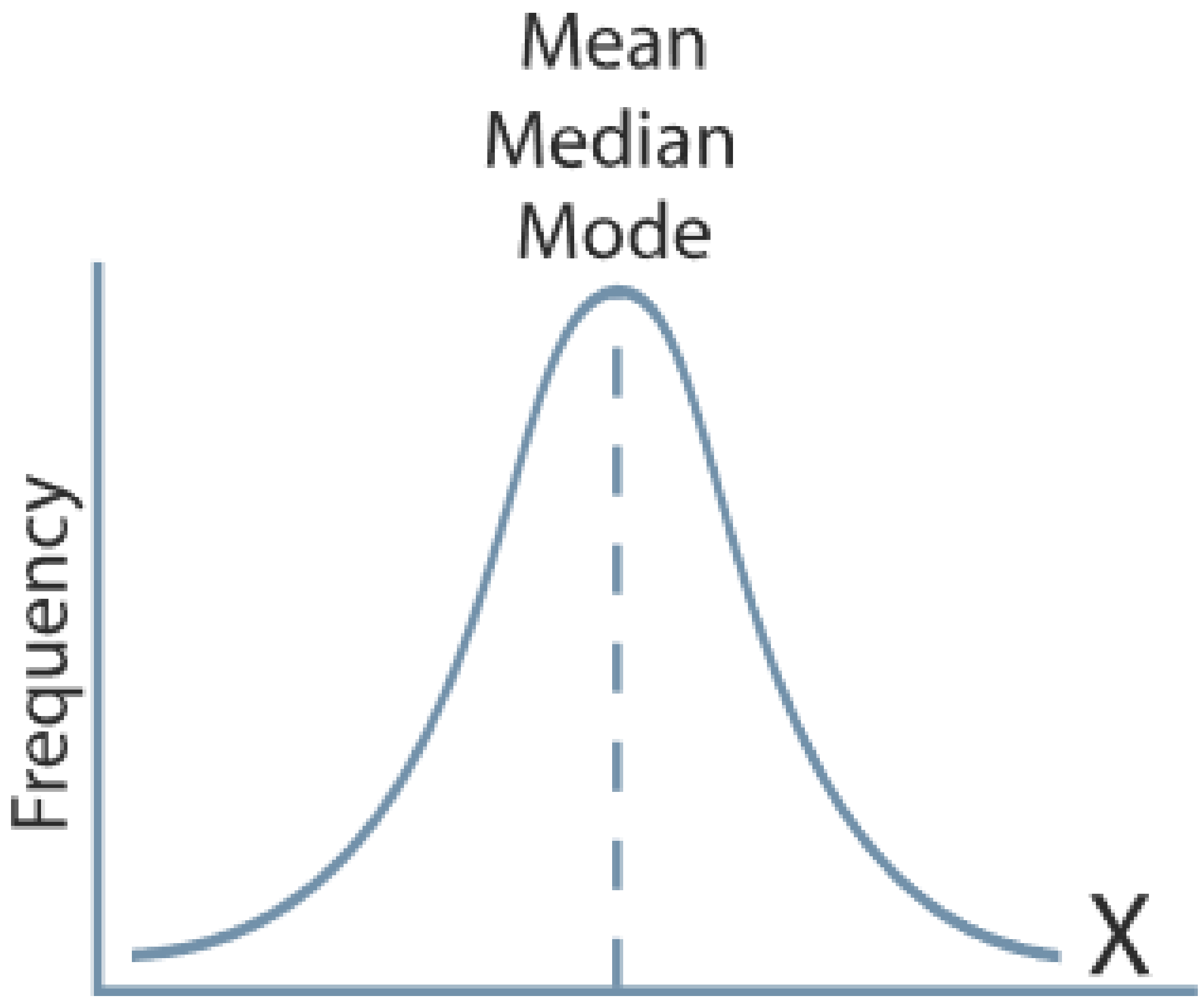
The central tendency is a single value that aids in the description of data by determining its center position.

Measures of central tendency are sometimes known as summary statistics or measures of central location.

The most popular measurements of central tendency are mean, median, and mode.

Measure of Central Tendencies

The normal distribution is a bell-shaped symmetrical distribution in which the mean, median, and mode are all equal.



Measure of Central Tendencies: Mean

Mean

It is calculated by dividing the sum of all data values by the total number of data values.

It gets affected when there are unusual or extreme values.

It is sensitive to outliers.

Measure of Central Tendencies: Mean

The formula for calculating mean is given below:

$$\text{Arithmetic Mean Formula} = \frac{X_1 + X_2 + X_2 + \ldots\ldots\ldots\ldots\ldots\ldots + X_n}{n}$$

Example

7, 3, 4, 1, 6, 7

Mean = 7+3+4+1+6+7

= 28 / 6

= 4.66

Measure of Central Tendencies: Median

Median

- It is the middle value in a set of data that has been sorted ascending or descending.
- It is a better alternative to mean since it is less impacted by outliers and skewness.
- It is closer to the central value.

Measure of Central Tendencies: Median

The formulas for calculating median are given below:

If the total number of values is odd, then:

$$median = (\frac{n+1}{2})^{th} term$$

Example

7, 3, 4, 1, 6
After sorting:
1, 3, 4, 6, 7
Median = 4

If the total number of values is even, then:

$$median = (\frac{(\frac{n}{2})^{th} term + (\frac{n}{2}+1)^{th} term}{2})^{th} term$$

Example

7, 3, 4, 1, 7, 6
After sorting:
1, 3, 4, 6, 7, 7
Median = 4+6 / 2 = 5

Measure of Central Tendencies: Mode

Mode

The mode represents the most common value in the dataset.

It is not at all affected by extreme observations.

It is the best measure of central tendency for highly skewed or non-normal distributions.

Measure of Central Tendencies: Mode

The formula for calculating mode is given below:

Mode Formula

$$= L + \frac{(f_m - f_1)Xh}{(f_m - f_1) + (f_m - f_2)}$$

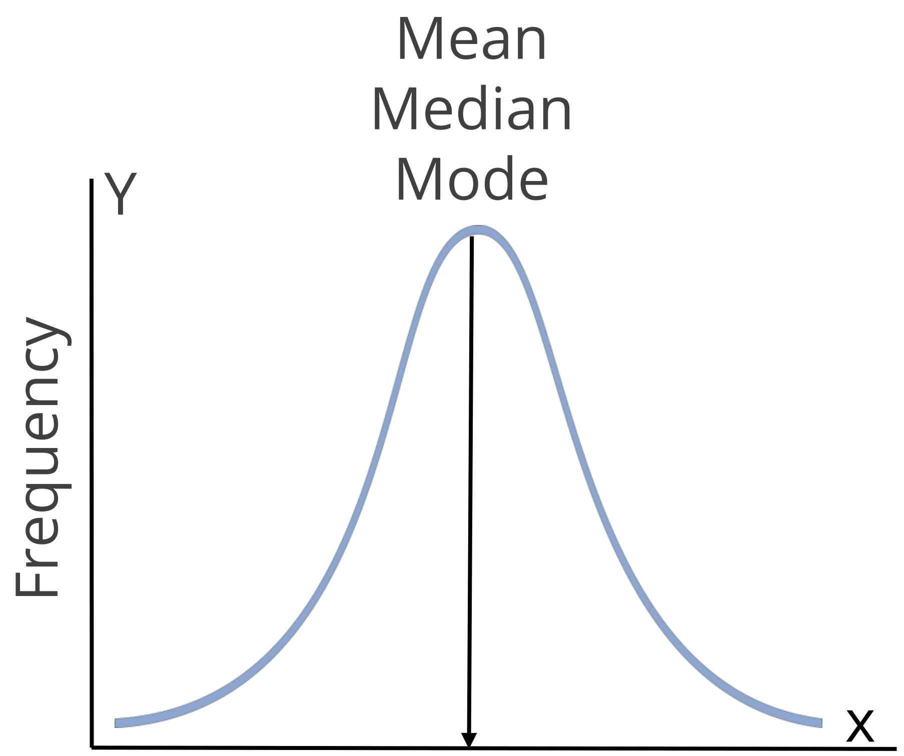
Example

7, 3, 4, 1, 6, 7
Mode = 7

Measures of Asymmetry

Measures of Asymmetry: Skewness

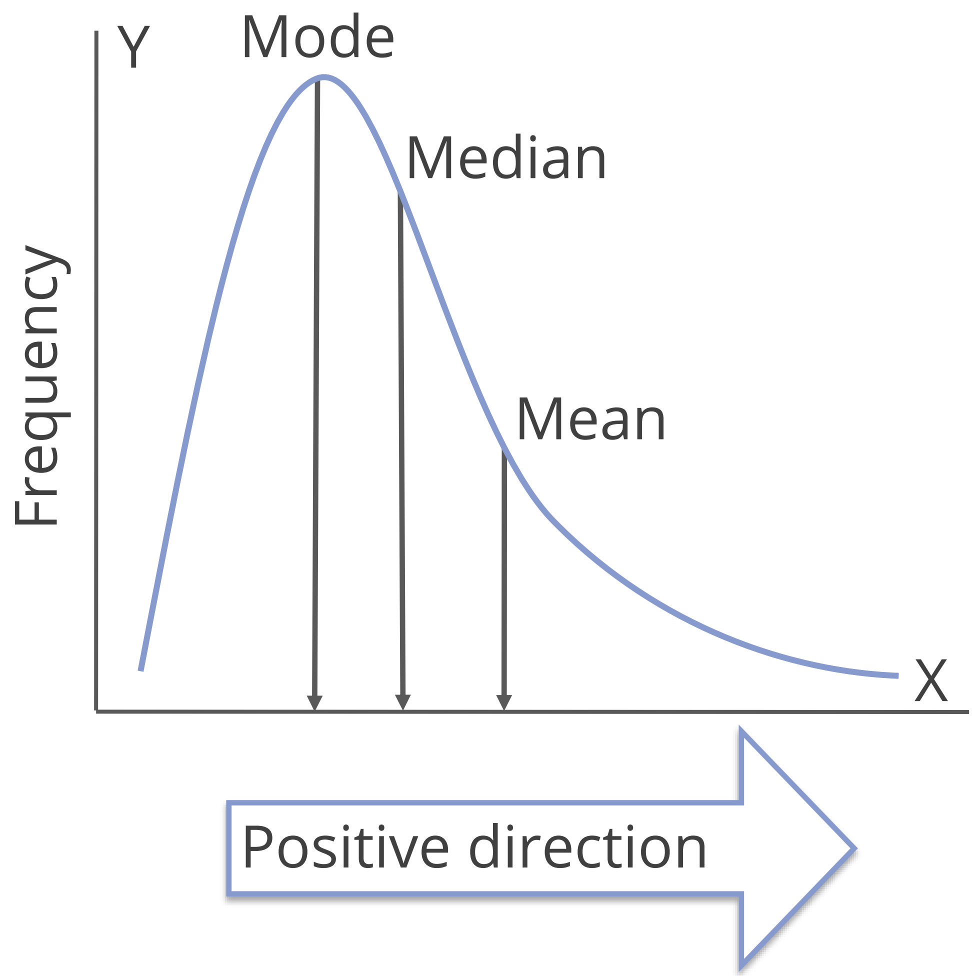
Skewness is a type of asymmetry in statistical data distribution that happens when curves are distorted or skewed to either the right or left side.



A normal distribution curve is a symmetrical bell curve.

Skewness is critical for data interpretation since it reveals how data is distributed.

Measures of Asymmetry: Positive Skewness

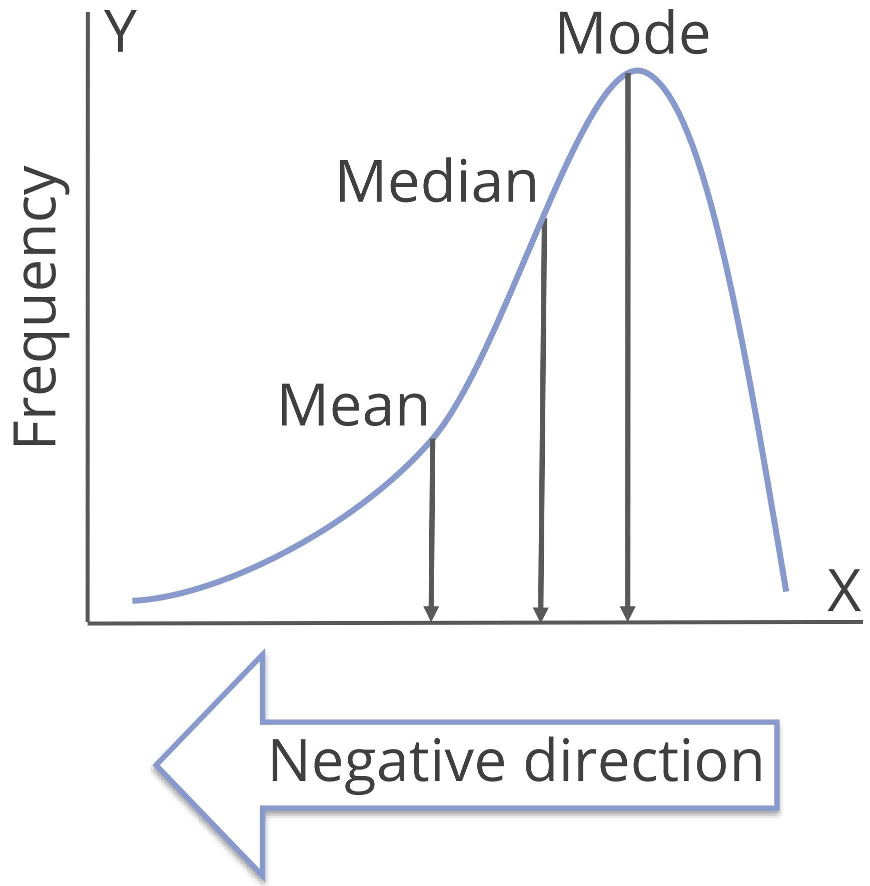


The tail is skewed to the right, while the outliers are also skewed to the right.

In the positive skewness:
mean > median > mode

The tail is skewed to the right, indicating that the outliers are skewed to the right.

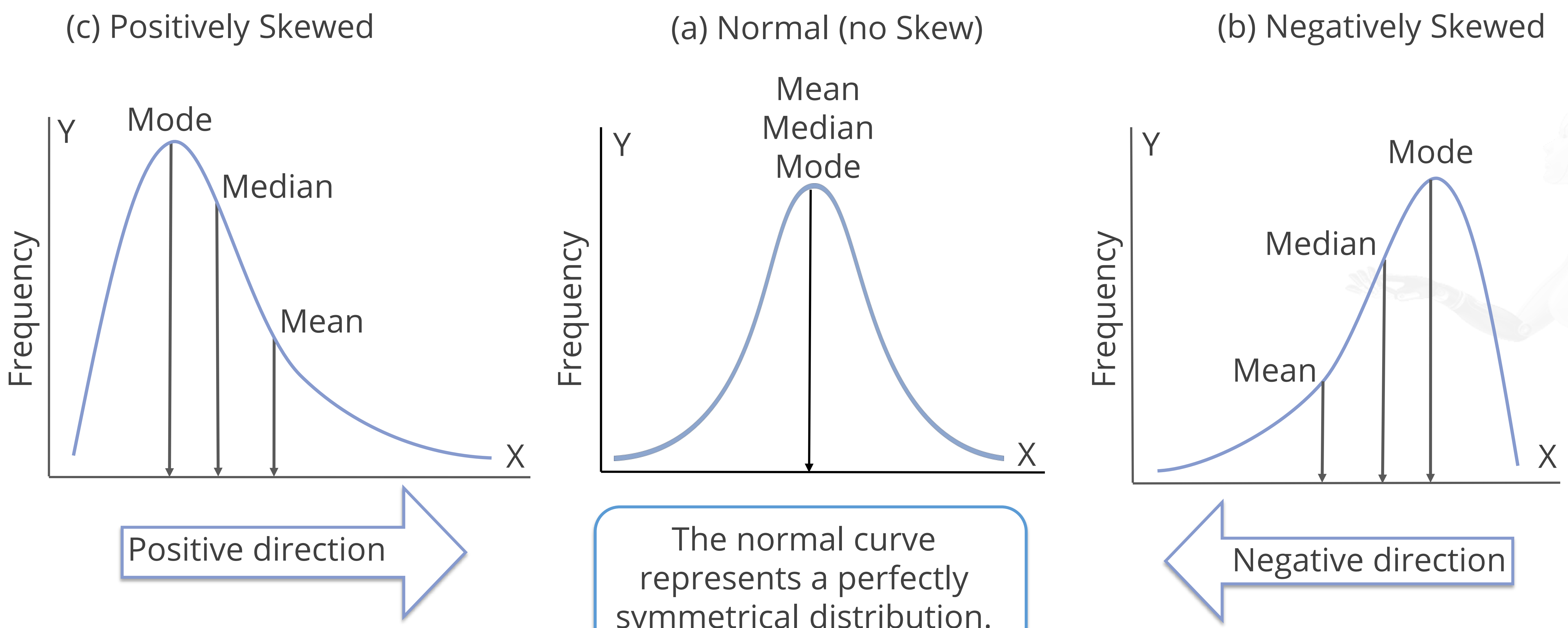
Measures of Asymmetry: Negative Skewness



- The right side of the curve contains the majority of the values.
- In negative skewness:
  mean > median > mode
- The tail is skewed to the left, indicating that the outliers are skewed to the left.

Measures of Asymmetry

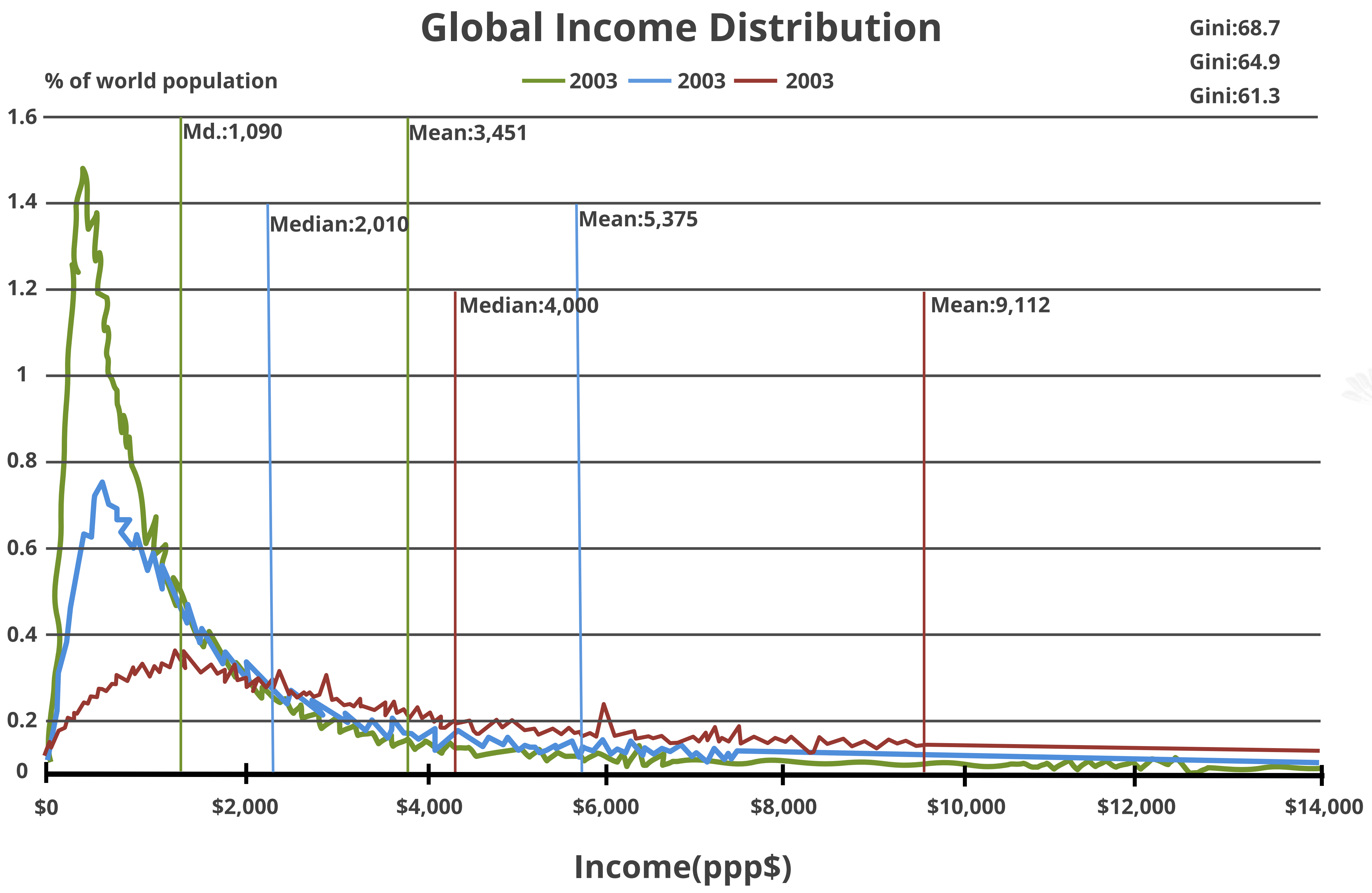
The difference between three distinct curves is shown in the image below:



The normal curve represents a perfectly symmetrical distribution.

Measures of Asymmetry: Example

The global income distribution statistics from 2003 is highly right-skewed.



Measures of Asymmetry: Example

The following are the key observations from the previous graph:

In 2003, the mean of $3,451 was higher than the median of $1,090.

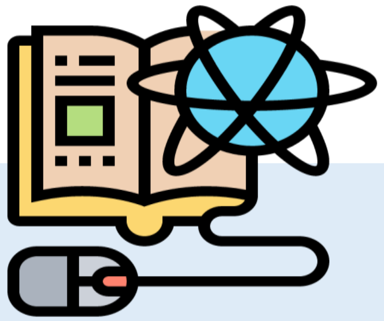
The global income is not evenly distributed.

The majority of the people make less than $2,000 each year.

Only a small percentage of the population earns more than $14,000.

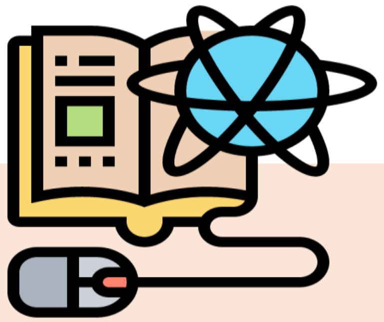
Measures of Variability

Measures of Variability: Dispersion



- The measure of central tendency provides a single value that addresses the full worth; however, the central tendency cannot depict the viewpoint entirely.
- The metric of dispersion helps us focus on the inconsistency in data spread.
- Measures of dispersion describe the spread of the data.
- The range, interquartile range, standard deviation, and variance are all examples of dispersion measures.

Measures of Variability: Range

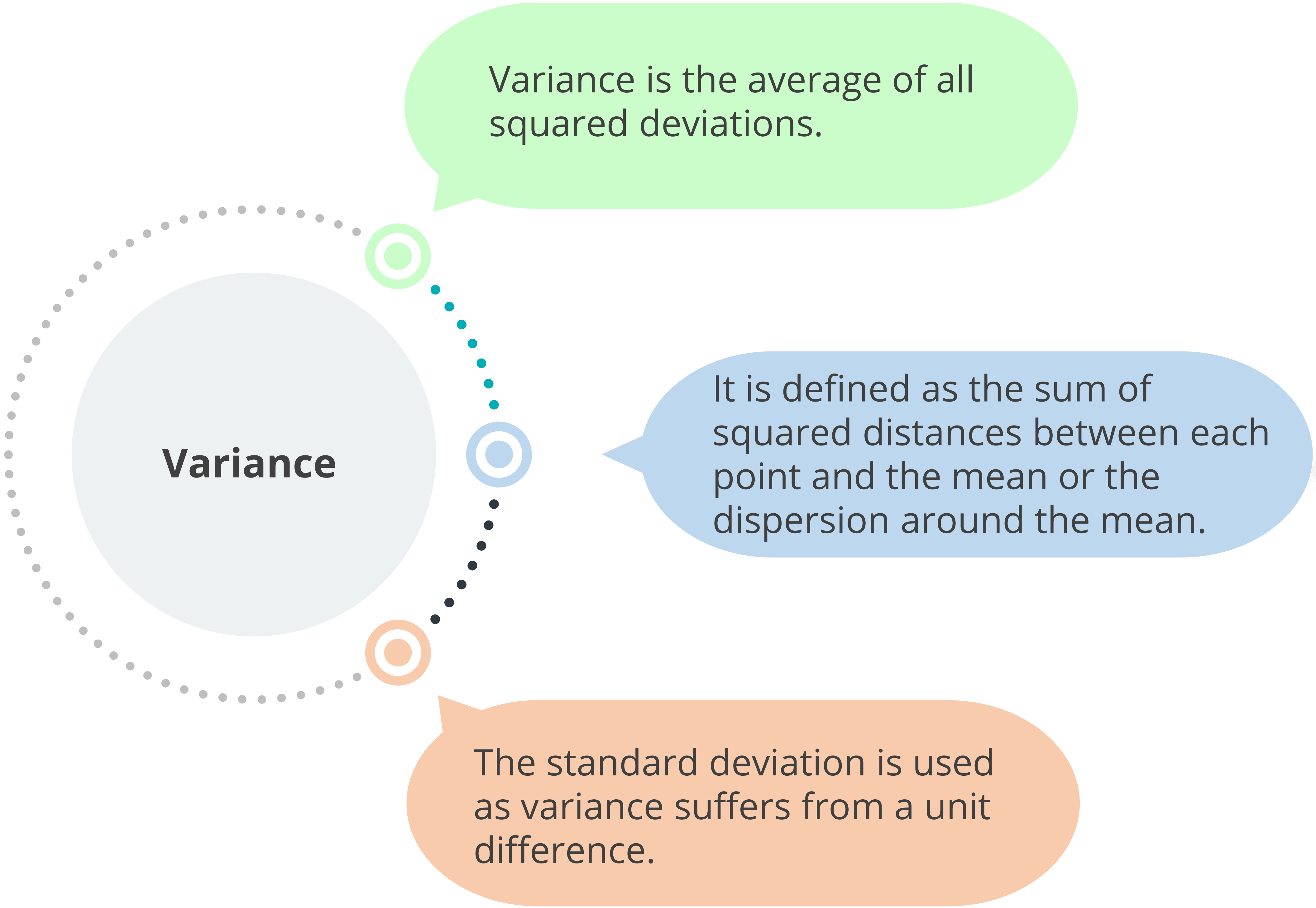


- The range of the distribution is the difference between the largest and the smallest amount of data.
- The range, for example, does not include all of a series' positive aspects.
- It concentrates on the most shocking aspects and ignores those that aren't considered critical.

Example

For {13,33,45,67,70} the range is 57 i.e., (70-13)

Measures of Variability: Variance



Measures of Variability: Variance

The formulas for calculating variance are given below:

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

$$s^2 = \frac{\sum (X - \overline{X})^2}{n - 1}$$

The units of values and variance are not equal, so another variability measure is used.

Measures of Variability: Standard Deviation

Standard Deviation

Standard deviation is a statistical term used to measure the amount of variability or dispersion around an average.

The standard deviation is calculated as the square root of the variance.

It depicts the concentration of the data around the mean of the dataset.

Measures of Variability: Standard Deviation

The formula for calculating variance is given below:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

Measures of Variability: Example

Finding the mean, variance, and standard deviation for the following dataset:

Example

Consider a dataset with the following values:
{3,5,6,9,10}

$$Mean = \frac{3+5+6+9+10}{5} = 6.6$$

$$Variance = \frac{(3-6.6)^2 + (5-6.6)^2 + (6-6.6)^2 + (9-6.6)^2 + (10-6.6)^2}{5}$$

$$= \frac{12.96 + 2.56 + 0.36 + 5.76 + 11.56}{5}$$

$$= \frac{33.2}{5} = 6.64$$

$$Standard\,Deviation = \sqrt{variance} = \sqrt{6.64} = 2.576$$

Measures of Relationship

Measures of Relationship: Covariance

The two variables are compared using relationship measures.

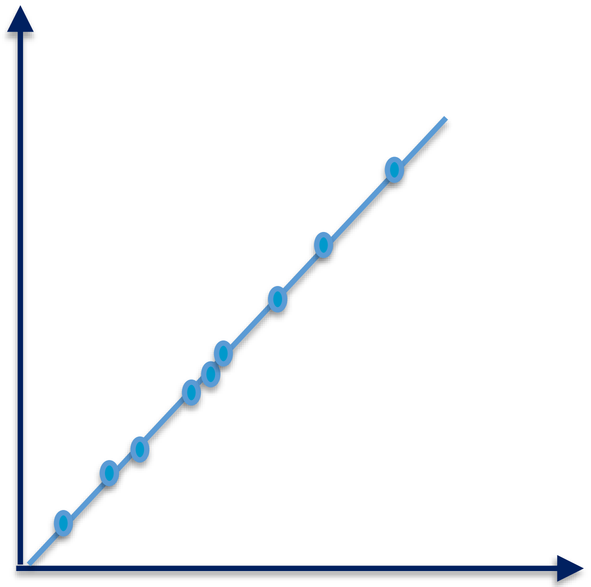
The covariance of two variables is a measure of their relationship.

It calculates the degree of change in the variables.

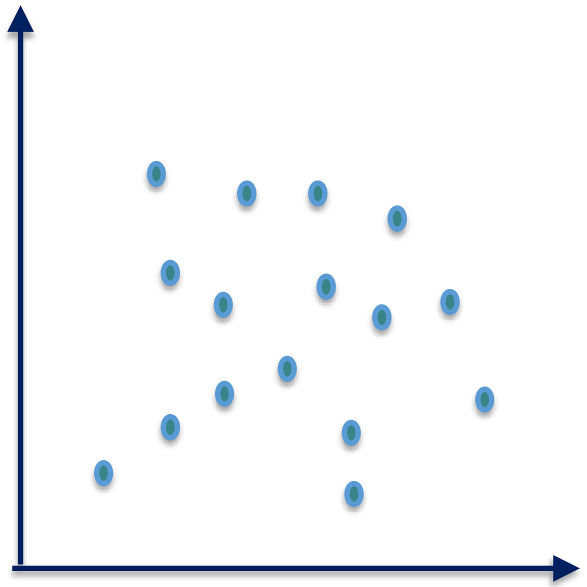
It determines if one variable will cause the other to alter in the same way.

Types of Correlation

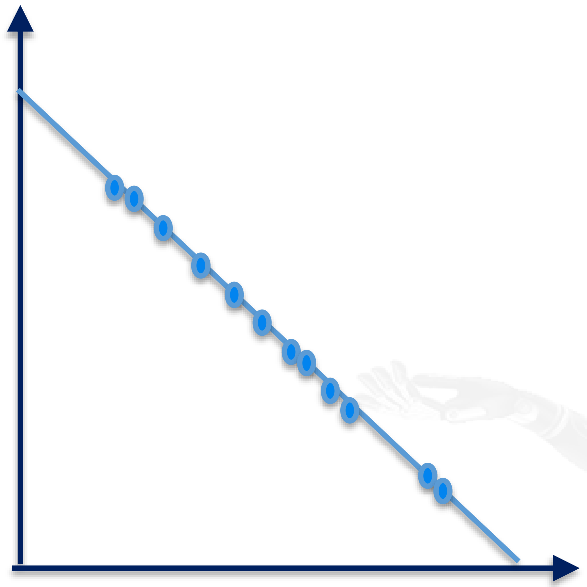
There are three types of correlation which are:



Perfect Positive
Correlation



Zero
Correlation



Perfect negative
Correlation

Measures of Relationship: Correlation

- Correlation provides a better understanding of covariance.
- It is normalized covariance.
- It calculates the degree of change in the variables
- It is also known as the Pearson Correlation Coefficient.

Measures of Relationship: Correlation

The formula for calculating correlation is as follows:

$$Correlation = \rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y})$$

The value of correlation ranges from -1 to 1.

Measures of Relationship: Correlation

Correlation = 1

- A correlation of 1 implies a positive relationship.
- When one independent variable increases, the other dependent variable increases.

Correlation = -1

- A correlation of -1 implies a negative relationship.
- When one independent variable increases, the other dependent variable decreases.

Correlation = 0

The value 0 shows that the variables are independent of one another.

Measures of Relationship: Example

Consider the following example for calculating the correlation:

Example

Height	Weight	$x - \overline{x}$	$y - \overline{y}$	$(x - \overline{x})(y - \overline{y})$	$(x - \overline{x})^2$	$(y - \overline{y})^2$
5	4.5	-0.14	-5	0.7	0.019	25
5.5	53	-0.36	3	-1.08	0.129	9
6	70	0.86	20	17.2	0.739	400
4.7	42	-0.44	-8	3.52	0.193	64
4.5	40	-0.64	-10	6.4	0.409	100

Measures of Relationship: Example

The formula for calculating correlation is given below:

$$Sum(Height) = 25.7 \quad Mean(Height) = 5.14$$
$$Sum(Weight) = 250 \quad Mean(Weight) = 50$$
$$\sum (x-\overline{x})(y-\overline{y}) = 26.74$$
$$\sum (x-\overline{x})^2 = 1.489$$
$$\sum (y-\overline{y})^2 = 598$$

$$correlation = \frac{\sum (x-\overline{x})(y-\overline{y})}{\sqrt{\sum (x-\overline{x})^2}\sqrt{\sum (y-\overline{y})^2}}$$

$$= \frac{26.74}{\sqrt{1.489}\sqrt{598}}$$

$$= \frac{26.74}{1.220 * 24.454}$$

$$0.889$$

Correlation of 0.889 indicates that height and weight have a positive relationship.
It is evident that as a person's height grows, weight also increases.

Introduction to Probability

Probability Theory

Probability is a measure of the likelihood that an event will occur.

Example

- One considers a coin toss scenario where the chances of getting heads on a coin toss are ½ or 50%.
- The probability of each given event is between 0 and 1 (inclusive).
- The sum of an event's cumulative probability cannot be greater than 1.
- Hence, the probability of x lies between zero and one, that is, $0 <= p(x) <= 1$.
- This means that $\int p(x)dx = 1$ (integral of p for a distribution over x).

Conditional Probability

The conditional probability of occurrence A given that event B has previously occurred is as follows:

This is also called **Bayesian probability**.

$$P(A|B) = \frac{P(A, B)}{P(B)}$$

https://en.wikipedia.org/wiki/Thomas_Bayes

Conditional Probability

The Bayes model specifies the probability of event A occurring if event B has already occurred.

If, P(A) = Probability of event A
And P(B) = Probability of event B

Then, P(A ∏ B) = Probability of both events happening

$$P(A \cap B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(A|B)P(B)$$

Conditional Probability: Example

Bayes Model Example

Consider the coin example where:
P(Coin1-H) = 2/4
P(Coin2-H) = 2/4
P(Coin1-H ∏ Coin2-H) = ¼

Two Coin Flip	
Coin 1	Coin 2
H	T
T	H
H	H
T	T

The probability of Coin1-H, given Coin2-H, can be calculated as follows:

P(Coin1-H | Coin2-H) = (1/4)/(2/4)
= ½
= 50%

Conditional Probability: Bayes Equation

Simplifying the Bayes Equation

Events A and B are statistically independent if:

$$P(A \cap B) = P(A|B)P(B)$$

$$P(A \cap B) = P(A)P(B)$$

$P(A|B) = P(A)$ assumes P(B) is not zero

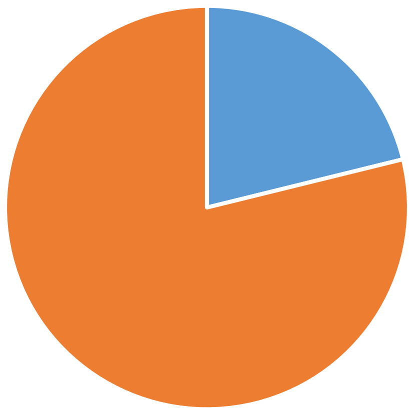
$P(B|A) = P(B)$ assumes P(A) is not zero

AI with Bayes Model: Example

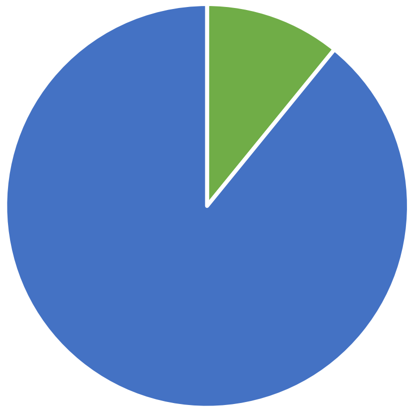
Consider an example for calculating the likelihood of having diabetes based on the frequency of fast-food consumption

Observed Data

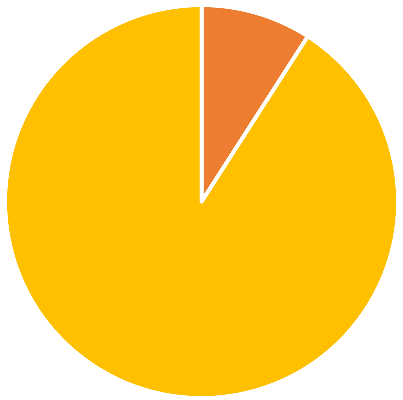
Fast food audience 20%



Diabetes prevalence 10%



Fast food and diabetes 5%



Chances of diabetes, given fast food: (conditional probability)

$\Rightarrow$ (D and F)/F = 5%/20% = ¼ = 25%

Analysis: Eating fast food increases the chance of having diabetes by 25%.

Chain Rule of Probability

Joint probability distributions over many random variables can be reduced into conditional distributions over a single variable.

Equation

It can be expressed as follows:

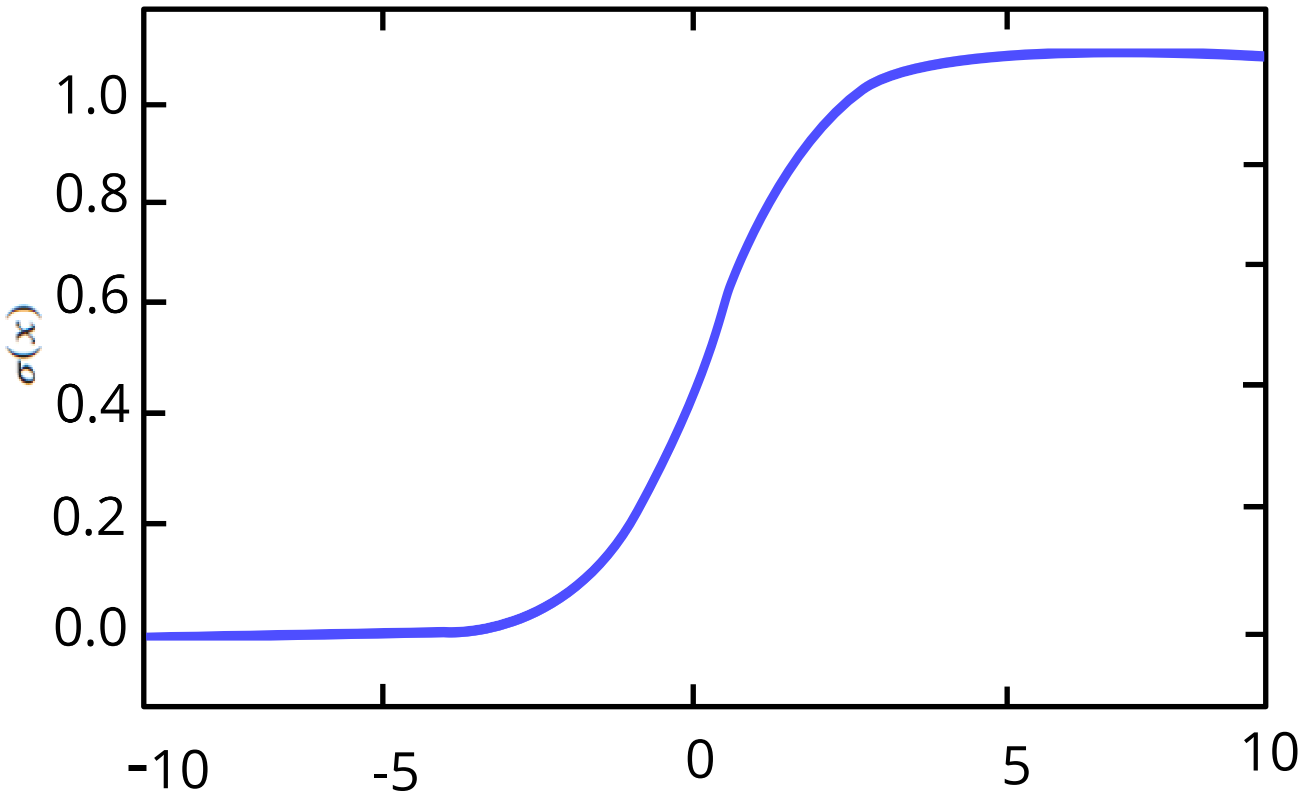
$$P(X^{(1)}, \ldots\ldots, X^{(n)}) = P(X^{(1)}) \Pi_{i=2}^{n} P(X^{(i)} | X^{(1)}, \ldots\ldots, X^{(i-1)})$$

Example

P(a, b, c) = P(a | b, c) * P(b | c) * P(c)

Logistic Sigmoid

The logistic function is a type of sigmoid function that aims to predict the class to which a particular sample belongs.



- Its outcome is a discrete binary value, a probability between 0 and 1.
- The Logistic Sigmoid is a useful function that follows the S curve.
- It saturates when the input is very large or very small.

Logistic Sigmoid

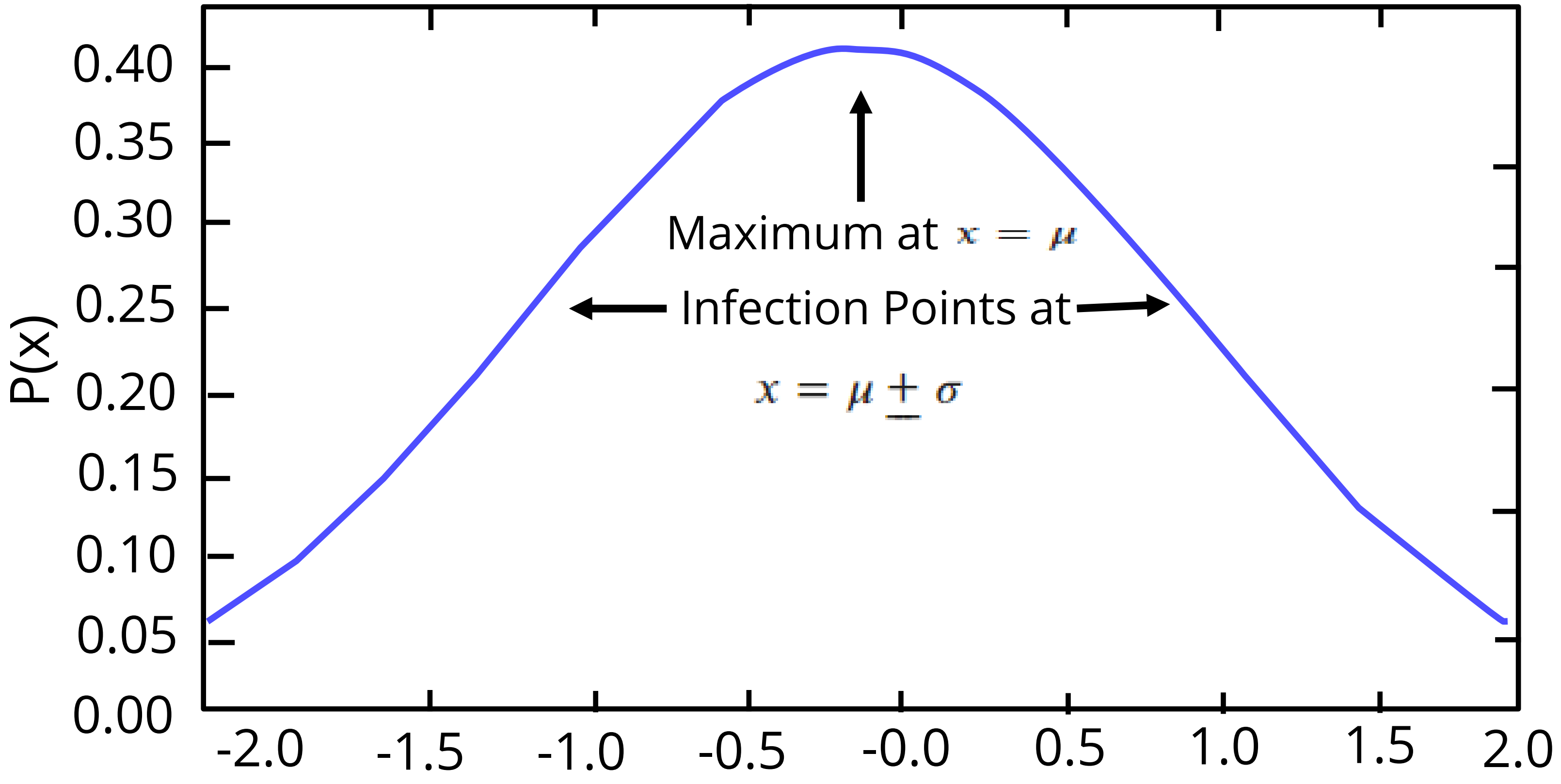
The formula for logistic sigmoid is given below:

Equation

$$\sigma(x) = \frac{1}{1 + exp(-x)}$$

Gaussian Distribution

The Gaussian distribution is a type of distribution in which data tends to cluster around a central value with little or no bias to the left or right.



It is often referred to as the normal distribution.

In the absence of prior information, the normal distribution is frequently a fair assumption in machine learning.

Gaussian Distribution: Equation

The formulas for calculating gaussian distribution are given below:

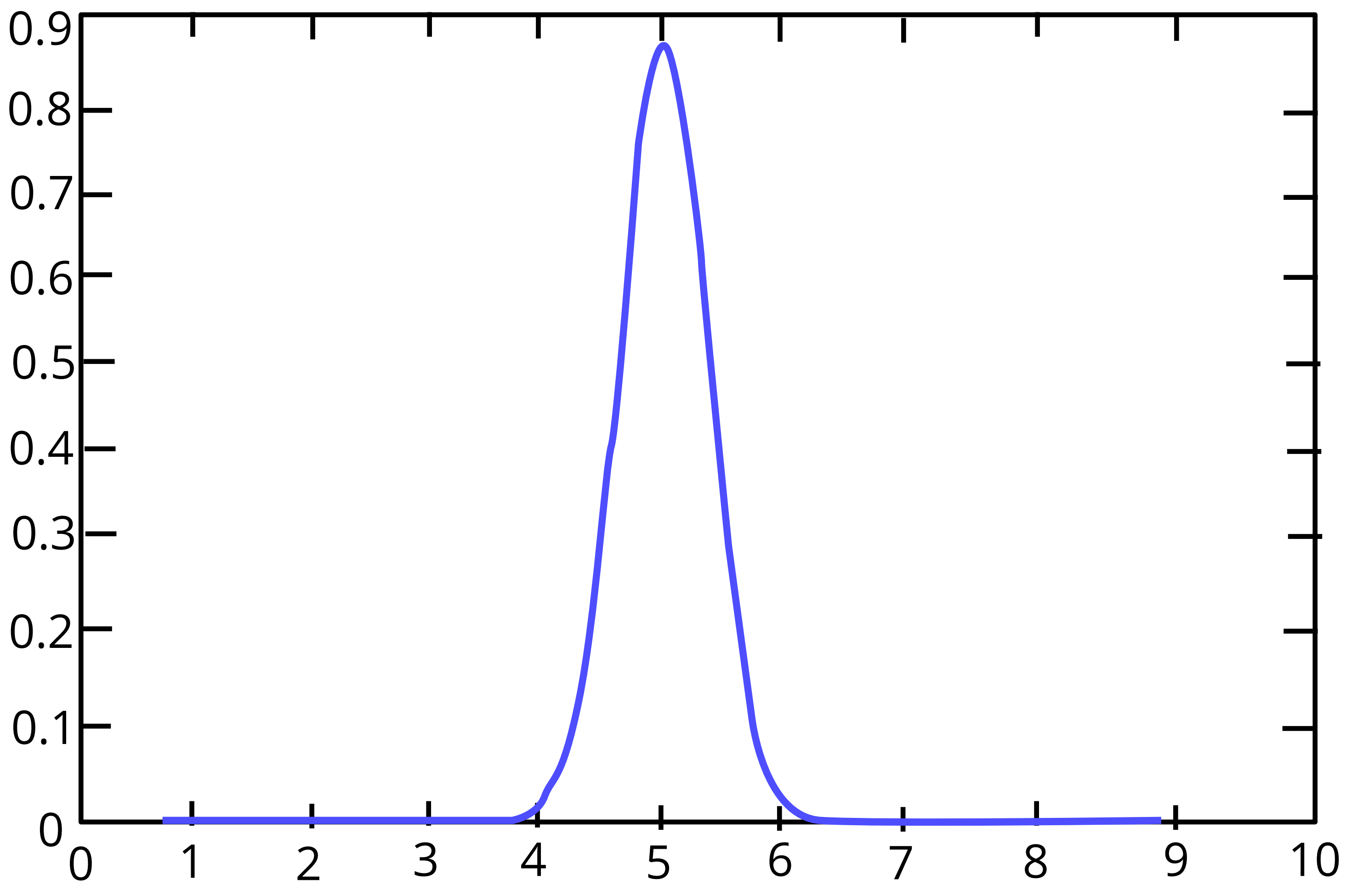
$$N(x; \mu, \sigma^2) = \sqrt{\frac{1}{2\Pi\sigma^2}} exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$$

$$N(x; \mu, \beta^{-1}) = \sqrt{\frac{\beta}{2\Pi}} exp(-\frac{1}{2}\beta(x-\mu)^2)$$

- μ = mean or peak value, which also means E[x] = μ
- σ = standard deviation, and σ^2 = variance
- A "standard normal distribution" has μ = 0 and σ = 1
- For efficient handling, invert σ and use precision β (inverse variance) instead

Types of Gaussian Distribution: Univariate

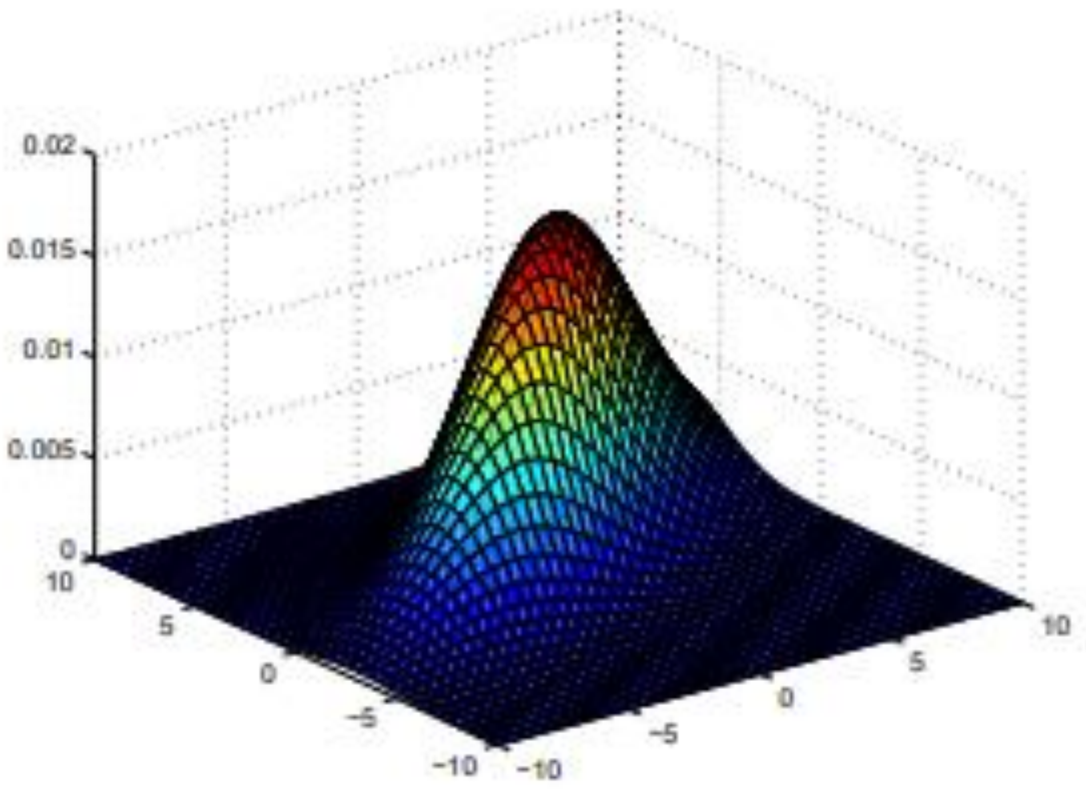
The distribution over a single variable is known as a univariate Gaussian distribution.



A univariate Gaussian distribution over a single variable x

Types of Gaussian Distribution: Multivariate

The multivariate normal distribution is an extension of the univariate normal distribution to several variables.



Multivariate Gaussian distribution over two variables x_1 and x_2

https://cs229.stanford.edu/section/gaussians.pdf

Key Takeaways

- Probability and statistics structure the premise of data.
- The data helps in anticipating the future or gauging in view of the past pattern of information.
- The central tendency is a single value that helps to describe data by identifying the central position. The mean, median, and mode are the measures of center tendencies.
- The distribution where the data tends to be around a central value with a lack of bias or minimal bias towards the left or right is called a Gaussian distribution.

Knowledge Check

Knowledge Check 1

If x1, x2, x3,….., xn are the observations of a given data. Then the mean of the observations will be:

A. Sum of observations/Total number of observations

B. Total number of observations/Sum of observations

C. Sum of observations + Total number of observations

D. Sum of observations + Total number of observations/2

Knowledge Check 1

If x1, x2, x3,….., xn are the observations of a given data. Then the mean of the observations will be:

A. Sum of observations/Total number of observations

B. Total number of observations/Sum of observations

C. Sum of observations + Total number of observations

D. Sum of observations + Total number of observations/2

The correct answer is **A**

Mean = Sum of observations/Total number of observations
In our example, mean = x1 + x2 + x3 +….+xn/n

Knowledge Check 2	Which of the following can be the probability of an event?

A. - 0.4

B. 1.004

C. 18/23

D. 10/7

Knowledge Check 2

Which of the following can be the probability of an event?

A. – 0.4

B. 1.004

C. 18/23

D. 10/7

The correct answer is **C**

The probability of an event is always between 0 and 1. In the above options, only 18/23 is between 0 and 1.