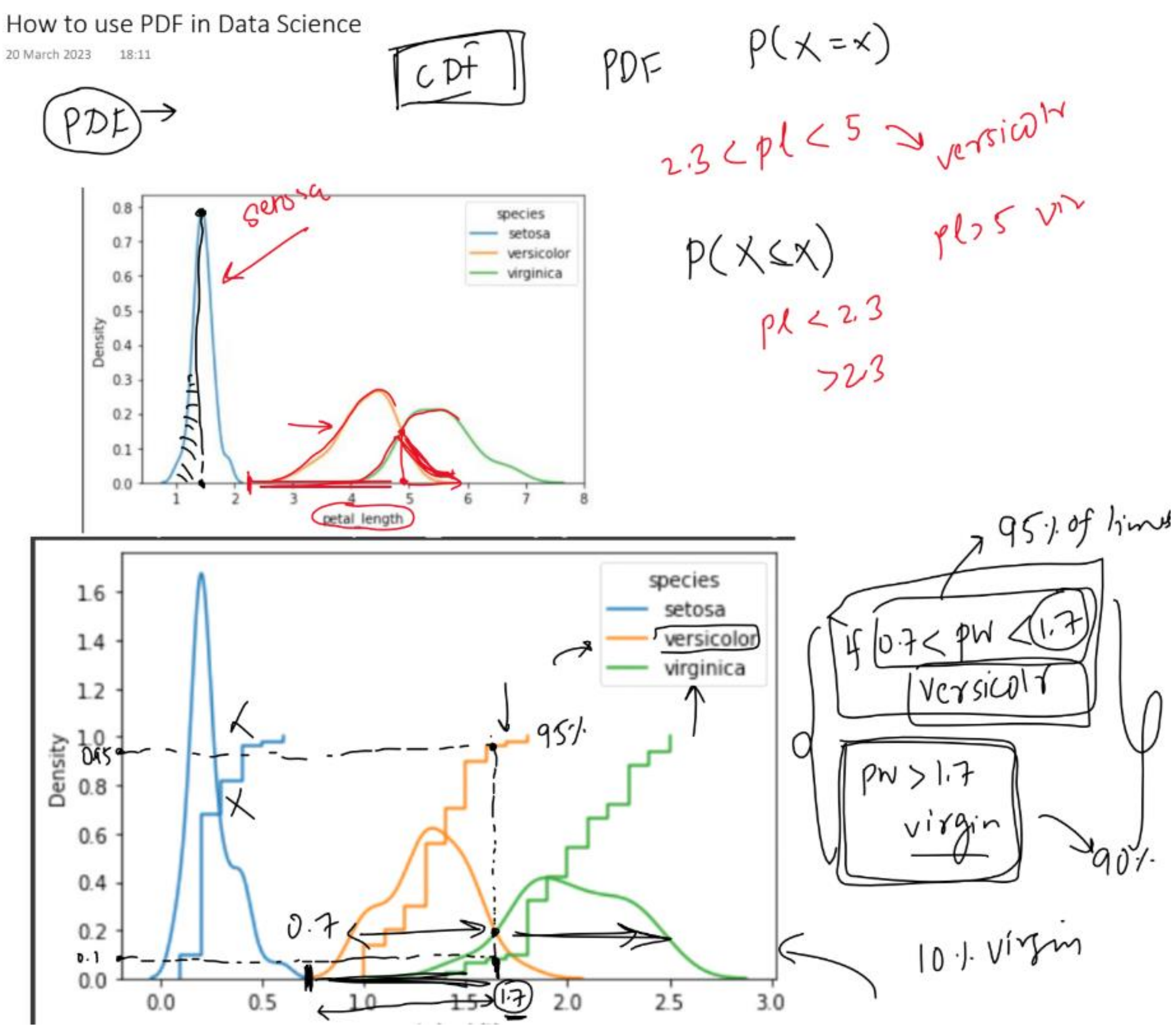
S 41 Normal Distribution

How to use PDF in Data Science :



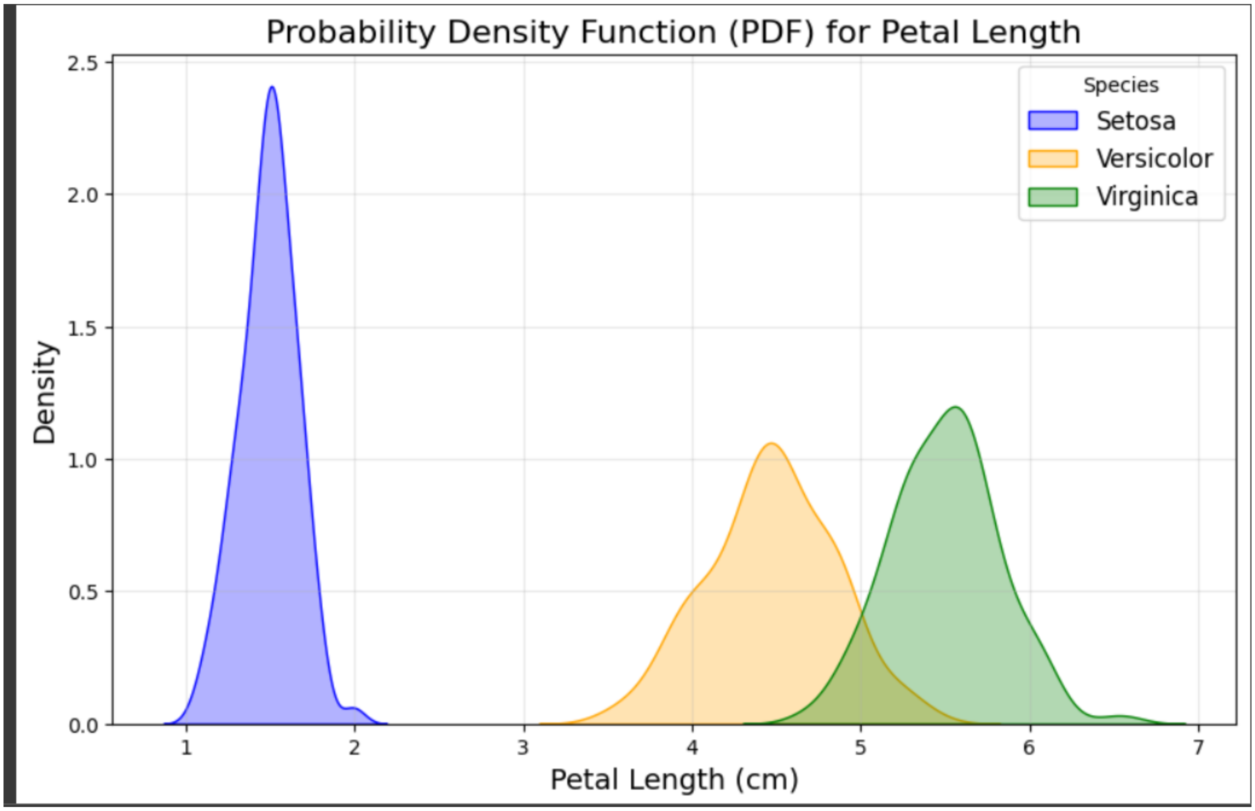
How to use PDF in Data Science

20 March 2023 18:11

PDF (Probability Density Function)

- PDF ek graph hai jo kisi variable ke probability distribution ko dikhata hai.
- Y-axis: Density (kitna frequent value hai).

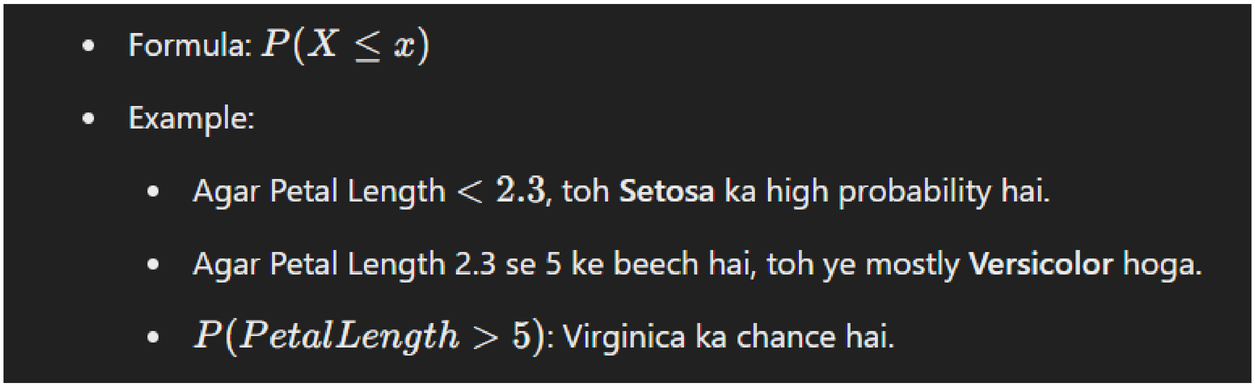
- X-axis: Variable values (e.g., **Petal Length**, **Petal Width**).



- Example:
  - Graph se pata chal raha hai ki **Setosa** species ka **petal length** 2 ke aas-paas zyada hota hai (peak near 2).
  - **Versicolor** aur **Virginica** ke petals ka length alag ranges mein aata hai.

CDF (Cumulative Distribution Function)

- CDF cumulative probability dikhata hai, yani kisi value ke neeche ke area ki probability kya hai.



- Formula: $P(X \leq x)$
- Example:
  - Agar Petal Length < 2.3, toh **Setosa** ka high probability hai.
  - Agar Petal Length 2.3 se 5 ke beech hai, toh ye mostly **Versicolor** hoga.
  - $P(PetalLength > 5)$: Virginica ka chance hai.

Image Insights

1. **Upper Graph (Petal Length)**
   - Blue line: Setosa ka petal length 1.5-2.5 ke beech zyada frequent hai.
   - Orange and Green lines overlap karte hain, jo dikhata hai ki **Versicolor** aur **Virginica** ka kuch values mein similarity hai.
   - **Use:** Is graph se hum ek range estimate kar sakte hain, e.g., agar 2.3 ≤ PetalLength ≤ 5, toh species **Versicolor** hoga.
2. **Lower Graph (Petal Width)**
   - Petal width ko analyze kiya gaya:
     - Agar 0.7<PetalWidth<1.7, hai toh **Versicolor** hone ka 95% chance hai.
     - Agar PetalWidth >1.7, toh **Virginica** hone ka 90% chance hai.
   - Blue peak shows Setosa ka petal width chota hota hai.

Visualization Importance

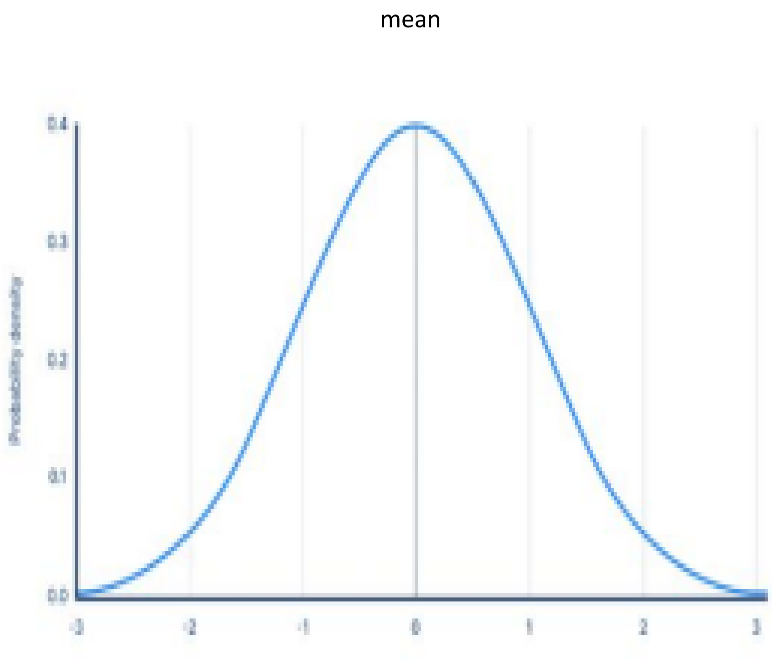
- **PDF aur CDF** ko use karke hum kisi dataset ke features ko samajh sakte hain.
- **Decision Boundaries:** Graphs dikhate hain ki ek feature (Petal Length/Width) ki kaunsi range ek species ko identify karne mein madad karti hai.

Real-Life Example

Imagine karo ek **flower classification app** banani hai, toh:

1. **Petal Width** aur **Petal Length** ke range ko analyze karke species predict karoge.
2. Agar width > 1.7, app suggest karega "Ye Virginica ho sakta hai.

Normal Distribution



What is Normal Distribution?

Normal distribution, also called **Gaussian distribution**, is a type of data distribution. It looks like a **bell-shaped curve**. It is **symmetrical**, which means the left and right sides of the graph are the same.

Key Features of Normal Distribution

1. Bell Shape

- The graph looks like a bell.
- **Middle part (mean)**: Most of the data points are near the middle.
- **Edges (tails)**: Very few data points are far away from the middle.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

---

2. Mean (μ):

- The **mean** is the **center** of the graph.
- It shows the **average value** of the data.
- **Example**: If most students in a class are 5.5 feet tall, then 5.5 is the mean height.

---

3. Standard Deviation (σ):

- Standard deviation tells us how **spread out the data is**.

    - **Small σ**: The data points are very close to the mean.
    - **Large σ**: The data points are far from the mean.
- **Example**:
    - If students' heights are between 5.3 and 5.7 feet, σ is small.
    - If students' heights range from 5 to 6 feet, σ is large.

---

4. Tails

- The graph's ends are called **tails**.
- They show the **extreme values** (very small or very large values).
- Example: Very short or very tall people are shown at the tails.

---

5. Asymptotic Nature

- The graph's tails never touch the x-axis. They come very close but go on forever.

Real-Life Examples

1. **Students' Marks**:
    - Most students score around the average marks (e.g., 50-60 out of 100).
    - Very few students score very high (90-100) or very low (0-10).
2. **Human Height**:
    - Most people are close to the average height (e.g., 5-6 feet).

- Very few people are very short or very tall.

3. **Daily Temperature**:
    - Most days, the temperature is near the average (e.g., 25-30°C).
    - Extreme temperatures (very cold or very hot) are rare.

How to Explain to Someone Else

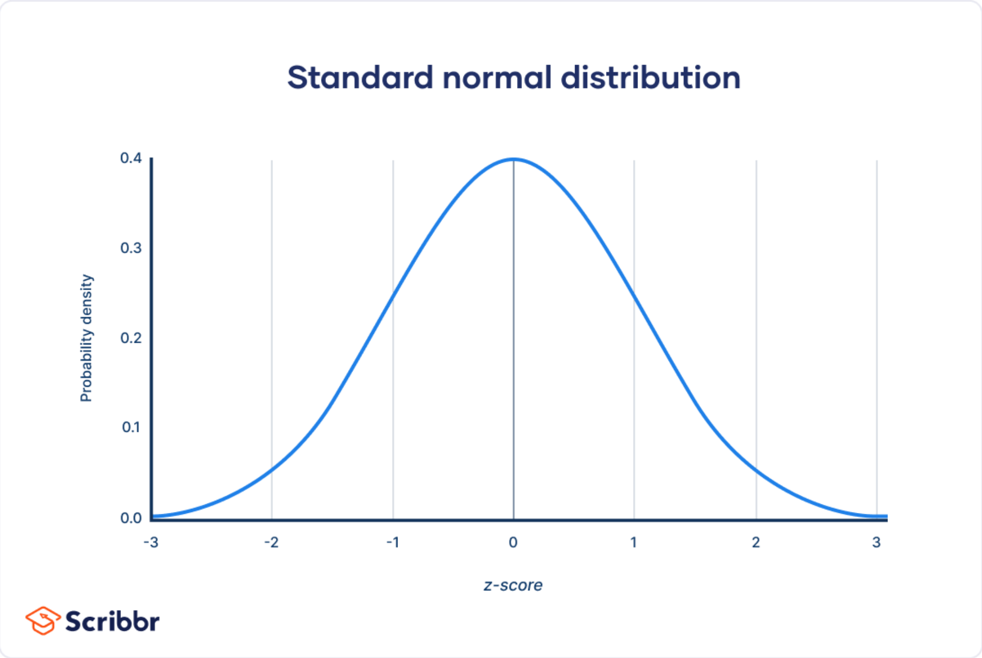
1. "Normal distribution is like a bell-shaped graph."
2. "Most of the data is near the middle, and very few are at the edges."
3. "The mean is the center, and standard deviation shows how spread out the data is."
4. "Examples are marks, height, and temperature—they follow a similar pattern."

What is Standard Normal Variate (Z)?

The **Standard Normal Variate (Z)** is a special version of the normal distribution where:

- The **mean (μ)** = 0
- The **standard deviation (σ)** = 1

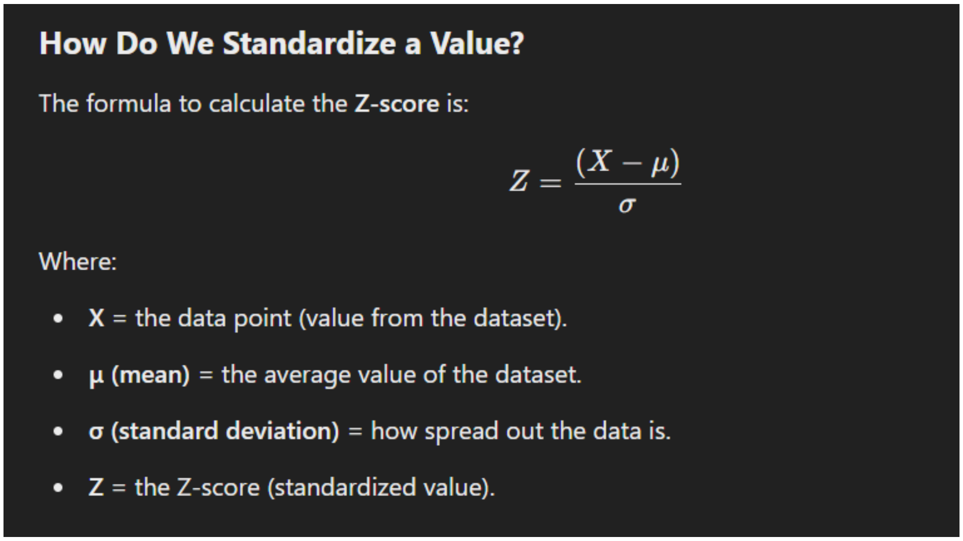
This is also called the **standard normal distribution**.



Why Standard Normal Variate is Useful?

When we **standardize** a normal distribution, we can:

1. **Compare different distributions** easily, even if their means and standard deviations are different.
2. Use **Z-tables** (or software) to quickly find probabilities and percentiles.



How Do We Standardize a Value?

The formula to calculate the **Z-score** is:

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

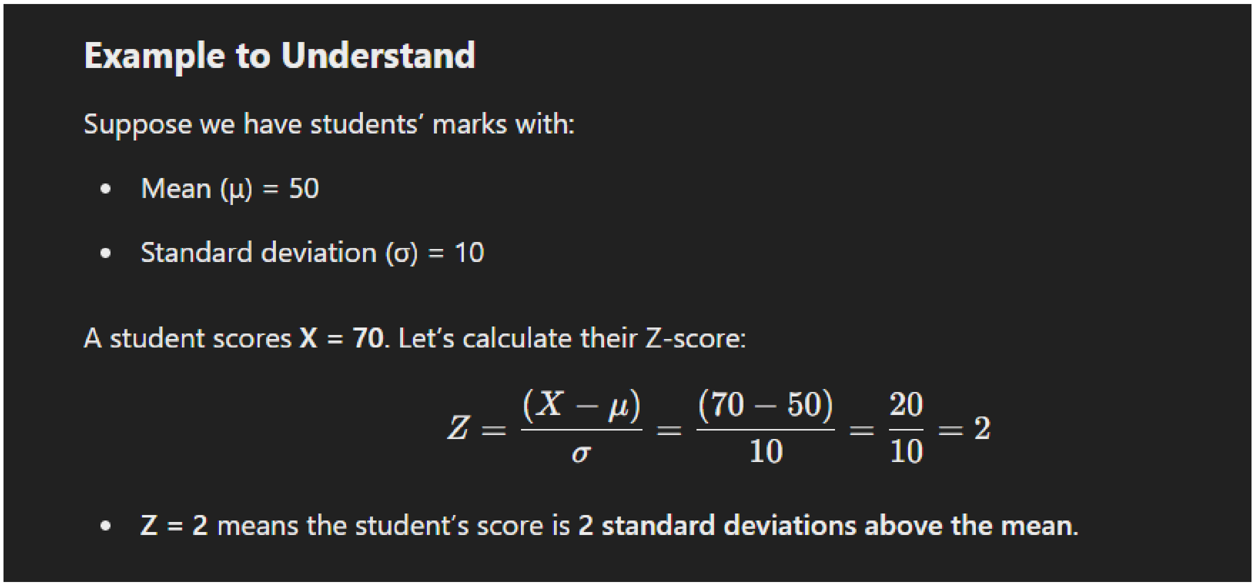
Where:

- **X** = the data point (value from the dataset).
- **μ (mean)** = the average value of the dataset.
- **σ (standard deviation)** = how spread out the data is.
- **Z** = the Z-score (standardized value).

What Does Z-Score Tell Us?

The Z-score tells us how **far a value (X)** is from the mean in terms of **standard deviations**.

- **Z = 0 :** The value is exactly at the mean.
- **Z > 0 :** The value is above the mean.
- **Z < 0 :** The value is below the mean.



Example to Understand

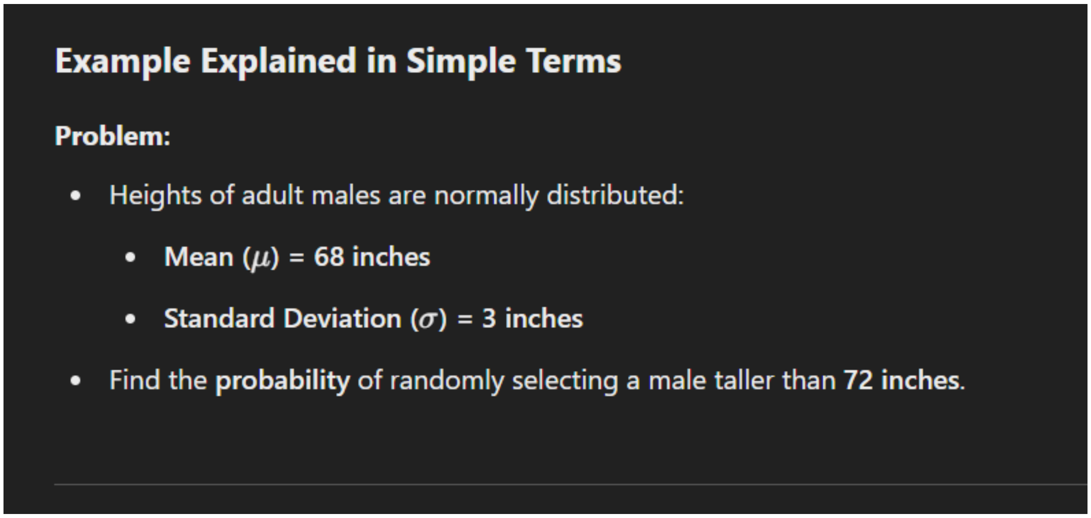
Suppose we have students' marks with:

- Mean (μ) = 50
- Standard deviation (σ) = 10

A student scores **X = 70**. Let's calculate their Z-score:

$$Z = \frac{(X - \mu)}{\sigma} = \frac{(70 - 50)}{10} = \frac{20}{10} = 2$$

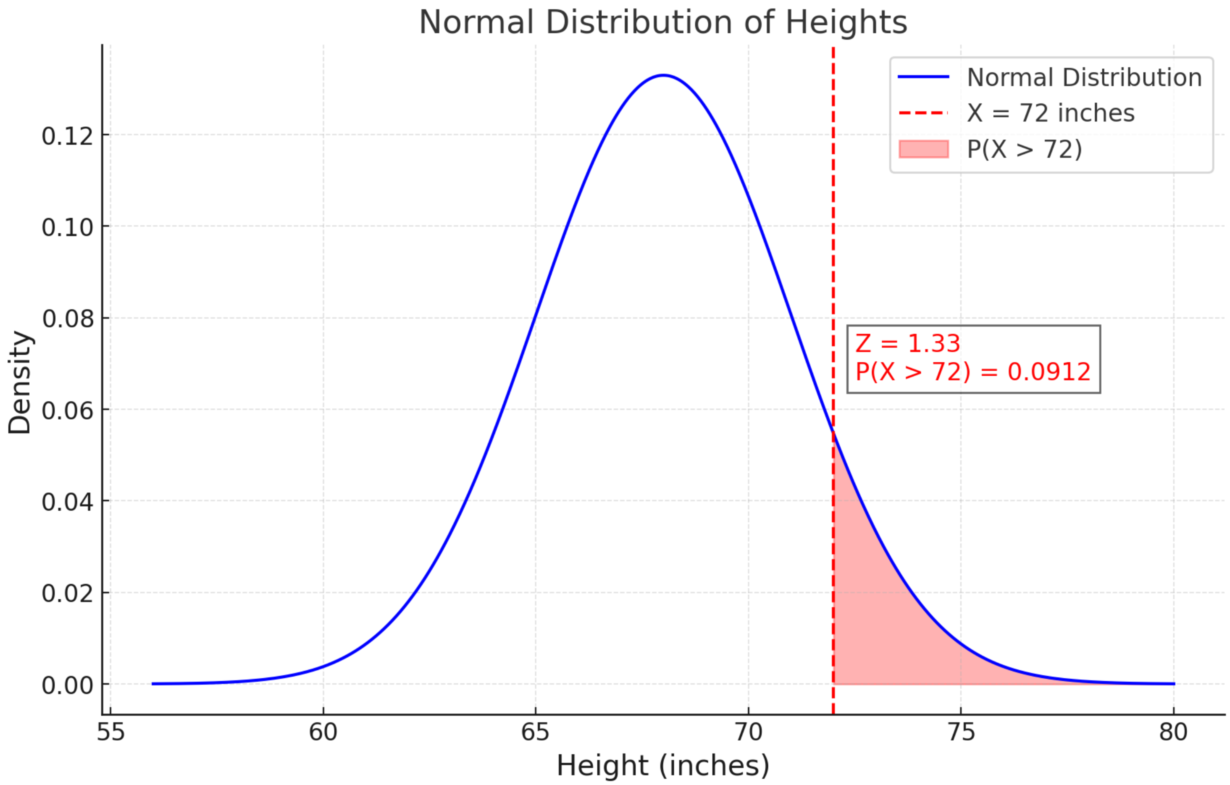
- Z = 2 means the student's score is **2 standard deviations above the mean**.

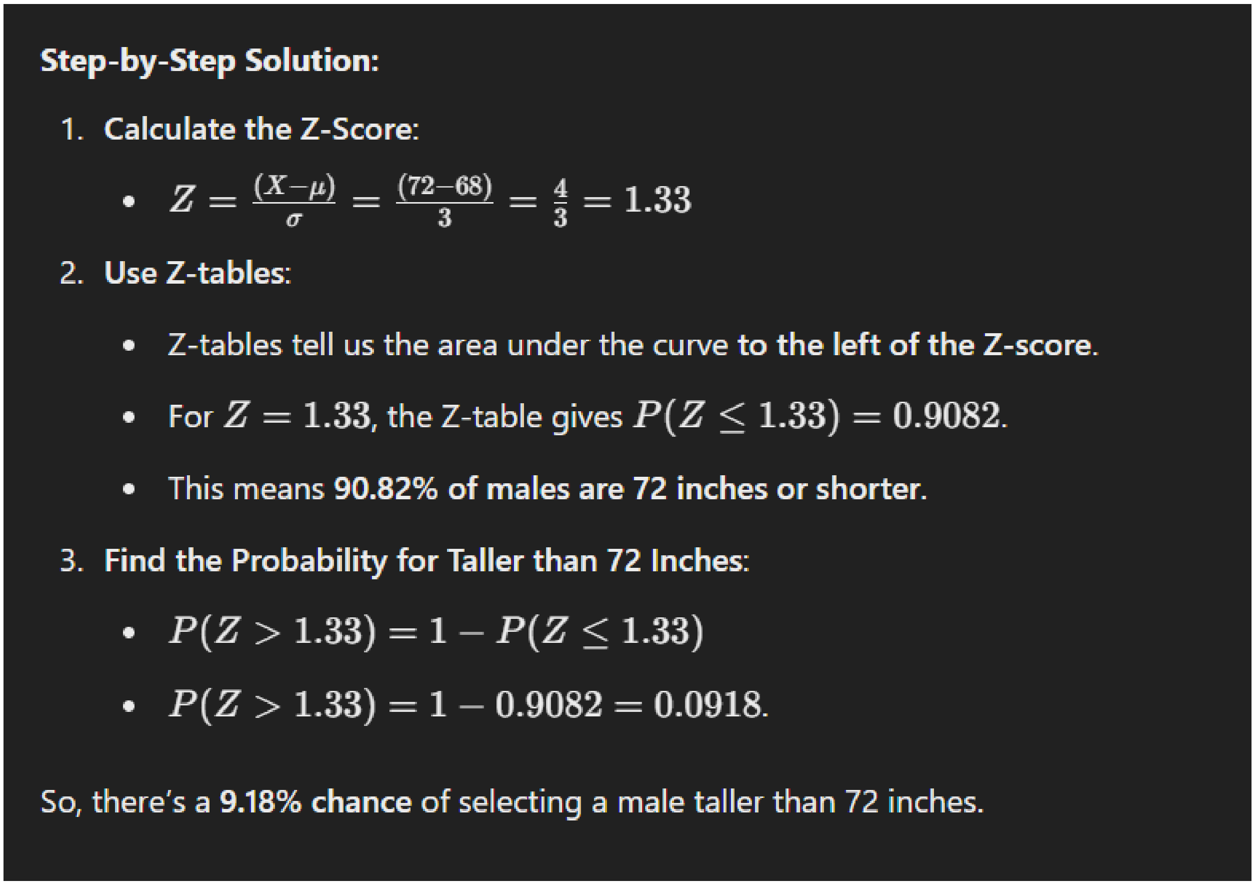


Example Explained in Simple Terms

Problem:

- Heights of adult males are normally distributed:
  - **Mean (μ) = 68 inches**
  - **Standard Deviation (σ) = 3 inches**
- Find the **probability** of randomly selecting a male taller than **72 inches**.



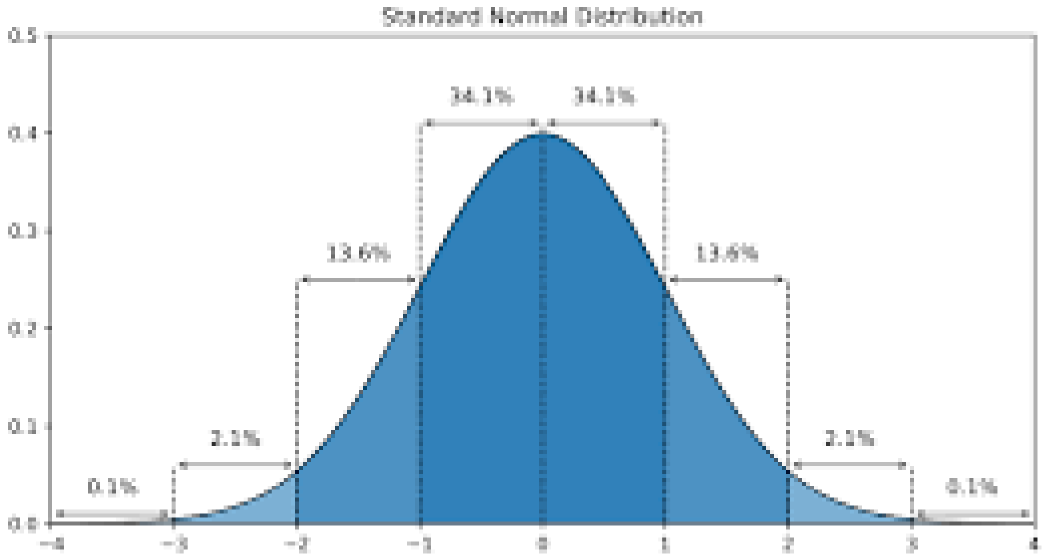


Step-by-Step Solution:

1. **Calculate the Z-Score:**
   - $Z = \frac{(X-\mu)}{\sigma} = \frac{(72-68)}{3} = \frac{4}{3} = 1.33$
2. **Use Z-tables:**
   - Z-tables tell us the area under the curve **to the left of the Z-score.**
   - For $Z = 1.33$, the Z-table gives $P(Z \leq 1.33) = 0.9082$.
   - This means **90.82% of males are 72 inches or shorter.**
3. **Find the Probability for Taller than 72 Inches:**
   - $P(Z > 1.33) = 1 - P(Z \leq 1.33)$
   - $P(Z > 1.33) = 1 - 0.9082 = 0.0918$.

So, there's a **9.18% chance** of selecting a male taller than 72 inches.

Properties of Normal distribution :



Properties of the Normal Distribution (in Easy Language)

1. **Bell-Shaped Curve**:
   - The graph of a normal distribution looks like a bell.
   - Most of the data is around the middle (mean), and the frequency decreases as you move further away.

Here's a simple explanation of the properties of the normal distribution:

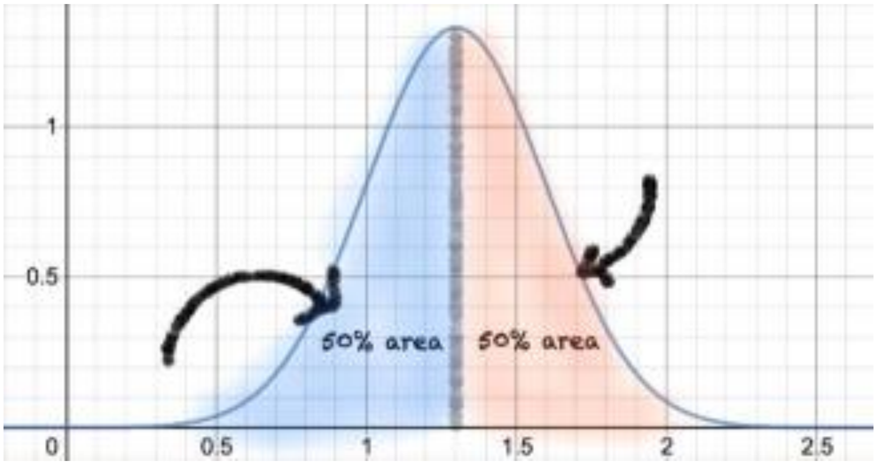
---

1. Symmetry

- **The normal distribution is symmetrical around the mean.**
- **This means the data on the left side of the mean is a mirror image of the data on the right side.**

- In simpler terms: The probability of getting a value greater than the mean is the same as the probability of getting a value smaller than the mean.

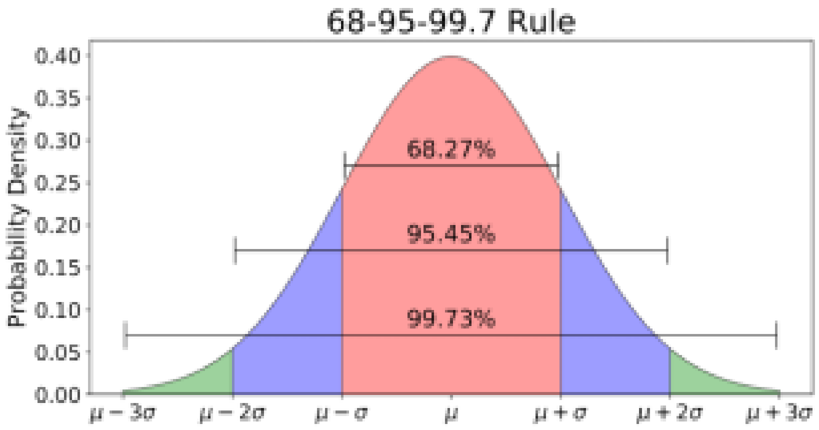
2. Measures of Central Tendencies are Equal



- In a normal distribution, the mean (average), median (middle value), and mode (most common value) are all located at the same point in the middle of the distribution.
- This means the center of the distribution is the same for all three measures, making it easier to understand the data.

3. Empirical Rule (68-95-99.7 Rule)

- The Empirical Rule tells us how data is spread out in a normal distribution:
    - 68% of the data lies within 1 standard deviation of the mean.
    - 95% of the data lies within 2 standard deviations of the mean.
    - 99.7% of the data lies within 3 standard deviations of the mean.



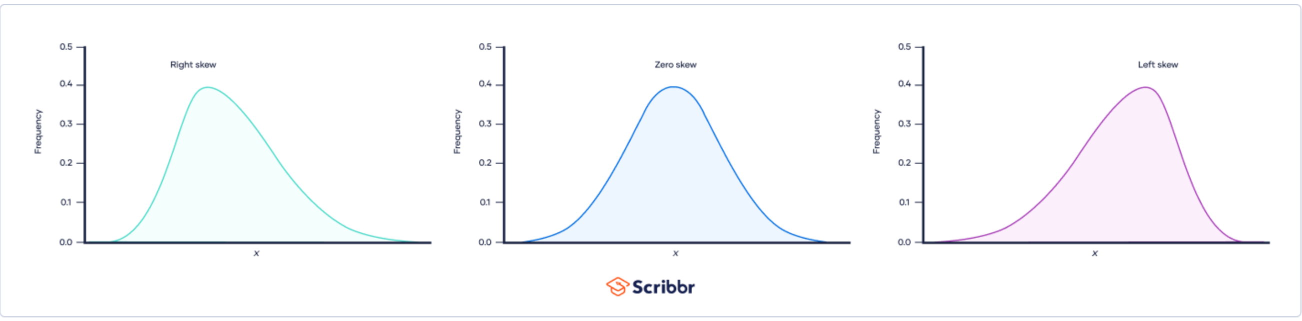
- **This rule helps us understand how most of the data is clustered around the mean and how extreme values are rare.**

4. Area Under the Curve

- **The total area under the bell-shaped curve is equal to 1 or 100%.**
- **This area represents the total probability of all possible outcomes.**
- **The area can be divided into different sections based on standard deviations to show probabilities of specific ranges of data.**

What is Skewness?

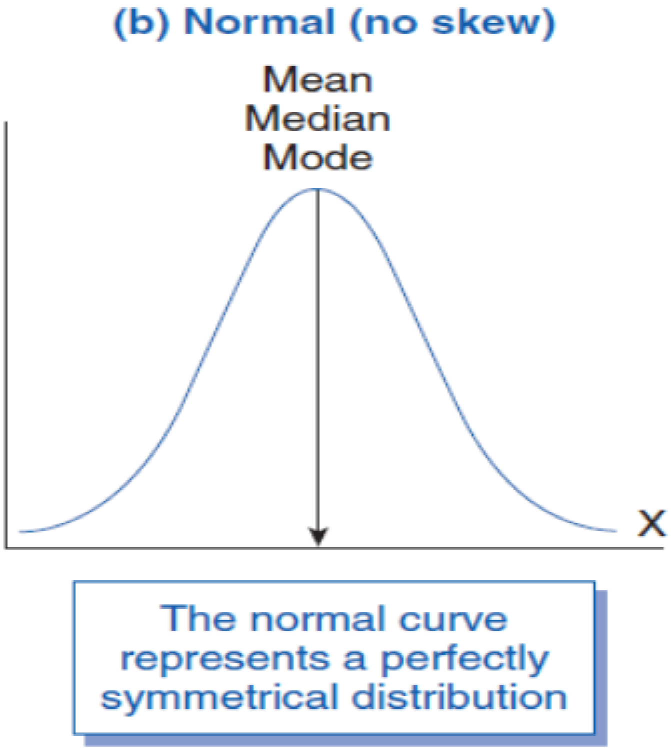
Skewness refers to how much a data distribution deviates from being symmetrical. In simpler terms, it tells us whether the data is lopsided or shifted to one side.



Click to enlarge

Understanding Skewness with an Example:

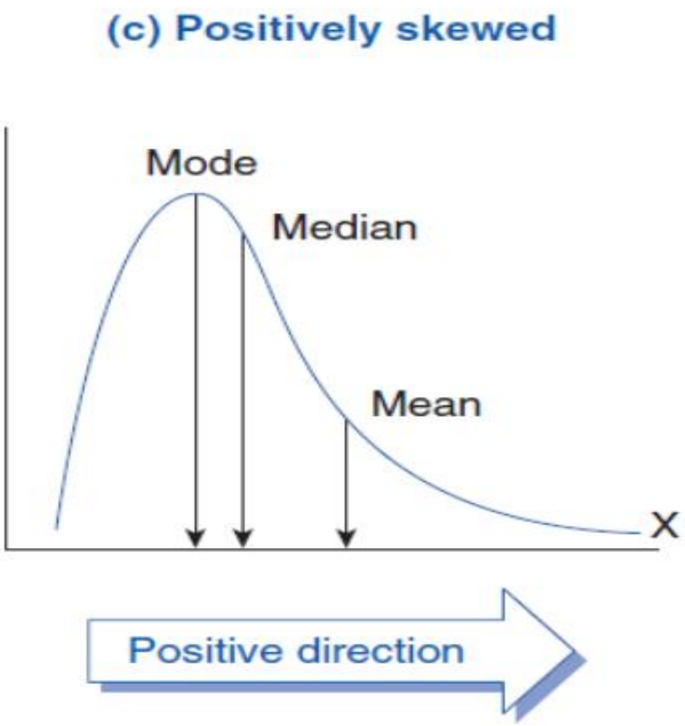
1. Symmetrical Distribution (No Skewness):


(b) Normal (no skew)
Mean
Median
Mode
X
The normal curve represents a perfectly symmetrical distribution

- Imagine you have a normal distribution (a bell-shaped curve).
- Here, the data is evenly spread around the mean.

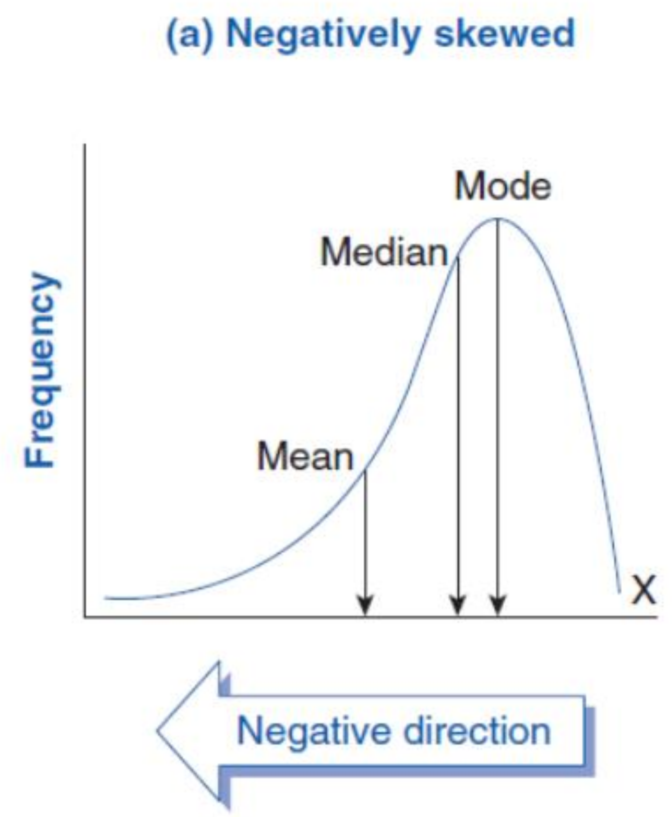
- In this case, the mean, median, and mode are all the same and are located in the center.

2. Positively Skewed Distribution (Right Skew):



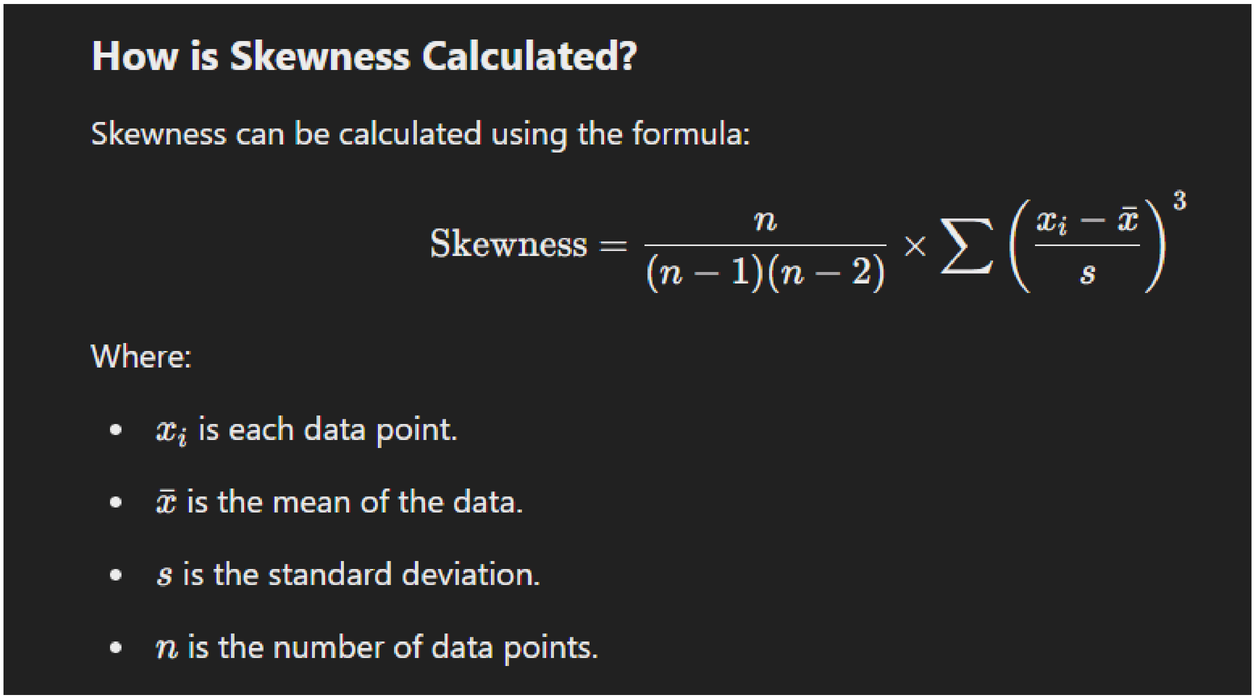
- In this case, the data is heavier on the left side, and the tail stretches out more on the right.
- Example: If you look at income data, most people might have average or below-average income, but a few people might earn very high incomes, stretching the data to the right.
- Mean > Median > Mode.

3. Negatively Skewed Distribution (Left Skew):



- In this case, the data is heavier on the right side, and the tail stretches out more on the left.
- Example: Age of retirement in a population could be negatively skewed if most people retire at the usual age (around 60-65), but a few people retire much earlier, creating a long left tail.
- Mean < Median < Mode.

4. Zero Skewness:
   - A perfectly symmetrical distribution (like a normal distribution) has zero skewness.
   - Here, the data is balanced on both sides.



How is Skewness Calculated?

Skewness can be calculated using the formula:

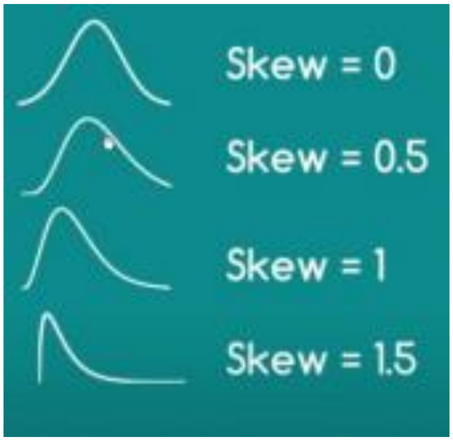
$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \times \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

Where:

- x_i is each data point.
- $\bar{x}$ is the mean of the data.
- s is the standard deviation.
- n is the number of data points.

Interpretation of Skewness:

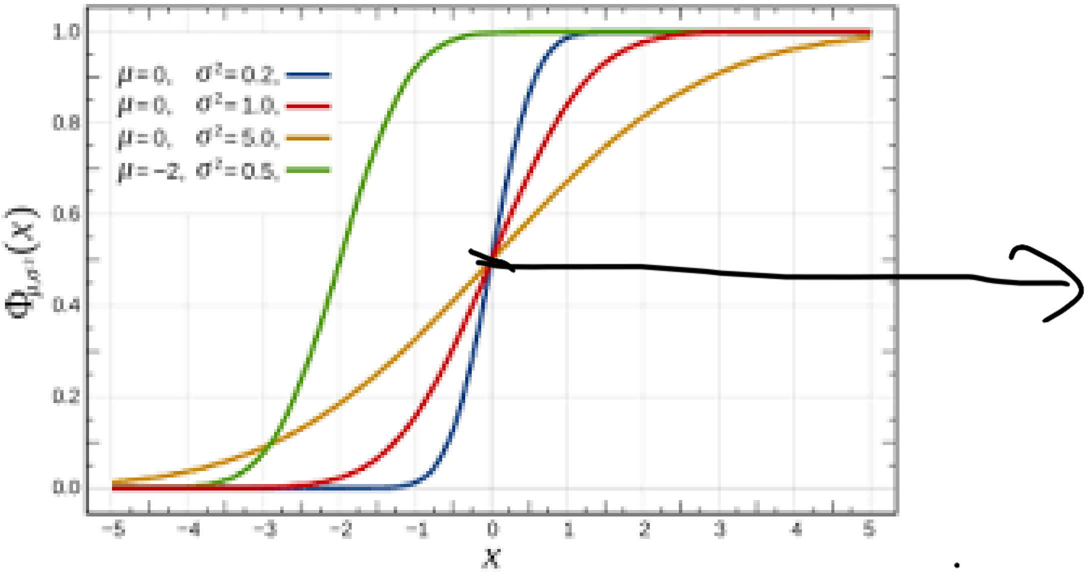
- If **skewness > 0**, the data is **positively skewed** (tail on the right).
- If **skewness < 0**, the data is **negatively skewed** (tail on the left).
- If **skewness = 0**, the data is **symmetrical**.



Why is Skewness Important?

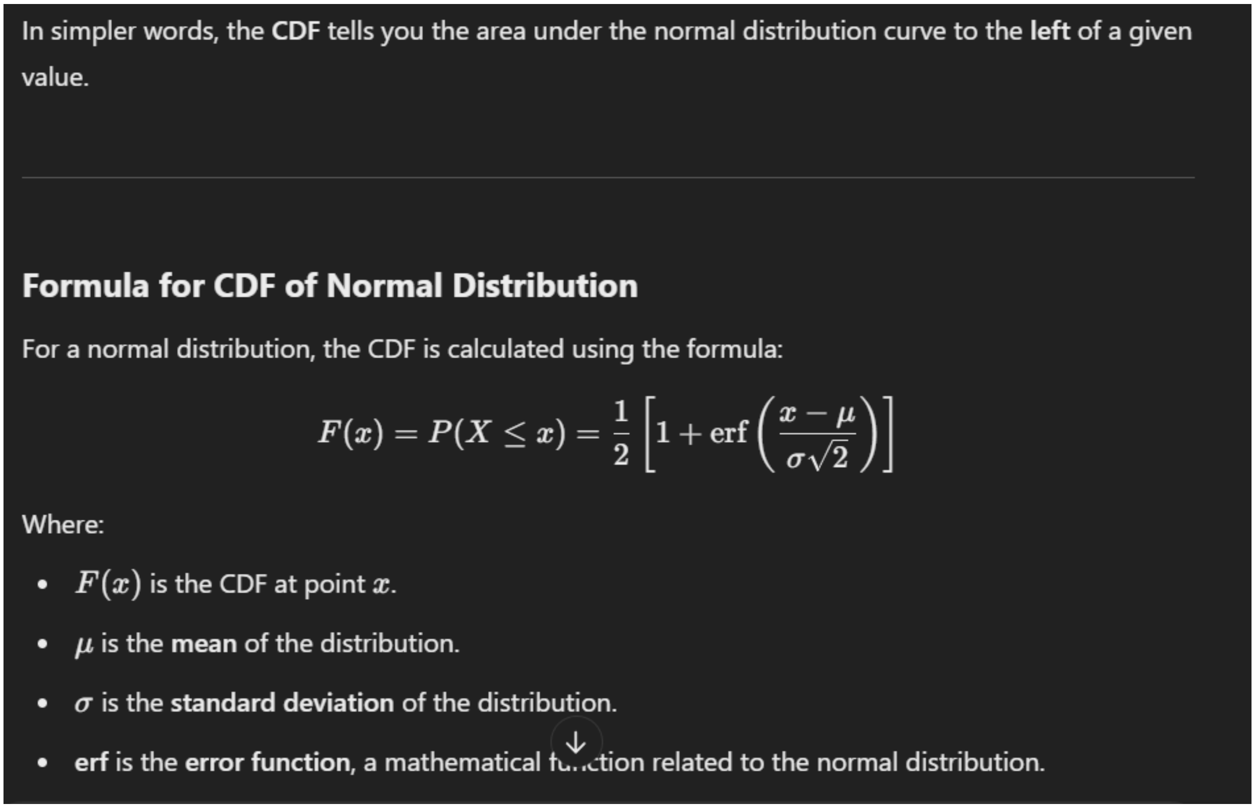
- **Understanding the Shape of the Data**: Skewness helps us know if the data follows a symmetrical pattern or if it's lopsided.
- **Identifying Outliers**: If the skewness is very high, it means that extreme values (outliers) might be affecting the data.
- **Better Analysis**: Knowing the skewness can help choose the right statistical methods or transformations for your data.

CDF of Normal Distribution



It essentially shows how much of the data is accumulated up to that point.

In simpler words, the **CDF** tells you the area under the normal distribution curve to the **left** of a given value.



In simpler words, the **CDF** tells you the area under the normal distribution curve to the **left** of a given value.

Formula for CDF of Normal Distribution

For a normal distribution, the CDF is calculated using the formula:

$$F(x) = P(X \leq x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

Where:

- $F(x)$ is the CDF at point x.
- μ is the **mean** of the distribution.
- σ is the **standard deviation** of the distribution.
- **erf** is the **error function**, a mathematical function related to the normal distribution.

Use of Normal Distribution in Data Science

The **normal distribution** is widely used in **data science** because many real-world datasets tend to follow a normal distribution or can be approximated by one. Here are some of the key areas where normal distribution plays a crucial role in data science:

---

1. Outlier Detection:

- **Outliers** are values that are significantly different from the majority of the data points. In a **normal distribution**, most data points are clustered around the **mean**, with fewer data points as you move away from the mean.

- **Outliers** are often identified as values that fall **far outside** of the normal range, typically beyond **3 standard deviations** from the mean (in a normal distribution, about 99.7% of data falls within 3 standard deviations).
- **Example**: If a dataset has a mean of 100 and a standard deviation of 10, any data point beyond **130** (100 + 3*10) or below* ***70*** *(100 - 3*10) could be considered an outlier.

---

2. Assumptions for Machine Learning Algorithms:

Many machine learning algorithms assume that the data is normally distributed. This assumption helps in better model performance and results. Some common algorithms where normal distribution plays a role are:

a. Linear Regression:

- Linear regression assumes that the **errors** or residuals (the differences between predicted and actual values) follow a normal distribution. This assumption allows us to apply statistical tests like **hypothesis testing** and **confidence intervals**.
- **Why it matters**: If errors are not normally distributed, the model's predictions may be biased, and statistical tests could give misleading results.

b. Gaussian Mixture Models (GMM):

- GMM is a probabilistic model used for clustering, where it assumes that data is generated from a mixture of several normal distributions (Gaussians).
- **Why it matters**: By modeling each cluster as a normal distribution, GMM can fit complex data distributions and cluster the data more effectively.

---

3. Hypothesis Testing:

- **Hypothesis testing** often relies on the assumption that the data follows a normal distribution (or can be approximated by it). In many statistical tests, such as the **t-test** and **Z-test**, we compare the means of two groups to determine if they are significantly different.

- If the data follows a normal distribution, we can calculate the probability (p-value) to decide whether to **reject** or **fail to reject** the null hypothesis.
- **Example**: In testing whether a new drug has a significant effect on blood pressure, if the blood pressure measurements follow a normal distribution, hypothesis testing will help determine if the change is statistically significant.

---

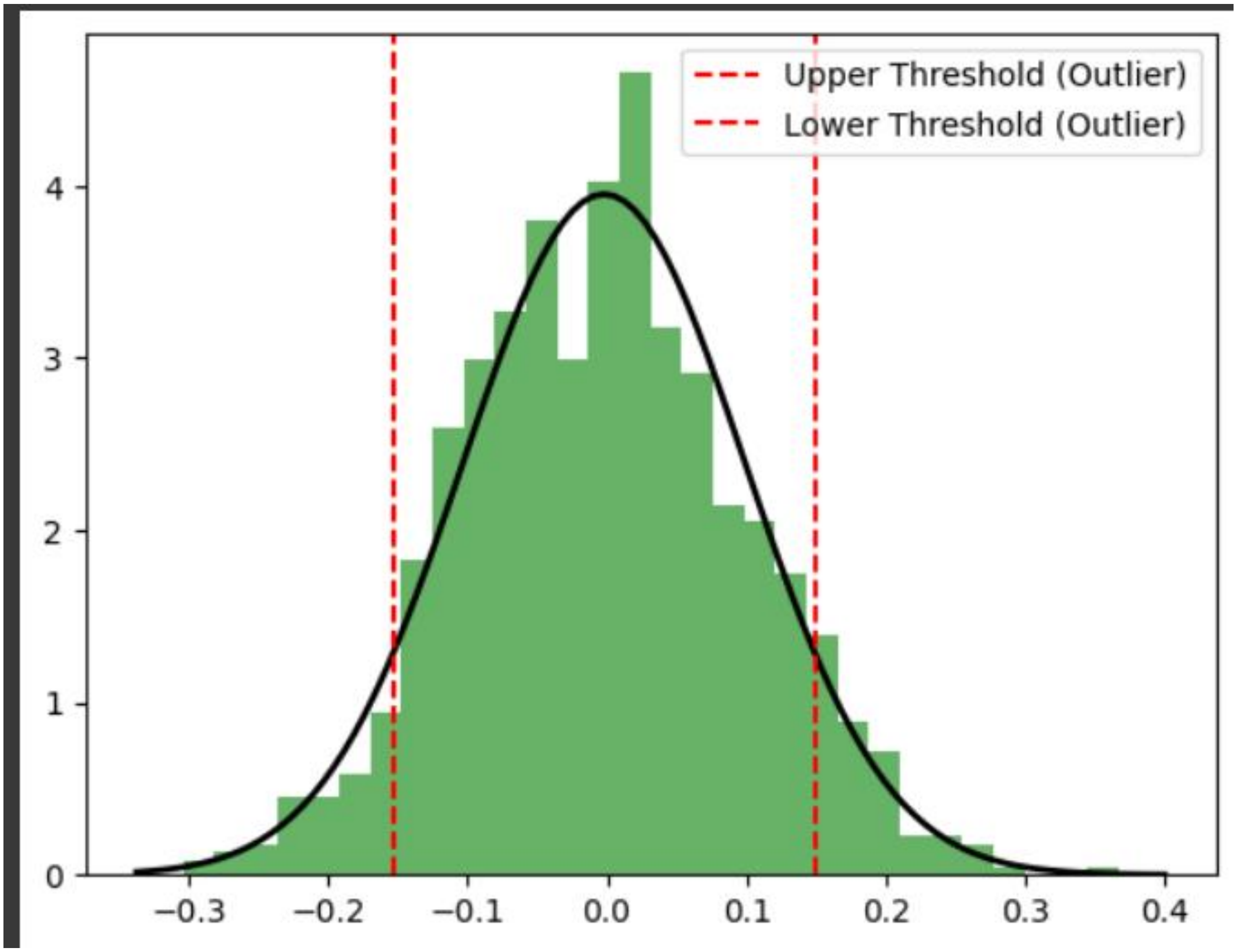
4. Central Limit Theorem (CLT):

- The **Central Limit Theorem** is one of the most powerful and important principles in statistics and data science.
- **What it says**: The Central Limit Theorem states that if we take **multiple random samples** from any population (no matter what distribution the population follows), the **sampling distribution of the sample mean** will be approximately **normal**, **as the sample size increases**.
- **Why it matters**: This allows us to perform statistical analysis even if the original data is not normally distributed. As long as the sample size is large enough, we can treat the sample mean as normally distributed and apply statistical methods like confidence intervals, hypothesis testing, etc.
- **Example**: If you're working with a dataset of customer satisfaction scores that are not normally distributed, as long as you take large enough random samples, the **average** of those samples will follow a normal distribution. This lets you perform further statistical analysis even on non-normal data.

Summary:

- **Outlier detection**: Identifying extreme values that don't fit the normal pattern of the data.
- **Machine Learning Assumptions**: Many ML algorithms, like linear regression and GMM, rely on the assumption that data follows a normal distribution.

- **Hypothesis Testing**: Statistical tests often assume normal distribution for evaluating hypotheses.
- **Central Limit Theorem**: Even if data is not normally distributed, sample means will tend to follow a normal distribution as sample size increases.



•