Understanding the Normal Distribution

A normal distribution is bell-shaped and symmetrical around the mean.

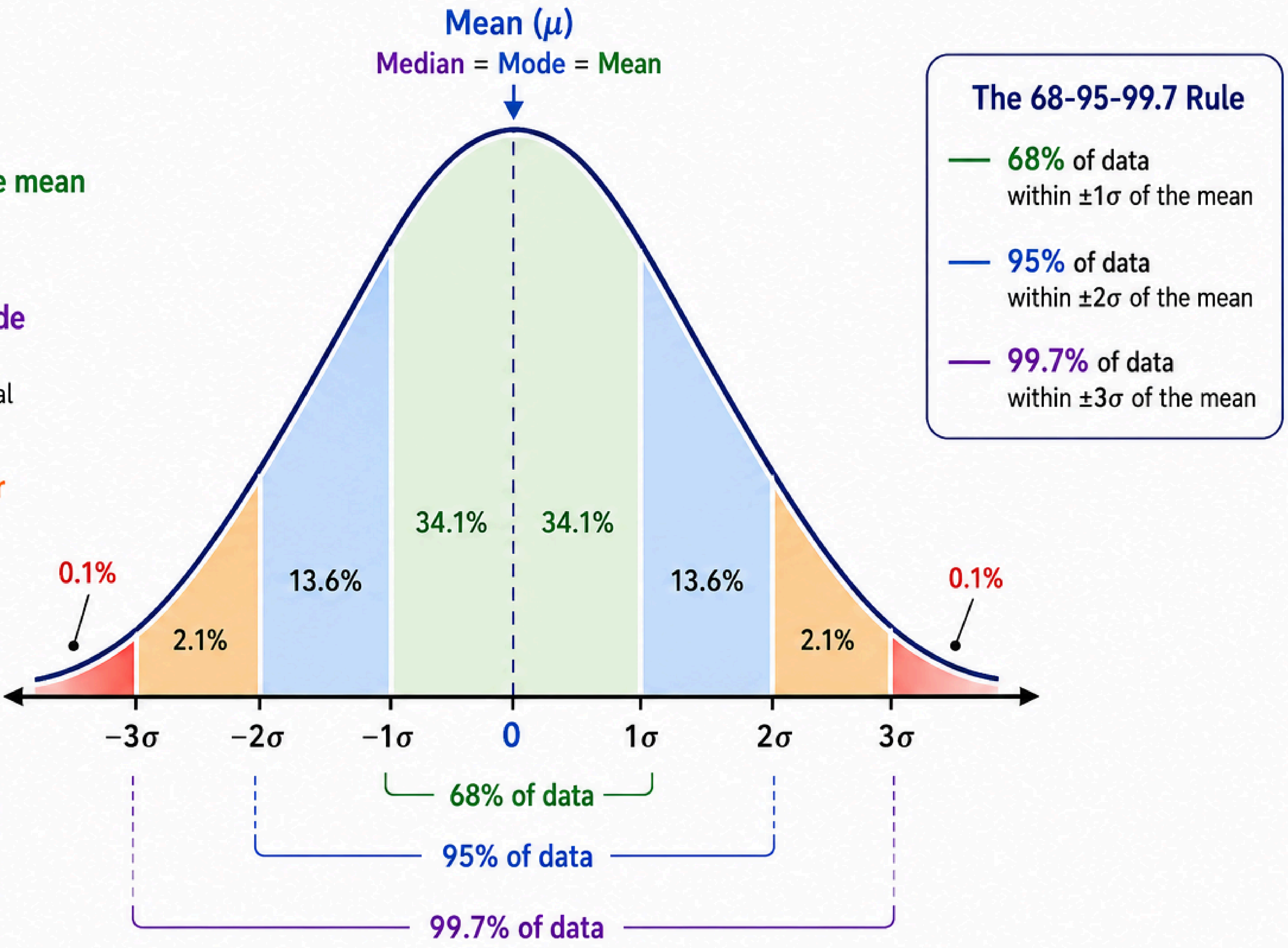
Bell-shaped curve
The normal distribution forms a smooth, bell shape.

Symmetrical around the mean
The left side is a mirror image of the right side.

Mean, median, and mode are equal
All three measures of central tendency are the same.

Most observations occur near the center
The highest frequency of values is around the mean.

Extreme values are rare
Values far from the mean are less likely to occur.



Pankaj Chouhan
www.codeswithpankaj.com
www.cwpc.in

Mean

The mean is the average value of a dataset.

$$\text{Mean} = \frac{\sum X}{N}$$

Example

Dataset: 2, 4, 6, 8, 10

$$\text{Mean} = \frac{2+4+6+8+10}{5} = \frac{30}{5} = 6$$

Key Points

- Uses all values in the dataset.
- Sensitive to extreme values.
- Represents the center or balance point of the data.

Pankaj Chouhan

www.codeswithpankaj.com

www.cwpc.in

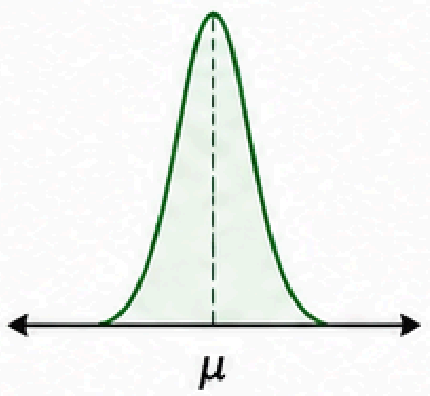
Standard Deviation

Standard deviation measures how spread out the data is around the mean.

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

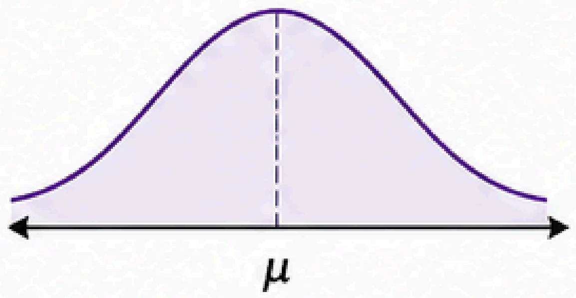
- Small standard deviation → values are close to the mean.
- Large standard deviation → values are spread out.

Small Standard Deviation



Values are close to the mean

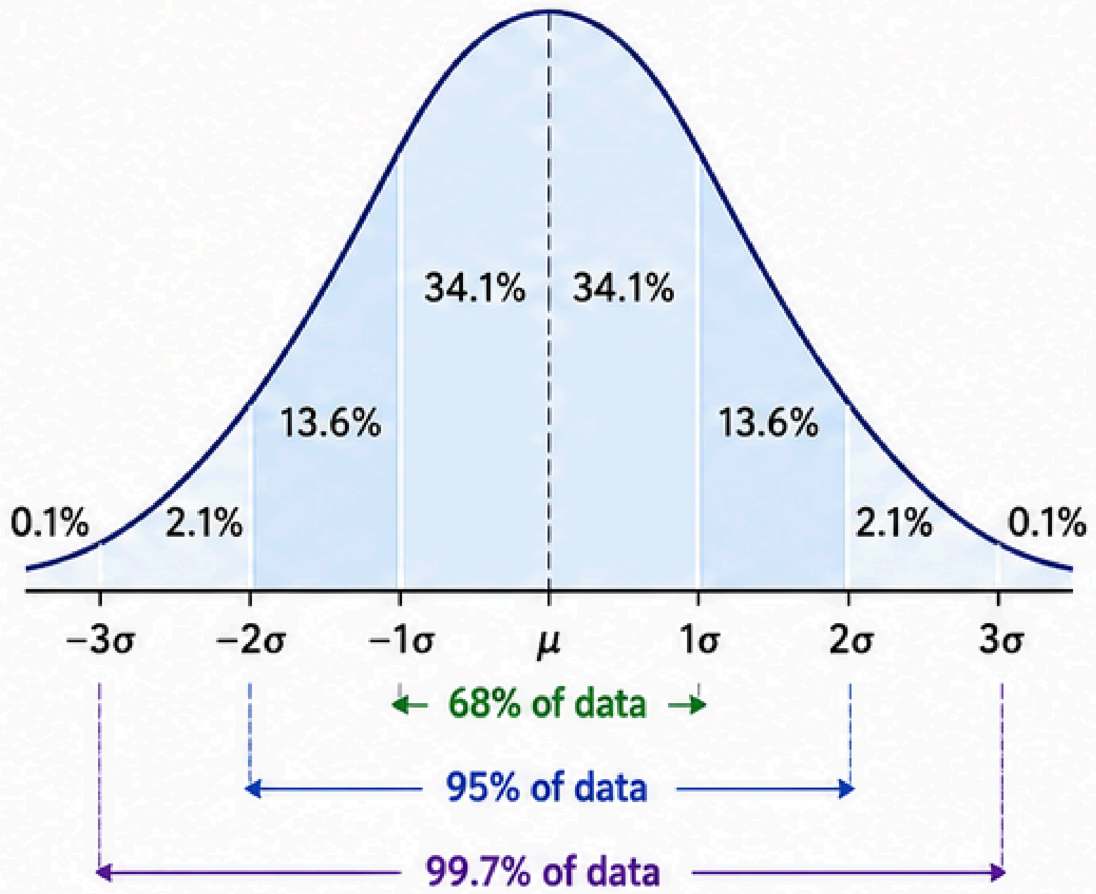
Large Standard Deviation



Values are spread out from the mean

Normal Distribution

A normal distribution is a symmetric, bell-shaped distribution where most values cluster around the mean.



Many real-world measurements approximately follow a normal distribution:

Student exam scores | Heights | Weights | IQ scores | Manufacturing measurements

The concept of **Z-score** is built directly on these ideas.

Mean and Standard Deviation Example

Let's understand mean and standard deviation with clear examples.

Calculate the Mean

Consider the following marks:

60, 70, 80, 90, 100

$$\text{Mean} = \frac{60 + 70 + 80 + 90 + 100}{5}$$

$$\text{Mean} = \frac{400}{5}$$

Mean = 80

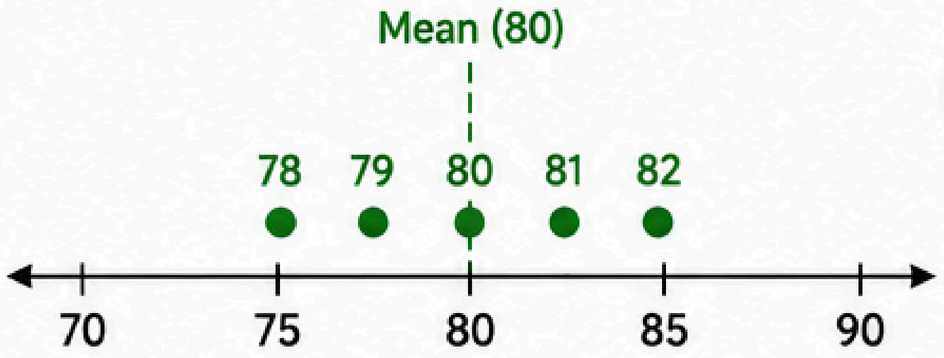
The average score is **80**.

Understanding Standard Deviation

Dataset with Small Spread

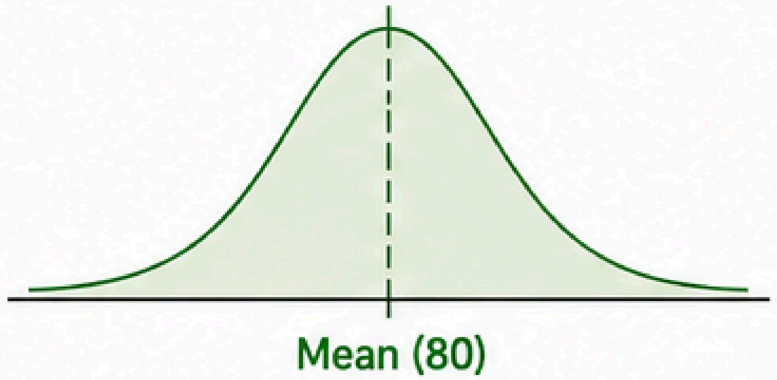
78, 79, 80, 81, 82

Values are very close to the mean.



Result:

- Low spread
- Small standard deviation

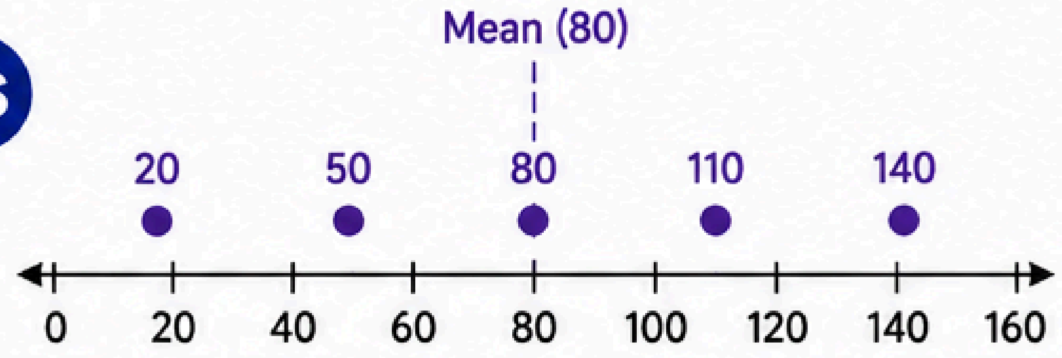


VS

Dataset with Large Spread

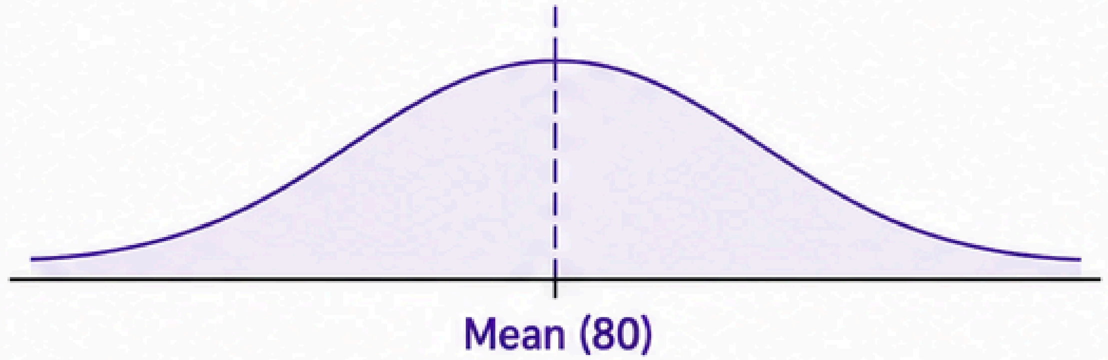
20, 50, 80, 110, 140

Values are far from the mean.



Result:

- High spread
- Large standard deviation



Pankaj Chouhan

www.codeswithpankaj.com

www.cwpc.in

The 68–95–99.7 Rule

For a normal distribution:

1 Within 1 Standard Deviation

Mean ± 1SD

Approximately **68%** of data falls within this range.

2 Within 2 Standard Deviations

Mean ± 2SD

Approximately **95%** of data falls within this range.

3 Within 3 Standard Deviations

Mean ± 3SD

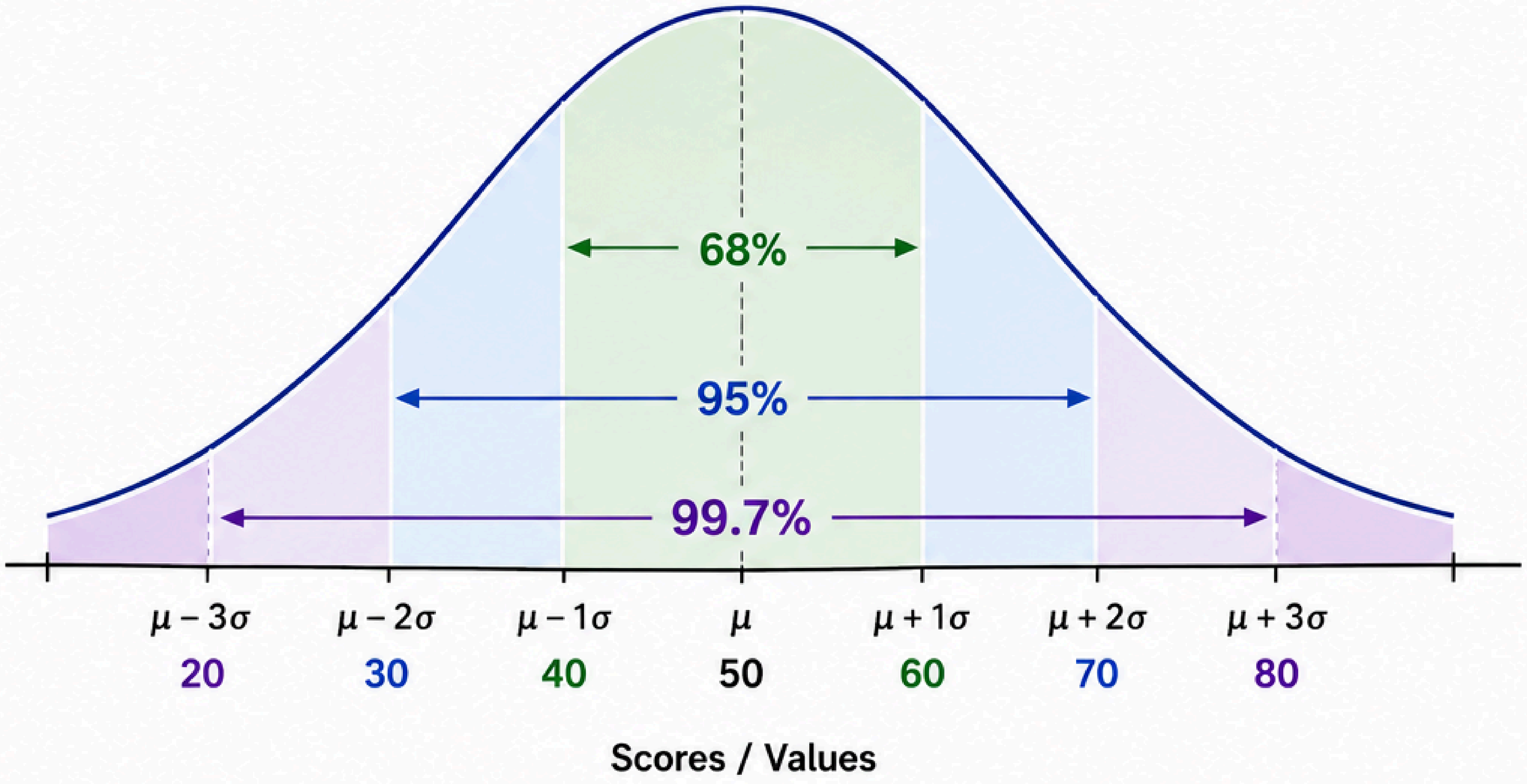
Approximately **99.7%** of data falls within this range.

Example of the Rule

Suppose:

Mean (μ)	= 50
Standard Deviation (σ)	= 10

68% of Data	40 to 60
95% of Data	30 to 70
99.7% of Data	20 to 80



Pankaj Chouhan

www.codeswithpankaj.com

www.cwpc.in

What is a Z-Score?

A Z-score measures how many **standard deviations** a value is away from the mean.

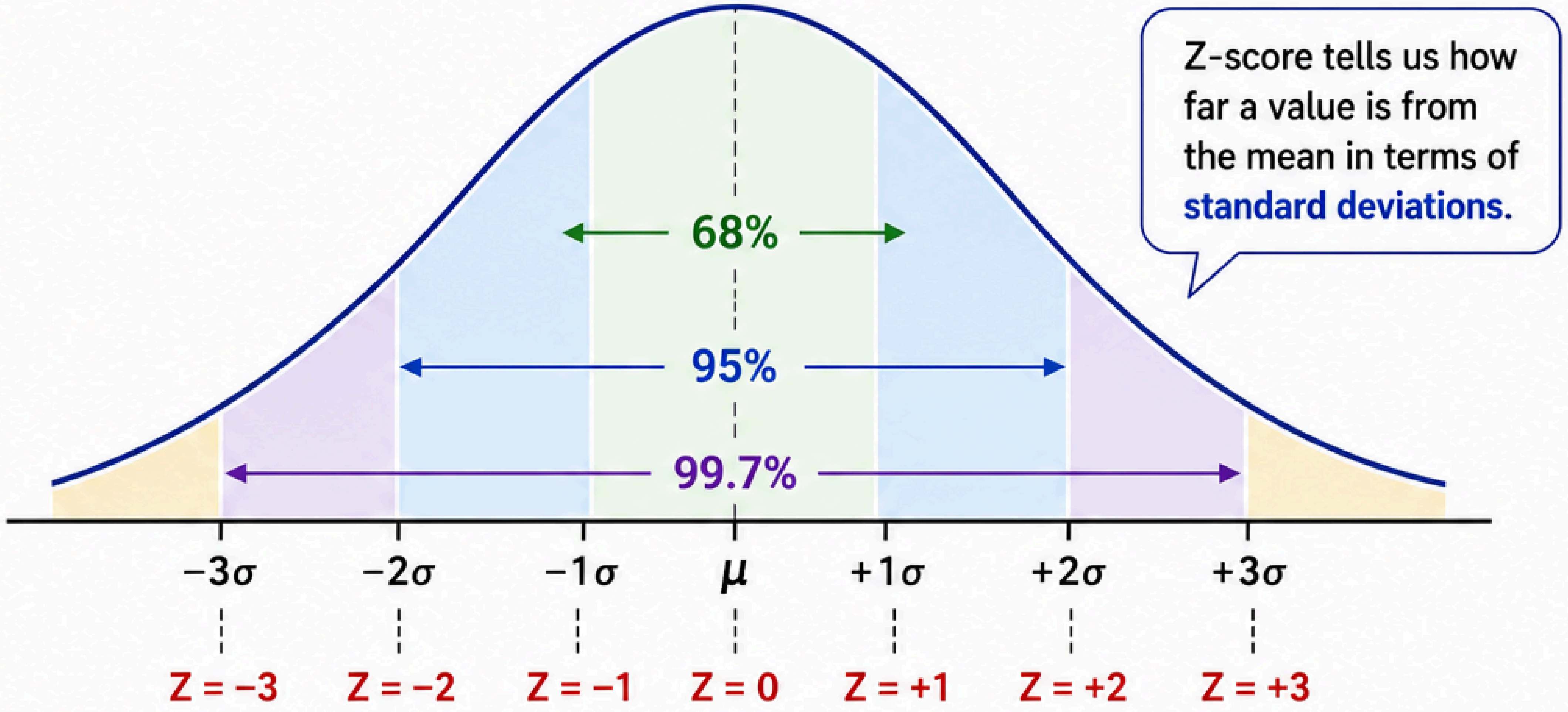
Formula

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

Where:

- X = Actual value
- μ = Mean
- σ = Standard deviation

Understanding Z-Score



Interpretation

- $Z = 0$ The value is equal to the mean.
- $Z > 0$ The value is above the mean.
- $Z < 0$ The value is below the mean.

The larger the absolute value of Z, the **farther the value** is from the mean.

Pankaj Chouhan

www.codeswithpankaj.com

www.cwpc.in

Calculating a Z-Score

Let's calculate the Z-score for a student's score.

Suppose:

- Mean (μ) = 70
- Standard Deviation (σ) = 10
- Student Score (X) = 90

Step 1: Apply the Formula

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

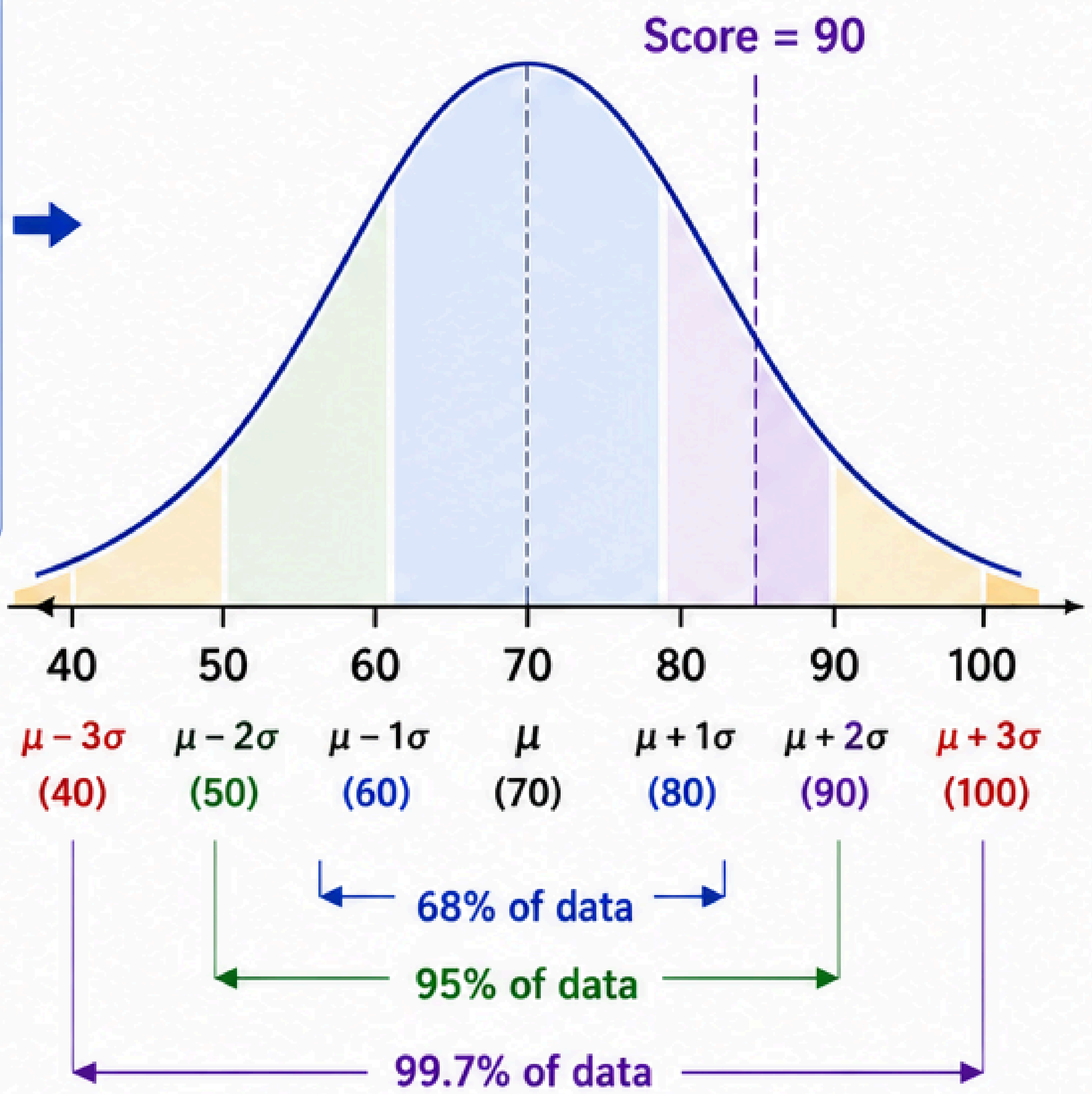
$$Z = \frac{90 - 70}{10}$$

Step 2: Simplify

$$Z = \frac{20}{10}$$

$$Z = 2$$

Visual Representation



Interpretation

- The score is 2 standard deviations **above** the mean.
- This indicates a significantly **above-average** performance.

Pankaj Chouhan

www.codeswithpankaj.com

www.cwpc.in

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

Positive Z-score → Above average | Z = 0 → Average (at the mean) | Negative Z-score → Below average

Negative Z-Score Example

Let's calculate and interpret a negative Z-score.

Suppose:

- Mean (μ) = 70
- Standard Deviation (σ) = 10
- Student Score (X) = 50

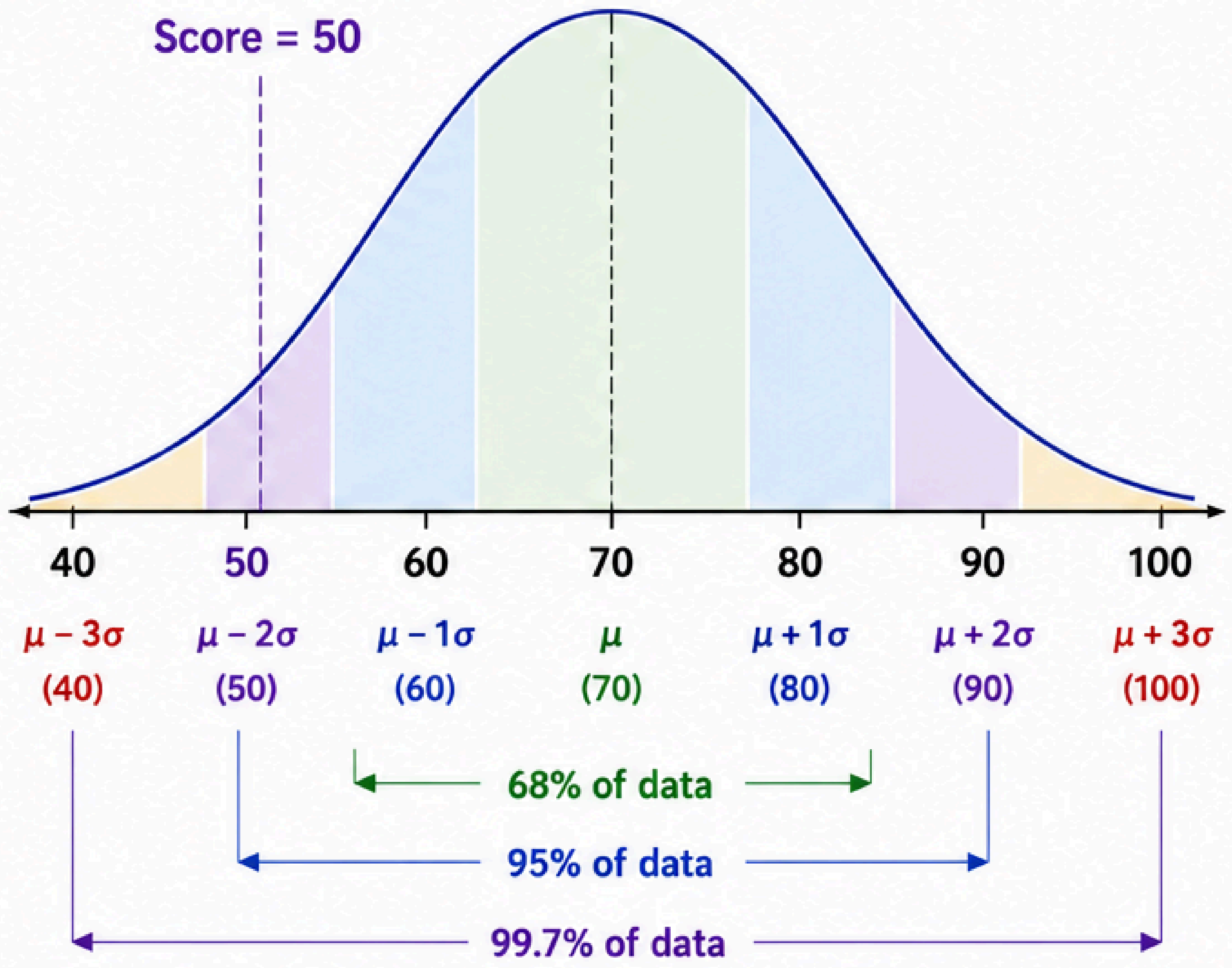
Calculation

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

$$Z = \frac{50 - 70}{10}$$

$$Z = -2$$

Position on the Normal Distribution



Interpretation

The score is 2 standard deviations **below the mean.**

This indicates **below-average** performance.

Pankaj Chouhan
www.codeswithpankaj.com
www.cwpc.in

What does a negative Z-score mean?

A negative Z-score means the value is **below the mean.**

Z = 0	At the mean (average)
Z > 0	Above the mean
Z < 0	Below the mean

Key Takeaway

The further the Z-score from 0, the farther the score is from the mean.

Why Do We Need Z-Scores?

Imagine two students from different schools.

Student A

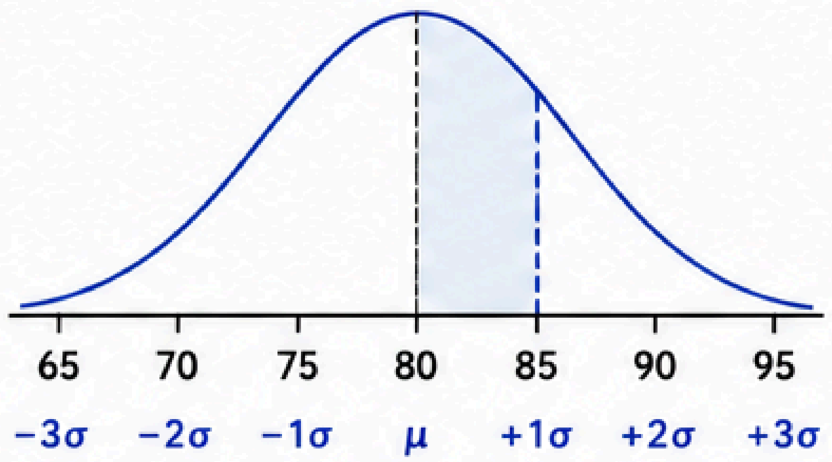
- Score (X) = 85
- Mean (μ) = 80
- Standard Deviation (σ) = 5

Z-Score Calculation

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

$$Z = \frac{85 - 80}{5}$$

Z = 1



Score is 1 standard deviation above the mean.

Student B

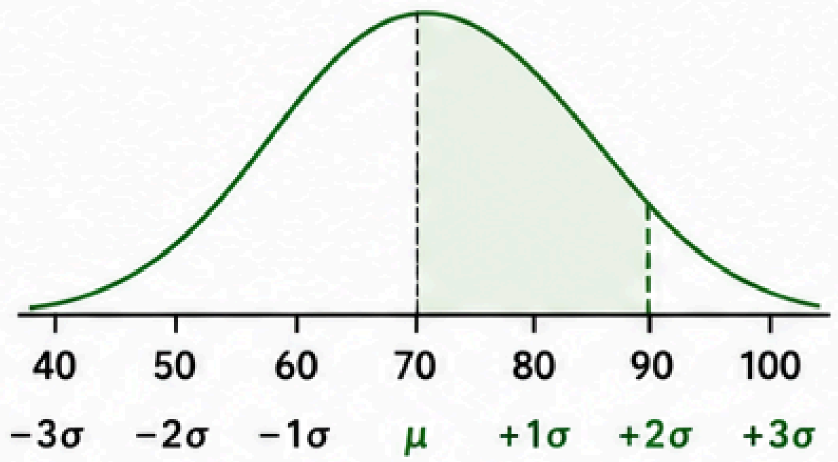
- Score (X) = 90
- Mean (μ) = 70
- Standard Deviation (σ) = 10

Z-Score Calculation

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

$$Z = \frac{90 - 70}{10}$$

Z = 2



Score is 2 standard deviations above the mean.

Although Student B has a higher raw score (90 > 85), Student B's performance is better relative to his school.

Z-scores allow us to compare performance across different datasets with different means and spreads.

Pankaj Chouhan

www.codeswithpankaj.com

www.cwpc.in

Comparison

Although **Student B** scored only 5 marks more, their performance relative to their school is **much stronger**.

Student A

- Score = 85
- Mean = 80
- Standard Deviation = 5

Z-Score = 1

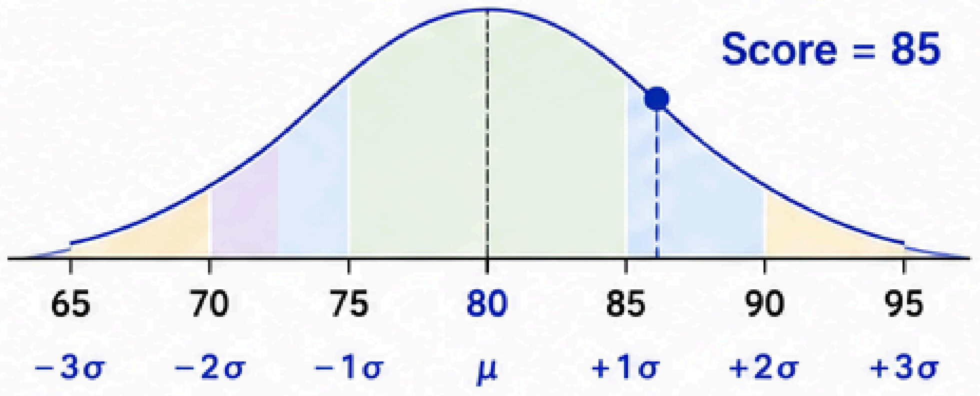
Student	Score (Marks)	Z-Score
A	85	1
B	90	2

Student B

- Score = 90
- Mean = 70
- Standard Deviation = 10

Z-Score = 2

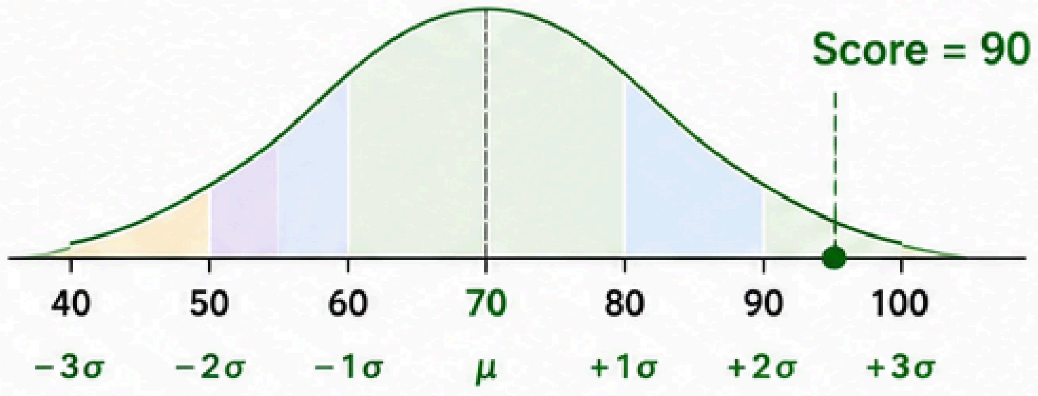
Student A (Mean = 80, SD = 5)



85 is 1 **standard deviation** above the mean.

Student **B** performed **better** compared to their peer group.

Student B (Mean = 70, SD = 10)



90 is 2 **standard deviations** above the mean.

This is why **Z-scores** are useful for **comparing different datasets**.

Pankaj Chouhan
www.codeswithpankaj.com
www.cwpc.in

Raw Scores Can Be Misleading
Different schools (or tests) have different means and spreads.

Z-Scores Standardize Performance
They tell us how far a score is from the mean in terms of standard deviations.

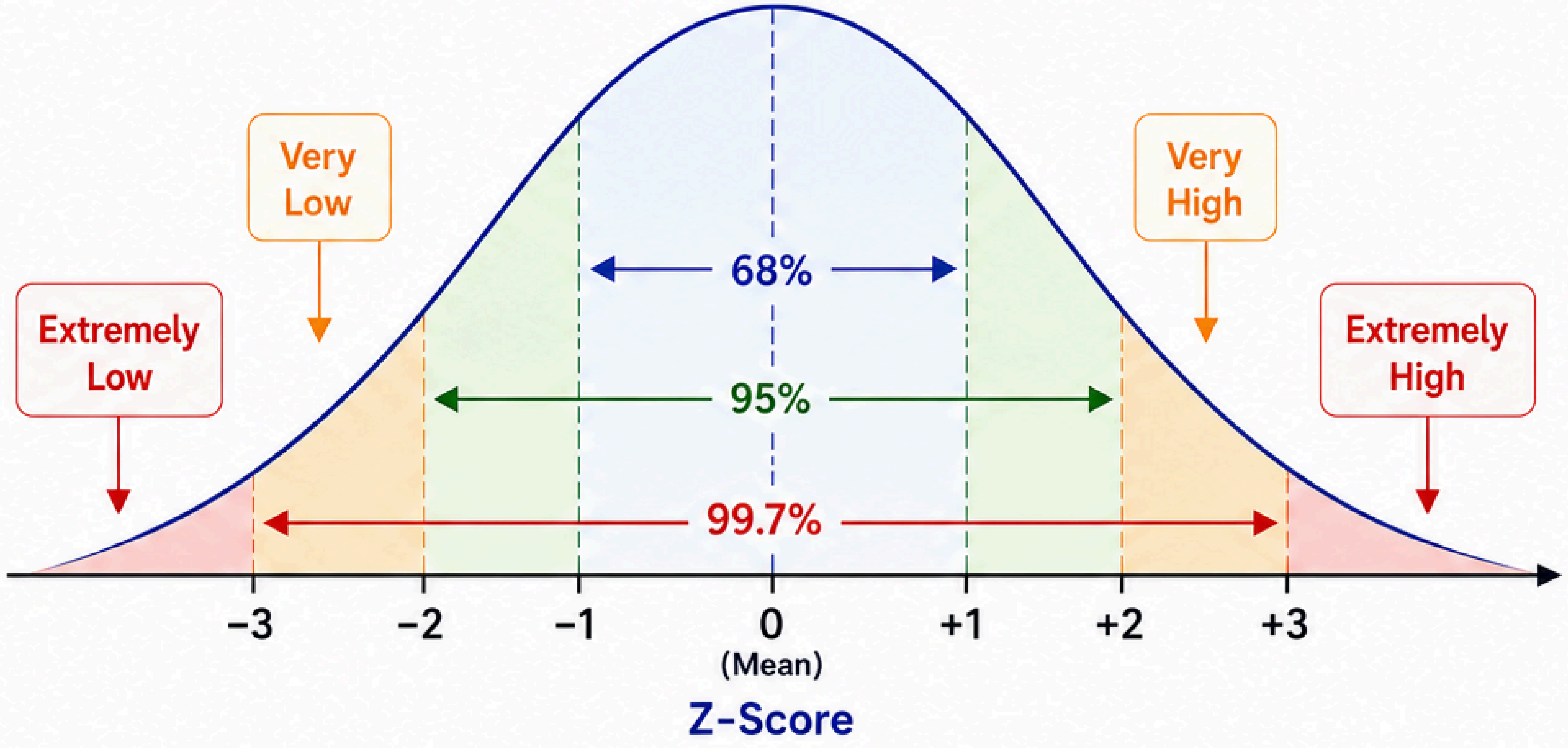
Better for Comparison
Z-scores allow fair comparison across different datasets.

Identifying Unusual Values

Z-scores can help detect **unusual observations.**

General Interpretation	
Z-Score	**Interpretation**
0	Average
+1	Above Average
−1	Below Average
+2	Very High
−2	Very Low
+3	Extremely High
−3	Extremely Low

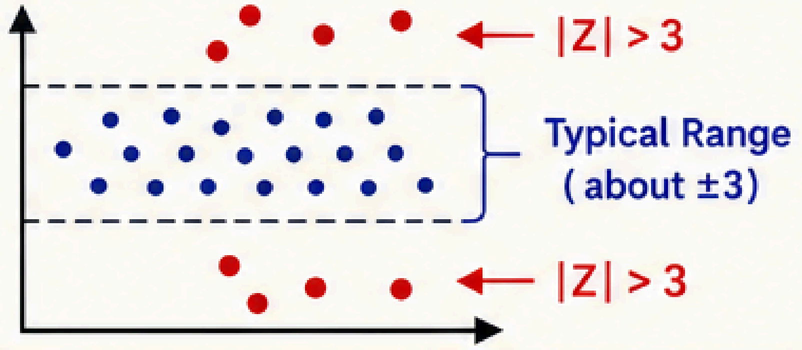
Normal Distribution and Z-Score Regions



A common rule is:

|Z| > 3

Values beyond **±3** are often considered **unusual** or potential **outliers.**



Statistics is the foundation of Machine Learning.
Understand the data. Build the future.

Pankaj Chouhan
www.codeswithpankaj.com
www.cwpc.in

Real-World Example: Salary Analysis

1 Suppose:

- Average Salary (μ) = ₹50,000
- Standard Deviation (σ) = ₹10,000
- Your Salary (X) = ₹70,000

2 Calculation

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

$$Z = \frac{70000 - 50000}{10000}$$

$$Z = 2$$

3 Interpretation

Your salary is **2 standard deviations above** the average salary.

This indicates that your income is **significantly higher** than the average employee in the dataset.

4 Standard Normal Distribution

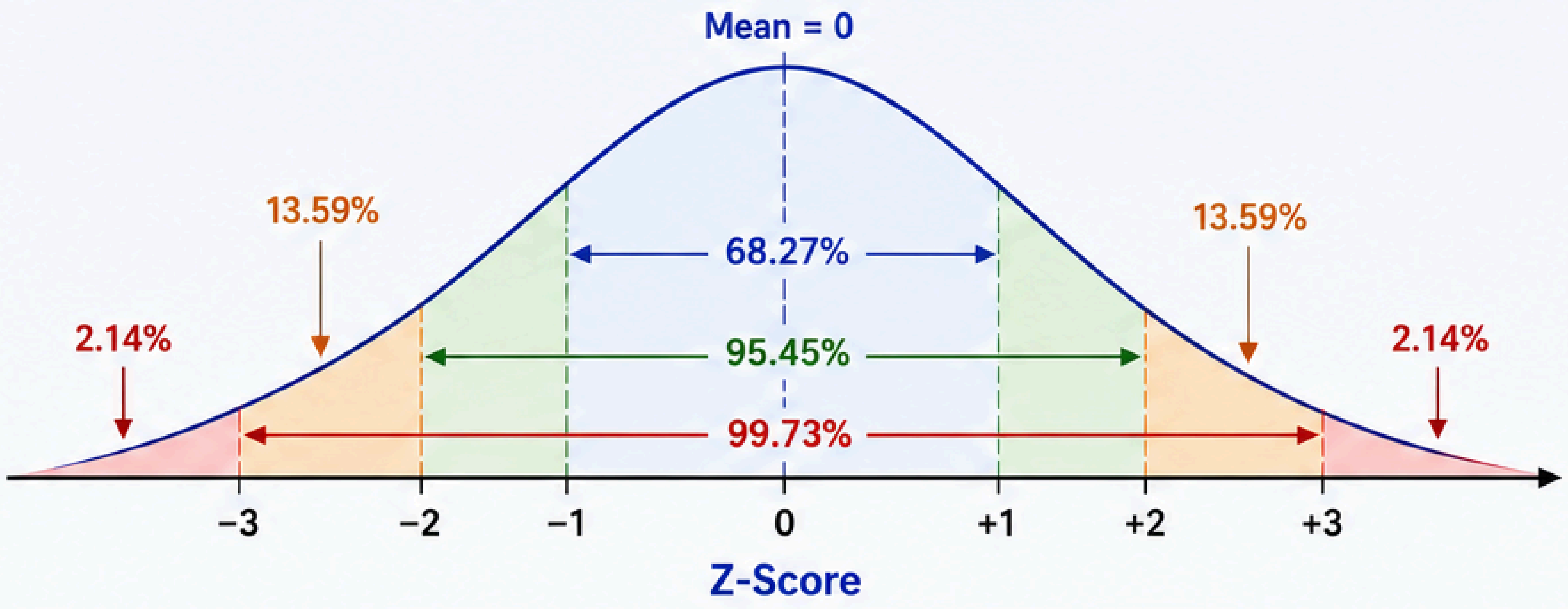
When data is converted into Z-scores:

- Mean becomes **0**
- Standard deviation becomes **1**

This transformed distribution is called the **Standard Normal Distribution.**

It allows different datasets to be **compared on the same scale.**

Standard Normal Distribution (Z)



Key Takeaway:
Z-scores help us **understand** how a value compares to the rest of the data, **no matter** what the original units or scale are.

Pankaj Chouhan

www.codeswithpankaj.com | www.cwpc.in

T-SCORES IN STATISTICS

T-Score is used when **sample sizes are small ($n < 30$)** and **population standard deviation (σ) is unknown.**

WHEN DO WE USE T-SCORE?

- Sample size is small (typically $n < 30$)
- Population standard deviation (σ) is **unknown**
- We estimate the population using **sample data**

T-SCORE FORMULA

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

$\bar{X}$ = Sample Mean

μ = Population Mean

S = Sample Standard Deviation

n = Sample Size

KEY CONCEPTS

- T-Score tells us how many **standard errors** a sample mean is away from the population mean.
- As sample size (n) increases, the T-Score approaches the **Z-Score**.
- T-Scores follow a t-distribution (not the normal distribution) with **$n - 1$ degrees of freedom**.

STANDARD ERROR OF THE MEAN (SEM)

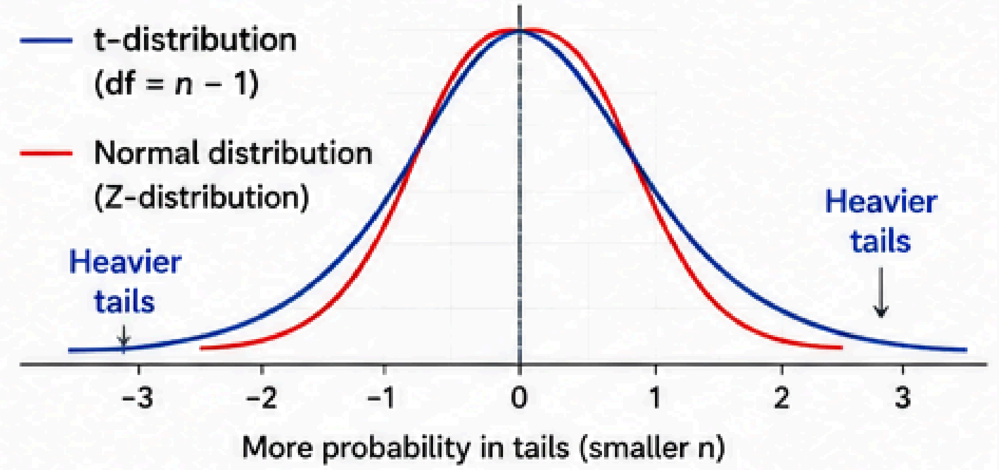
The Standard Error (SE) measures how far the sample mean is likely to be from the population mean. →

$$SE = \frac{S}{\sqrt{n}}$$

Smaller SE ⇒ more precise estimate

Larger SE ⇒ less precise estimate

t-DISTRIBUTION (vs NORMAL DISTRIBUTION)



EXAMPLE

Sample Mean ($\bar{X}$)	=	72
Population Mean (μ)	=	70
Sample Standard Deviation (S)	=	8
Sample Size (n)	=	16

$$T = \frac{72 - 70}{8 / \sqrt{16}} = \frac{2}{8/4} = \frac{2}{2} = 1.00$$

NOTE: When $n \geq 30$, the t-distribution is very close to the normal distribution, so we can use **Z-Score** instead of T-Score.

Pankaj Chouhan

www.codeswithpankaj.com

www.cwpc.in

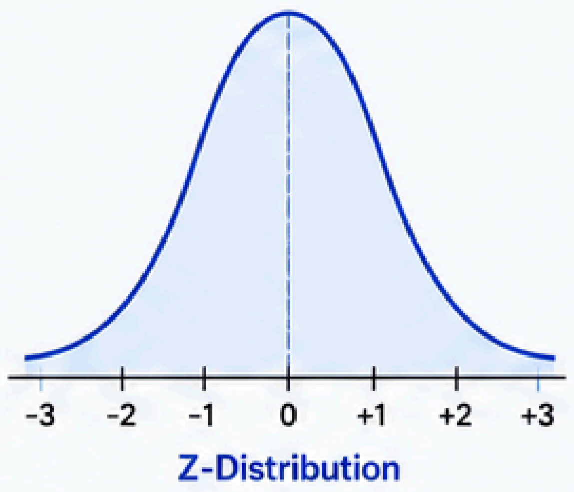
T-Scores help us make reliable inferences from **small samples** when population standard deviation is **unknown**.

Why Do We Need a T-Score?

T-Score helps us make reliable inferences when we have a **small sample size** and the **population standard deviation (σ) is unknown.**

1. Z-SCORE: When It Works

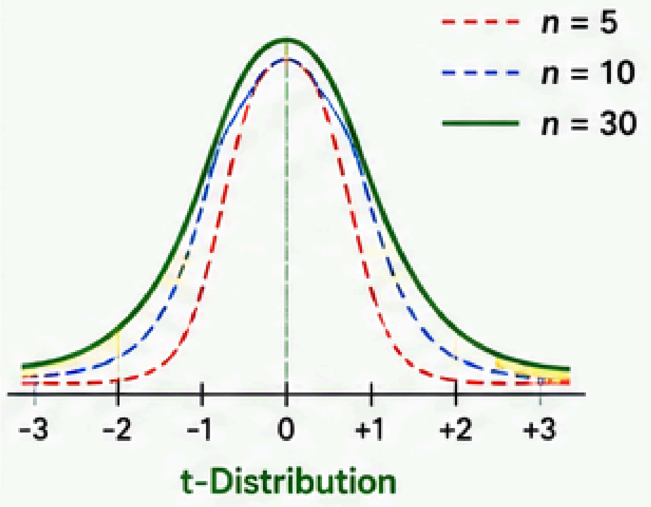
- Large sample size (typically $n \geq 30$)
- Population standard deviation (σ) is known
- We can use Z-Score (Standard Normal Distribution)



VS

2. T-SCORE: When It's Needed

- Small sample size (typically $n < 30$)
- Population standard deviation (σ) is unknown
- We estimate σ using sample standard deviation (s) and use T-Score (t-Distribution)



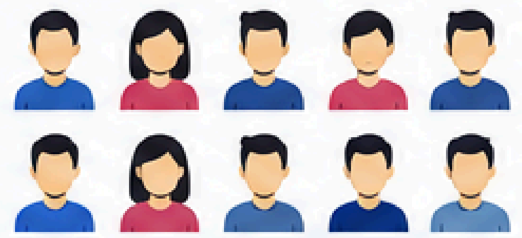
Real-World Example

Scenario

A company wants to know if the average salary of 10 new employees is different from the population average. The population standard deviation is unknown.

→

Sample Data (n = 10)



We calculate sample mean ($\bar{x}$) and sample standard deviation (s)

→

What if we used Z-Score?

We don't know the true σ

- Z-Score would give incorrect results and misleading conclusions.

Why T-Score is the Right Choice?

T-Score adjusts for extra uncertainty in small samples by using t-distribution (which has heavier tails).

→

Result

T-Score gives more accurate and reliable inference for small samples with unknown σ.

Key Takeaway

Use T-Score when:

- Sample size is small ($n < 30$)
- Population standard deviation (σ) is unknown

T-Score Formula

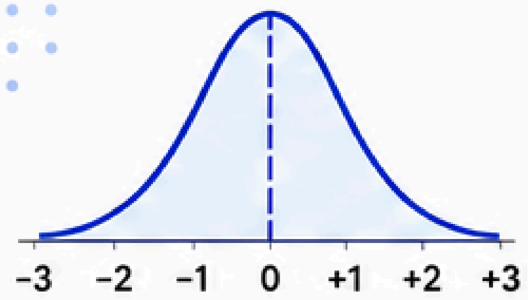
$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

$\bar{x}$ = Sample Mean
μ = Population Mean
s = Sample Standard Deviation
n = Sample Size

As sample size (n) increases, t-distribution approaches the normal distribution (Z-distribution).

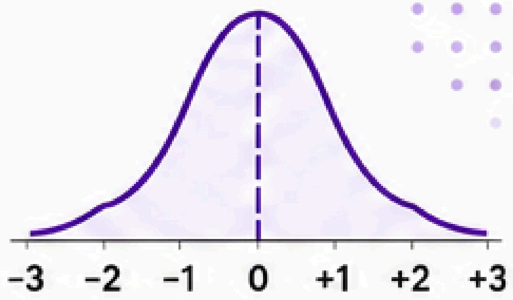
Pankaj Chouhan
www.codeswithpankaj.com
www.cwpc.in

T-Score ensures **better decisions** when data is limited and uncertainty is high. It's a powerful tool for **real-world** statistical analysis!



Z-SCORE vs T-SCORE

VS



Feature	Z-Score	T-Score
Population SD Known	Yes	No
Sample Size	Large	Small
Distribution Used	Normal Distribution	T Distribution
More Accurate for Small Samples	No	Yes

In Simple Words: Use **Z-Score** when you know the population standard deviation and have a large sample.

Use **T-Score** when the population standard deviation is unknown and the sample size is small.

What is a T-Distribution?

The t-distribution looks similar to a normal distribution but has:

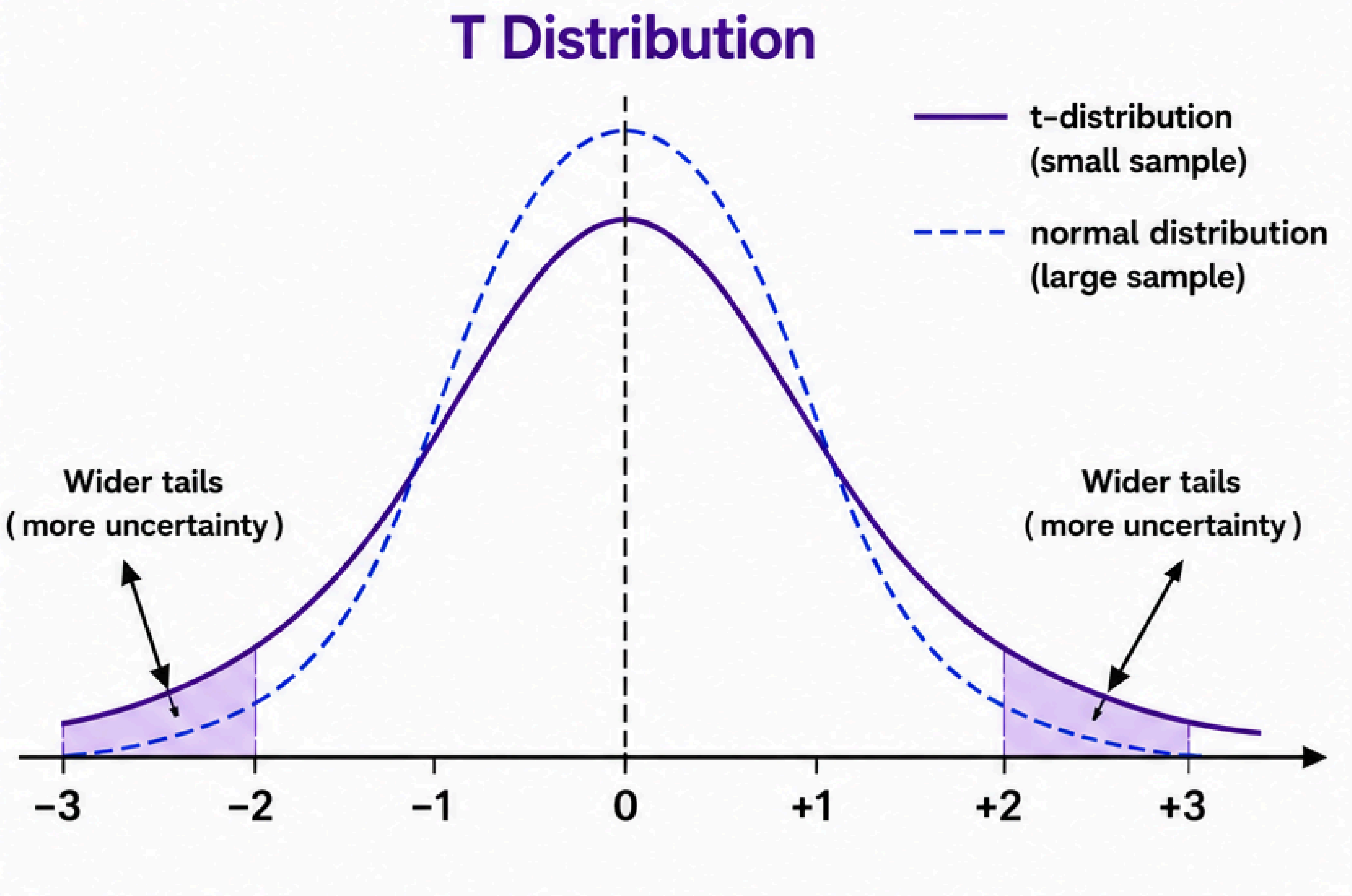
A bell shape
It is symmetric and centered like the normal distribution.

More spread in the tails
The tails are wider, meaning more extreme values are possible.

More uncertainty
Wider tails represent greater uncertainty in estimates.

As sample size increases, the t-distribution becomes almost identical to the normal distribution.

Visual Idea



Key Takeaway: The t-distribution is used when the sample size is small and the population standard deviation is unknown. It accounts for extra uncertainty by having wider tails.

Pankaj Chouhan
www.codeswithpankaj.com
www.cwpc.in

Formula for T-Score

T-SCORE (t-STATISTIC) FORMULA

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

The t-score (t-statistic) tells us how far the **sample mean** is from the **population mean**, in terms of **standard error**.

WHERE:

$\bar{x}$ = Sample mean

μ = Population mean (hypothesized value)

s = Sample standard deviation

n = Sample size

Note: The t-score is used in hypothesis testing when the population standard deviation is **unknown** and the sample size is **small** (typically $n < 30$).

Pankaj Chouhan

www.codeswithpankaj.com | www.cwpc.in

Example 1: Manual Calculation

Suppose:	
Sample Mean ($\bar{x}$)	= 75
Population Mean (μ)	= 70
Sample SD (s)	= 10
Sample Size (n)	= 25

Step 1: Calculate Standard Error

$$SE = \frac{s}{\sqrt{n}}$$

$$SE = \frac{10}{\sqrt{25}}$$

$$SE = \frac{10}{5}$$

$$SE = 2$$

Step 2: Calculate T-Score

$$t = \frac{\bar{x} - \mu}{SE}$$

$$t = \frac{75 - 70}{2}$$

$$t = \frac{5}{2}$$

$$t = 2.5$$

Result | T-Score = 2.5

The sample mean is 2.5 standard errors above the population mean.

Pankaj Chouhan

www.codeswithpankaj.com

www.cwpc.in

T-SCORE & DEGREES OF FREEDOM

Understanding Significance, df, and Their Role in Hypothesis Testing

DEGREES OF FREEDOM (df)

Degrees of freedom (df) determine the shape of the t-distribution.

$$df = n - 1$$

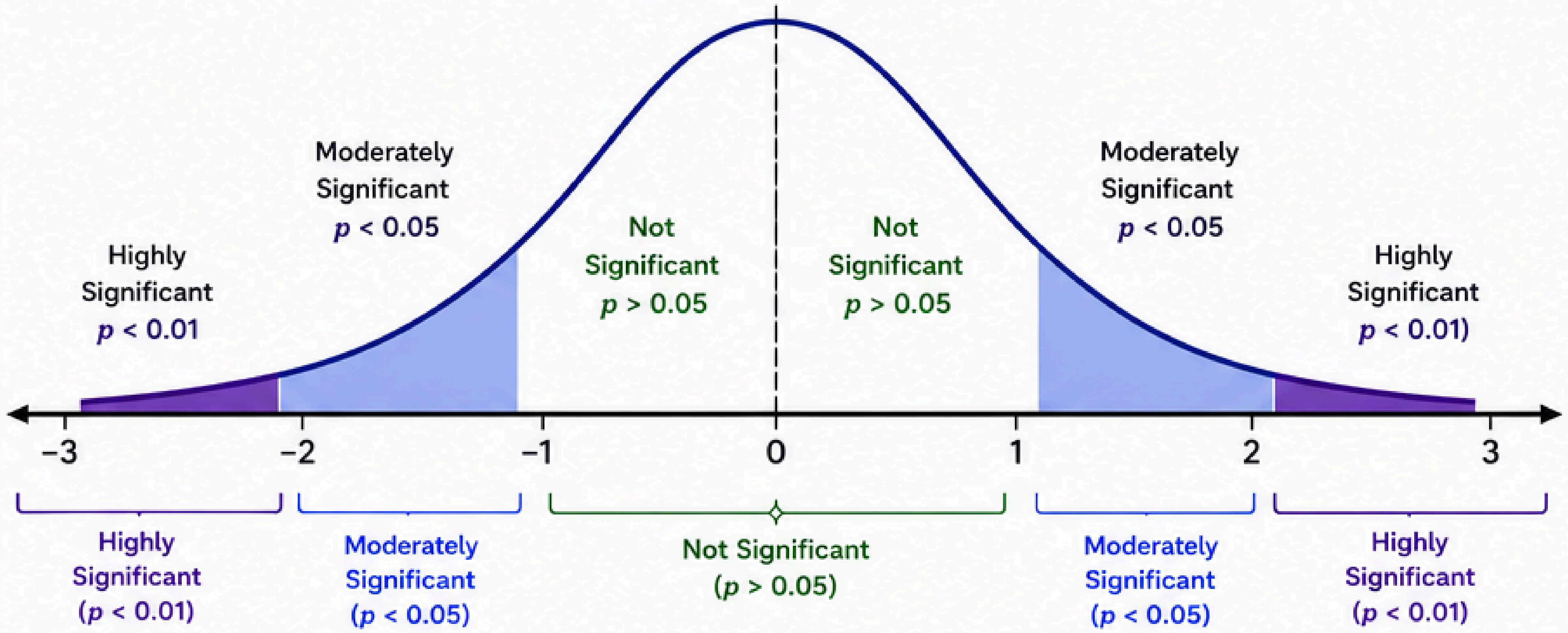
Where, n = Sample Size

EXAMPLE:

- Sample Size (n) = 25
- df = 25 – 1
- df = **24**

T-SCORE SIGNIFICANCE SCALE (T-DISTRIBUTION)

T-scores help determine how unusual a sample mean is under the null hypothesis.



EXAMPLE: CRITICAL T-VALUES (TWO-TAILED TEST)

df (n-1)	Confidence Level		
	90% (α = 0.10)	95% (α = 0.05)	99% (α = 0.01)
10	1.812	2.228	3.169
20	1.725	2.086	2.845
30	1.697	2.042	2.750
∞ (Z)	1.645	1.960	2.576

As df increases, critical t-value decreases and approaches the Z-critical value.

HOW T-SCORES ARE INTERPRETED?

The number of degrees of freedom (df) and the confidence level determine the critical t-value.

If |t-score| > critical t-value → Reject the Null Hypothesis (H_0)

If |t-score| ≤ critical t-value → Fail to Reject the Null Hypothesis (H_0)

KEY TAKEAWAY

T-scores, **degrees of freedom**, and confidence levels work together to decide whether a result is statistically significant.

Pankaj Chouhan

www.codeswithpankaj.com | www.cwpc.in

Real-Life Example

A company claims:

Average employee salary = ₹50,000

You collect data from:

25 employees

We will test if the true average salary is higher than ₹50,000.

Results:

Sample Mean ($\bar{x}$) = ₹55,000

Sample SD (s) = ₹8,000

Sample Size (n) = 25

1. Standard Error (SE)

$$SE = \frac{s}{\sqrt{n}}$$

$$SE = \frac{8000}{\sqrt{25}}$$

$$SE = \frac{8000}{5}$$

$$SE = 1600$$

2. T-Score (t)

$$t = \frac{\bar{x} - \mu}{SE}$$

$$t = \frac{55000 - 50000}{1600}$$

$$t = \frac{5000}{1600}$$

$$t = 3.125$$

Conclusion: The sample mean is significantly higher than the claimed average salary.

Pankaj Chouhan

www.codeswithpankaj.com | www.cwpc.in

Python Example

1. Calculating T-Score Manually

```python
import math

sample_mean = 75
population_mean = 70
sample_sd = 10
n = 25

t_score = (sample_mean - population_mean) / (sample_sd / math.sqrt(n))

print("T-Score:", round(t_score, 2))
```

Output

```
T-Score: 2.5
```

The sample mean is **2.5 standard errors** above the population mean.

2. T-Test in Python

```python
from scipy.stats import ttest_1samp

scores = [70, 75, 80, 72, 78, 76, 74, 77]

t_stat, p_value = ttest_1samp(scores, 70)

print("T-Statistic:", t_stat)
print("P-Value:", p_value)
```

A T-score is commonly used in a **T-Test**.

This helps determine whether the **sample mean differs significantly** from the population mean.

Z-Score vs T-Score Example

Suppose:

- Sample Mean ($\bar{x}$) = 75
- Population Mean (μ) = 70
- Standard Deviation (SD) (σ) = 10
- Sample Size (n) = 25

1. If Population SD Known (Use Z-Score)

Use:

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$Z = \frac{75 - 70}{10 / \sqrt{25}}$$

$$Z = \frac{5}{2}$$

Result:

$$Z = 2.5$$

2. If Population SD Unknown (Use T-Score)

Use:

$$T = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

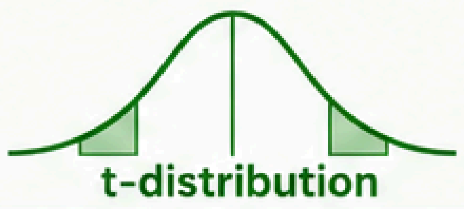
$$T = \frac{75 - 70}{10 / \sqrt{25}}$$

$$T = \frac{5}{2}$$

Result:

$$T = 2.5$$

The formula may look similar, but the interpretation uses the **t-distribution** instead of the **normal distribution**.



Pankaj Chouhan | www.codeswithpankaj.com | www.cwpc.in

T-SCORE: KEY TAKEAWAYS

Understand the Concept. Apply with Confidence.

KEY TAKEAWAYS

1. T-score is used when the population standard deviation is unknown.
2. It is especially useful for small sample sizes.
3. T-score uses the t-distribution.
4. The t-distribution has wider tails than the normal distribution.
5. Degrees of freedom = $n - 1$.
6. As sample size increases, the t-distribution approaches the normal distribution.
7. T-scores are the foundation of hypothesis testing and t-tests.

QUICK MEMORY TRICK

Z = Population Known

T = Population Unknown

OR SIMPLY

Use **Z** when you **know** the population standard deviation.

Use **T** when you **estimate** it from a sample.

Pankaj Chouhan | www.codeswithpankaj.com | www.cwpc.in