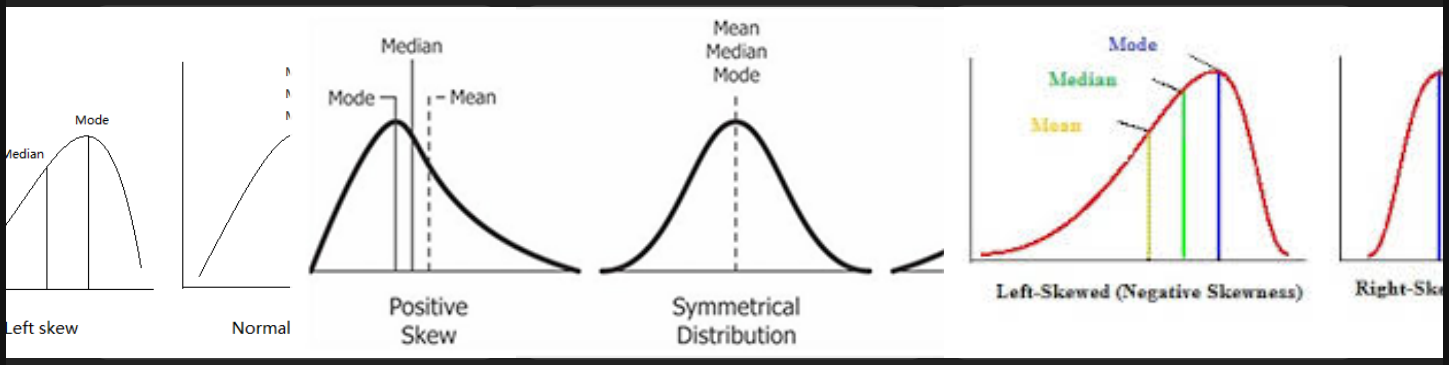
explain these topics one by one try to cover hidden aspects of them

Statistics Basics
Mean, median, mode
Standard deviation, variance
Correlation vs causation

Great — we will go **one by one**, with **deep intuition**, **hidden aspects**, **derivations**, **real-life meaning**, and **mathematical clarity**.

⭐ 1. Mean, Median, Mode



These three are **measures of central tendency** — ways to express what is "typical" in data. But each measures **different** types of "typical".

1.1 Mean (Average)

Formula

$$\text{Mean } \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Intuition

- Mean is the **balance point** of the data.
- Think of each data point as a weight on a number line → mean is the point where the line balances.

Hidden Aspect

- Mean minimizes **sum of squared errors**:

$$\mu = \arg\min_x \sum (x_i - x)^2$$

This is why the mean is used in:

- Linear regression (least squares)
- Gaussian distributions
- Optimization problems

Real-life Examples

1. **Average marks**, **average salary**, **average speed**.
2. In machine learning, mean is used to calculate **loss functions** for gradients.

Weakness

- Very sensitive to **outliers**
  Example: Salaries = {30k, 35k, 40k, 300k} → mean looks very high because of one outlier.

1.2 Median

Definition

Middle value when data is ordered.

Hidden Aspect

- Median minimizes **sum of absolute errors**:

$$\text{Median} = \arg\min_x \sum |x_i - x|$$

This makes the median **robust** against outliers.

Use Cases

- Income analysis
- House price analysis
- Any skewed distribution

Why more stable?

If you add a huge value (1,000,000) to a dataset, the median barely moves.

1.3 Mode

Definition

Most frequent value.

Hidden Aspects

- Mode is the only measure that makes sense for **categorical** data.
  Examples:
  - Most common blood group
  - Most bought product
- A distribution can have:
  - 1 mode (unimodal)
  - 2 modes (bimodal)
  - many modes (multimodal)

Use Cases

- Detecting clusters in data
- Fraud signals
- Customer segmentation

Comparing Mean, Median, Mode in Skewed Distributions

Right-skewed data (income, house prices)

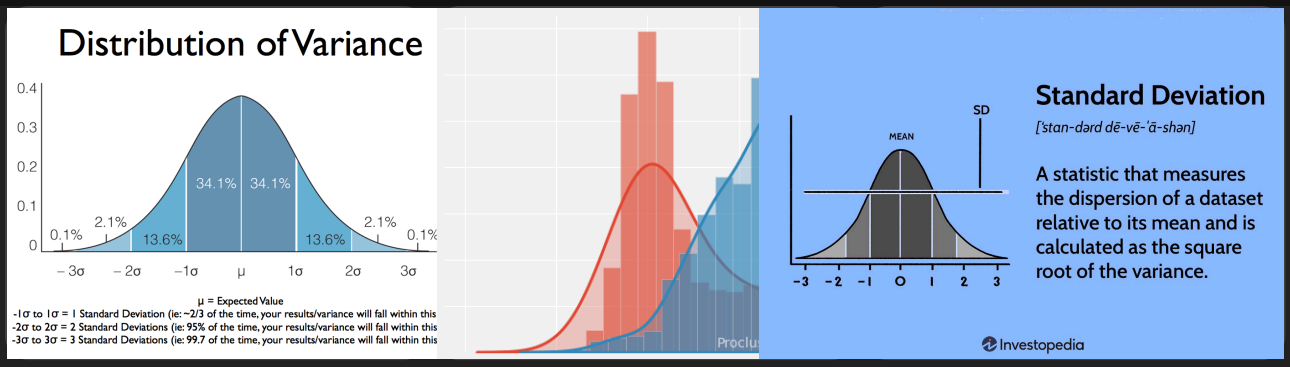
Mode < Median < Mean

Left-skewed data (life expectancy in ancient times)

Mean < Median < Mode

This ordering reveals distribution shape!

⭐ 2. Variance & Standard Deviation



Both measure **spread** — how far data points are from the center.

2.1 Variance

Formula

$$\sigma^2 = \frac{1}{n}\sum(x_i - \mu)^2$$

Meaning

Variance tells you **how much data points vary from the mean**.

Hidden Aspects

1. Squaring emphasizes **larger deviations** → large errors = very high variance.
2. Variance appears everywhere in:
   - Gaussian distribution
   - Maximum likelihood estimation
   - Covariance matrices
   - Portfolio risk calculation in finance
3. Units are **squared units**, e.g., meters^2, dollars^2
   → This is why it's not intuitive.

2.2 Standard Deviation (SD)

$$\sigma = \sqrt{\sigma^2}$$

Meaning

Standard deviation brings variance back to **original units**.

Intuition

SD measures the **typical distance from the mean**.

Empirical Rule (68–95–99 Rule)

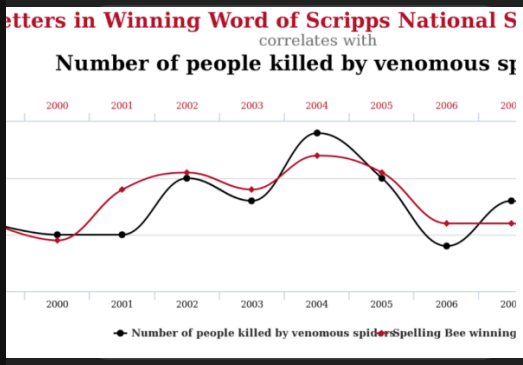
For normal distributions:

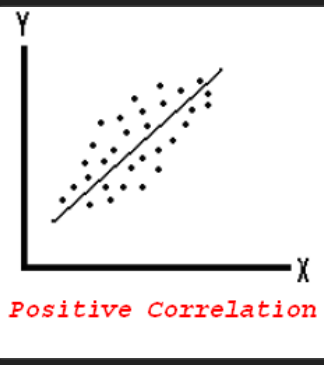
- 68% data within 1 SD
- 95% within 2 SD
- 99.7% within 3 SD

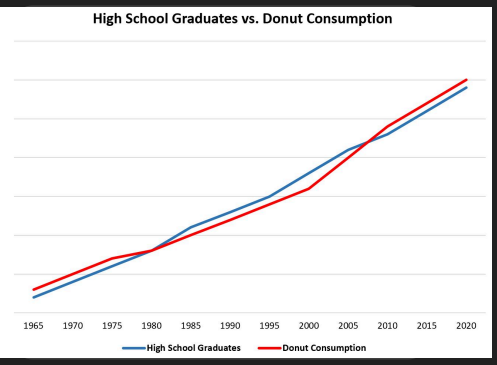
Real-Life Uses

- Stock volatility (higher SD → more risky)
- Quality control in manufacturing
- Signal processing noise analysis
- ML feature normalization

⭐ 3. Correlation vs Causation







3.1 What is Correlation?

Correlation measures **strength & direction** of linear relationship.

Pearson correlation coefficient

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

- r = +1 → perfect positive line
- r = −1 → perfect negative line
- r = 0 → no linear relationship

Hidden Aspects

1. **Correlation does NOT detect nonlinear relationships**.
   If $Y = X^2$, correlation may be 0 even though perfectly related.
2. **Outliers can destroy correlation**.
3. **Correlation does not mean causation**

- Ice cream sales ↑ & drowning deaths ↑ → correlated because of summer, not each other.

3.2 What is Causation?

X causes Y if:

- Manipulating X changes Y
- There is no other confounding variable
- Relationship is stable & consistent

Methods to Prove Causation

- Controlled experiments (A/B testing)
- Randomization
- Longitudinal studies
- Instrumental variables
- Granger causality (time series)

Spurious Correlation Examples

- People who eat more chocolate win more Nobel Prizes
- Shoe size correlates with reading ability in children (age is hidden variable)

✔ Summary Table

Concept	What it Measures	Hidden Insight	Weakness
Mean	Balance point	Minimizes squared error	Sensitive to outliers
Median	Middle value	Minimizes absolute error	Harder to use algebraically
Mode	Most frequent	Works for categories	Not stable

Concept	What it Measures	Hidden Insight	Weakness
Variance	Spread of squared distances	Captures energy, appears in ML & finance	Squared units
Std. Deviation	Average deviation	Interpretable	Affected by outliers
Correlation	Linear relationship	Not causation	Misses nonlinear
Causation	Cause-effect	Needs experiments	Hard to prove

If you'd like, I can continue with:

✅ derive the formulas (variance minimization, correlation formula origin)
✅ show 20 practice problems for intuition
✅ show hidden symbols in statistics (μ, σ^2, ρ, r, cov, Σ, etc.)

Which part should we go deeper into next?