**Probability Set Theory**

Today I am going to document my learnings on sets, Conditional Probability, Union, Intersection, Bernoulli’s Distribution, Normal Distribution, Binomial Distribution, Variance, Standard Deviation, Mean, Median, Mode, Expected Value. Taking Walmart customer behavior as a use case.

**Business Scenario**

Walmart wants to increase sales and customer satisfaction by understanding **who their customers are, how they shop, and how different departments interact** (for example, electronics and groceries).

We as an analytics team gathers **six months of transaction data** across departments, including customer purchases, online delivery performance, and responses to marketing campaigns

**Set** -A well-defined collection of distinct, unique elements, where you can clearly tell whether an item belongs or not.

**Non-well-defined examples**

Set of “all beautiful paintings” → What counts as beautiful? It is subjective, so it is not well-defined.

Set of “nice customers” → Without a clear rule, it is subjective.

**Why does this Well-defined collection of elements matter in datasets**?

When we work with data, we want clear, defined groups:

All unique customer IDs → well-defined set.

All orders placed in June → well-defined set.

This clarity is essential for clean data analysis, filtering, and modeling.

Suppose we Consider

**Set A:** All customers who bought electronics. → well-defined set- A= {Sneha, Tanuja, Hruthik, Prudhvi}

**Set B:** All customers who bought groceries. → well-defined set- B= {Sneha, Tanuja, Lasya, Navya, Mahesh babu}

**Subset:** Customers who only bought electronics, or only groceries, or both.

**Why set and subset matters?**

We as a team can segment Walmart ‘s **customer base**:

Electronics-only shoppers (**Set A**) → may want promotions on electronics accessories.

Grocery-only shoppers (**Set B**) → may be influenced by food deals or health-related offers.

Both **(A ∩ B**) → {Sneha, Tanuja}- valuable cross-category shoppers, strong loyalty potential high value customers

**Union:** Combines two sets, keeping all unique items. - {Sneha, Tanuja, Hruthik, Lasya, Prudhvi, Mahesh Babu, Navya}

**Union (A ∪ B):** All customers who bought either electronics or groceries (no duplicates).

**Intersection:** Keeps only items common to both sets.

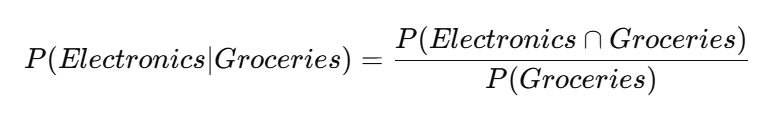
**Intersection (A ∩ B):** Only the customers who bought both.

**Conditional Probability**

The probability that one event occurs given that another event has already occurred.

Walmart Example

**What is the probability that a customer buys electronics given they buy groceries?**

****

**Why it matters**

If data shows grocery shoppers are 50% more likely to also explore electronics, Walmart can:

Place electronics promotions on grocery receipts.

Push cross-category recommendations in the Walmart app.

This increases basket size and cross-department sales.

**Bernoulli Trials → *Models repeated binary experiments***

**Bernoulli trials let you simulate or count how often the success shows up.**

**Definition:**  
A **single trial** or experiment that results in:

* Success (1)
* Failure (0)
* Observe repeated independent events (like many customers making decisions).
* Want to **count** or **model the probability of success/failure** over multiple trials.

**Example (Walmart):**

* Did this individual customer buy groceries? Yes (1) or no (0).
* Run this check over thousands of customers.

Modeling **the process** of repeated, independent binary outcomes.

**Normal Distribution**

The normal distribution models continuous variables that follow a **bell-shaped** curve, centered around a mean (average), with most values near the middle and fewer at the extremes.

**Walmart Example: Normal Distribution on Spending Amounts**

Let us shift focus from **binary outcomes** (yes/no) to a **continuous outcome**.

**Example:**

* We track the **amount of money customers spend on groceries** when they also buy electronics.

From past data, we find:

* Mean grocery spends: $50.
* Standard deviation: $10.

We discover this spending follows a **normal distribution**.

**What We Can Do with This**

**Predict likelihoods:**

* What is the probability a customer spends between $40 and $60 on groceries?

Use the area under the normal curve between 40 and 60.

**Flag unusual behavior:**

* Spending below $30 or above $70 might be rare (outliers).

These customers could be targeted for special promotions.

**realistic business expectations:**

* Know that ~68% of customers will spend within ±1 standard deviation (i.e., $40–$60).
* Plan inventory and promotions accordingly.
* **Conditional probability** tells you: *Given they buy electronics, what is the chance they also buy groceries?*
* **Bernoulli trials** model: *For many electronics customers, who buys groceries or not?*
* **Normal distribution** models: *For those who buy, how much do they spend (as a continuous spread)?*

**Now I am changing my Business scenario to Retail domain (Amazon) for Variance and Standard deviation**

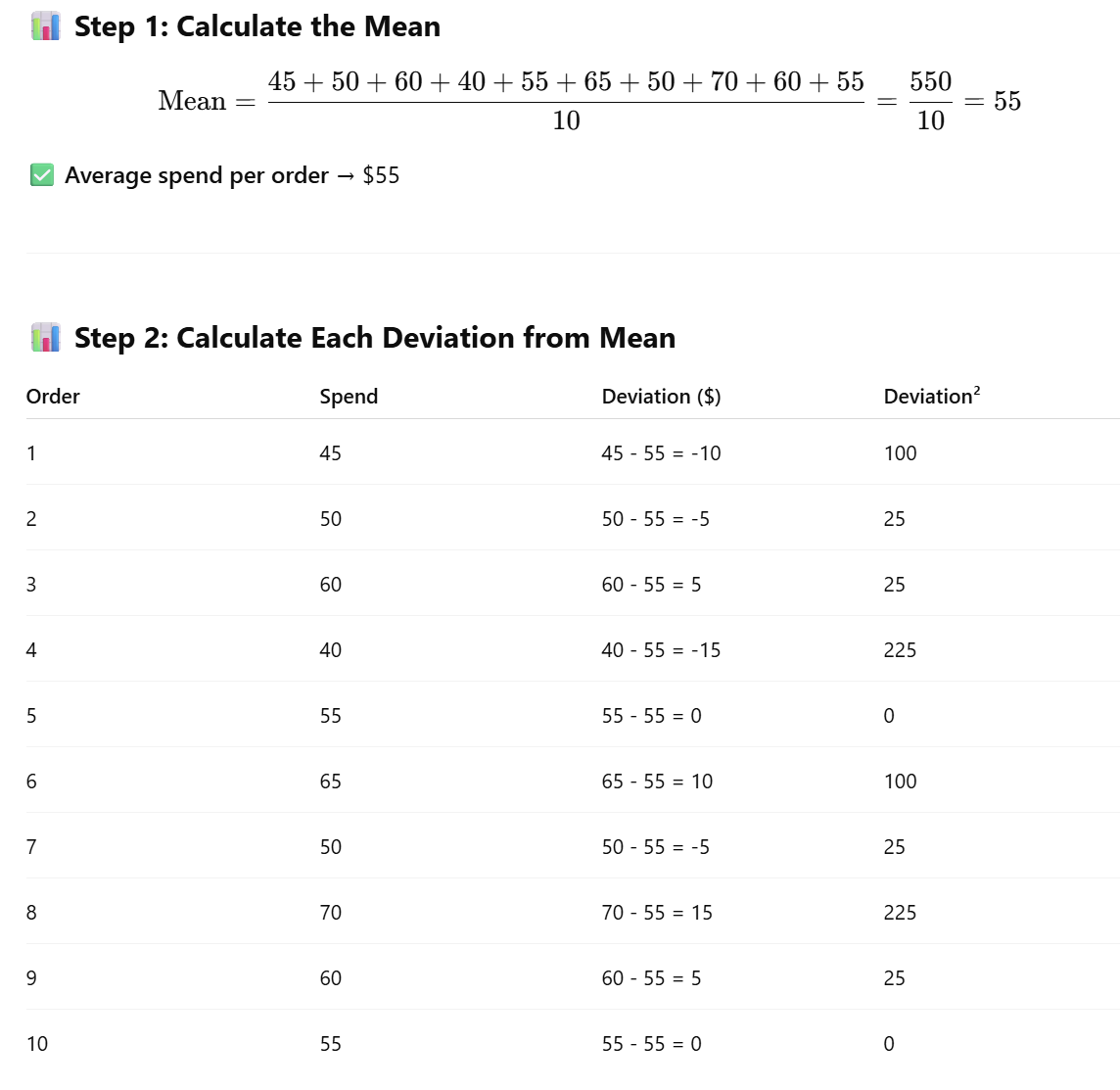
**Variance**

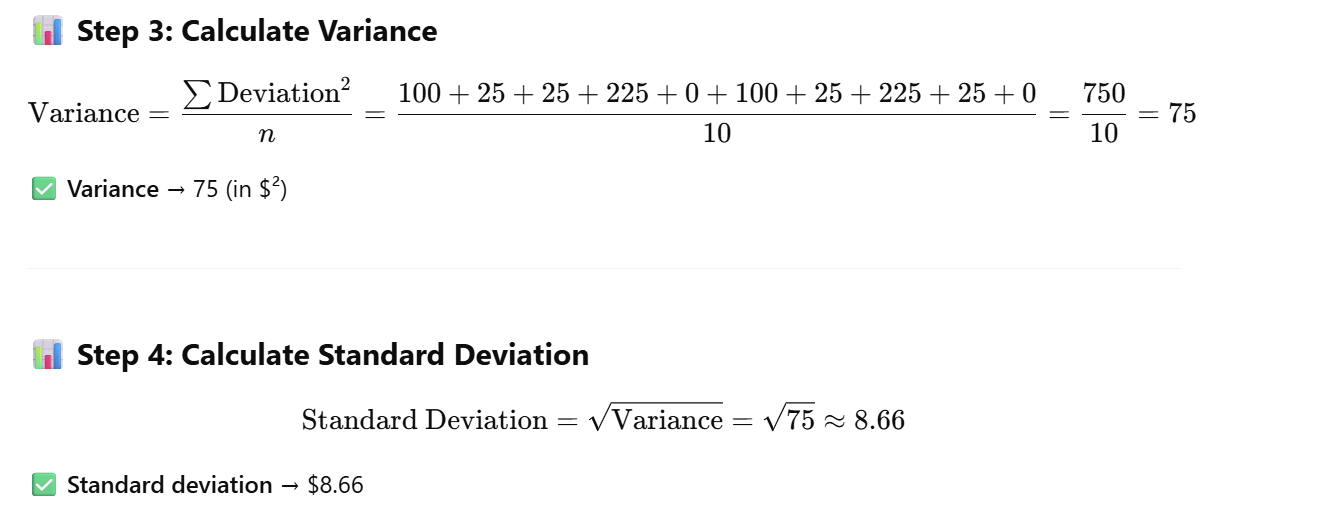
We want to measure how spread-out customer data is not just averages, but how much individual data points deviate from the average.

Let us pick **customer spending per order**.

Imagine we analyze **10 customer orders** on Amazon:

| **Order** | **Spend ($)** |
| --- | --- |
| 1 | 45 |
| 2 | 50 |
| 3 | 60 |
| 4 | 40 |
| 5 | 55 |
| 6 | 65 |
| 7 | 50 |
| 8 | 70 |
| 9 | 60 |
| 10 | 55 |





* On average, customers spend $55 per order.
* But the spread (standard deviation) is about $8.66 →  
  Most orders fall between $55 ± $8.66 → between $46.34 and $63.66.

**Business Use Cases:**

* Predict typical order ranges.
* Spot unusually high or low spenders (possible fraud, big buyers).
* Compare spending patterns between product categories or regions

| **Metric** | **Meaning** |
| --- | --- |
| Mean (average) | Central value → average spend per order ($55) |
| Variance | Average squared deviation → overall spread ($²) |
| Standard deviation | Typical spread in original units ($8.66) → easy to interpret |

**What is Expected Value (EV)?**

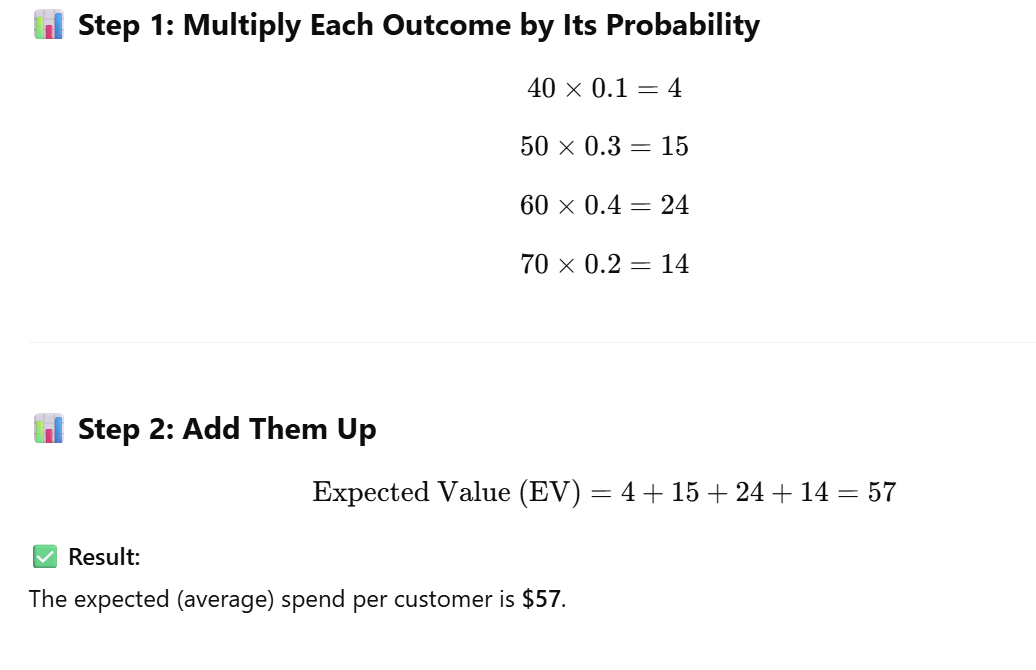
**Definition:**  
The expected value is the **average or mean outcome you would expect over many trials or observations**, weighted by their probabilities.

It answers:

*If you repeated this process many times, what is the average result you would expect?*

*Imagine Amazon has the following* ***probabilities*** *for customer spending on electronics orders:*

| **Spend per Order ($)** | **Probability** |
| --- | --- |
| 40 | 0.1 (10%) |
| 50 | 0.3 (30%) |
| 60 | 0.4 (40%) |
| 70 | 0.2 (20%) |
|  |  |



**What Does This Mean for Amazon?**

* Even though individual customers spend $40, $50, $60, or $70,  
  if Amazon looks across thousands of customers, it should expect an **average** revenue of **$57 per customer**.

This helps Amazon:

* Forecast future revenue.
* Set benchmarks and targets.
* Compare against actual performance to detect anomalies