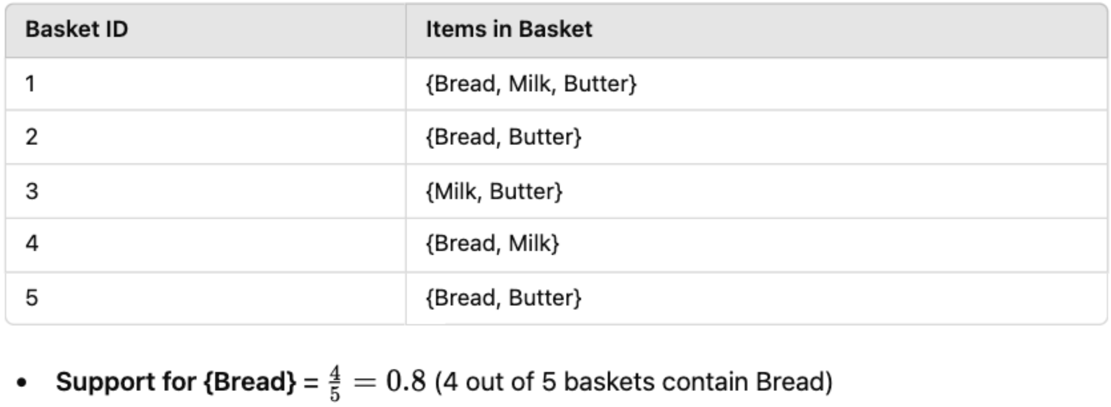
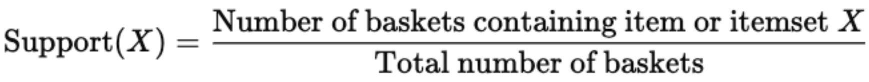
- Define frequent itemsets
- Support for item X: the number of baskets containing all items in X. Given as a percentage.



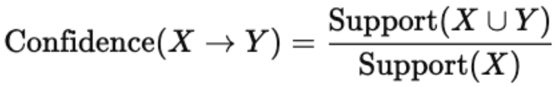
Basket ID	Items in Basket
1	{Bread, Milk, Butter}
2	{Bread, Butter}
3	{Milk, Butter}
4	{Bread, Milk}
5	{Bread, Butter}

- **Support for {Bread}** = $\frac{4}{5} = 0.8$ (4 out of 5 baskets contain Bread)

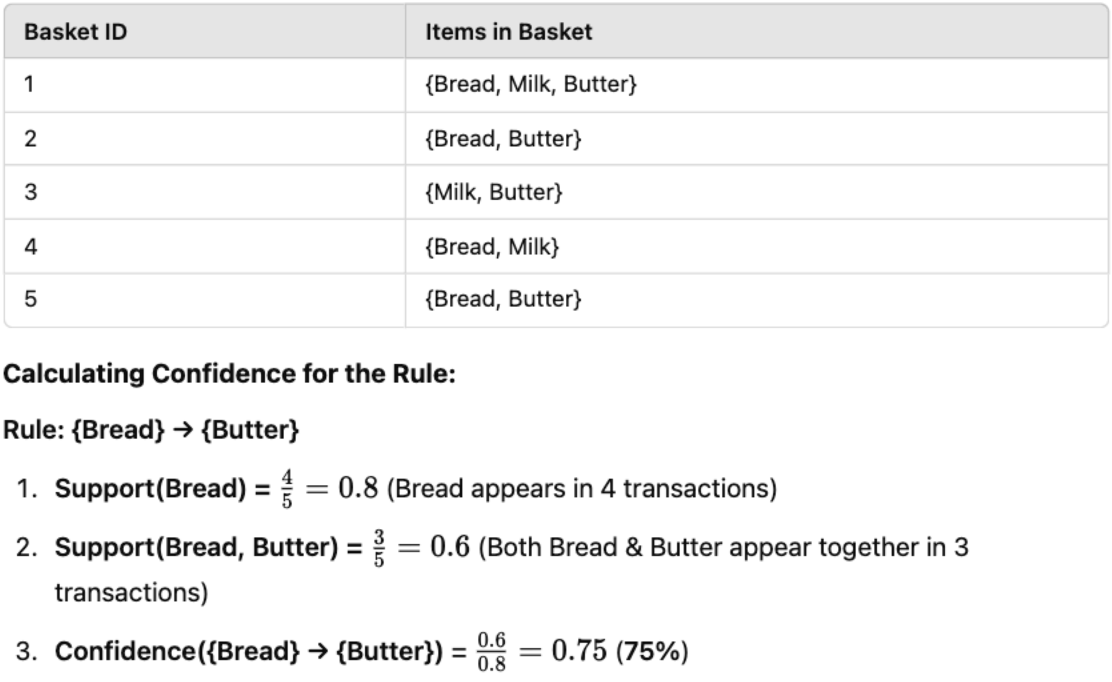


$$\text{Support}(X) = \frac{\text{Number of baskets containing item or itemset } X}{\text{Total number of baskets}}$$

- Confidence: the ratio of support for I I{j} with support for I
  - **Confidence** is used to measure the reliability of a rule in predicting the occurrence of an item based on another.



$$\text{Confidence}(X \to Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$$



Basket ID	Items in Basket
1	{Bread, Milk, Butter}
2	{Bread, Butter}
3	{Milk, Butter}
4	{Bread, Milk}
5	{Bread, Butter}

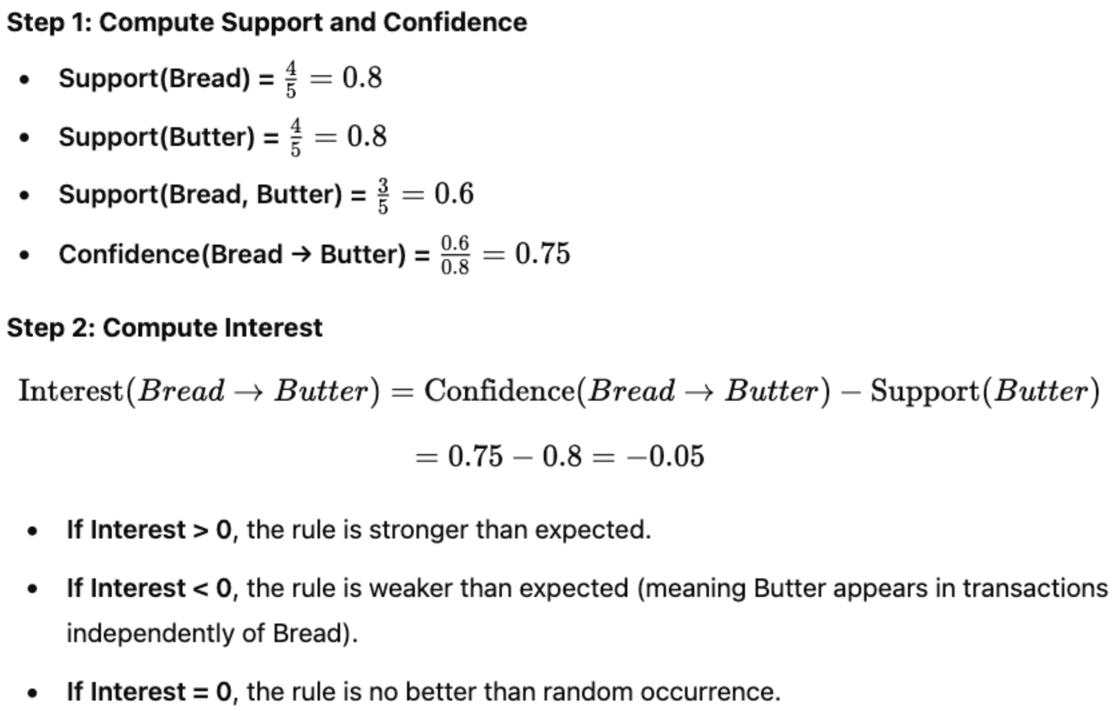
Calculating Confidence for the Rule:

Rule: {Bread} → {Butter}

1. **Support(Bread)** = $\frac{4}{5} = 0.8$ (Bread appears in 4 transactions)
2. **Support(Bread, Butter)** = $\frac{3}{5} = 0.6$ (Both Bread & Butter appear together in 3 transactions)
3. **Confidence({Bread} → {Butter})** = $\frac{0.6}{0.8} = 0.75$ **(75%)**

- Interest
    - evaluates how much the confidence of an association rule **X → Y** deviates from the expected probability of **Y** occurring independently.
        - In other words, it measures how likely when you buy two items together than buy one independently.

$$\text{Interest}(X \to Y) = \text{Confidence}(X \to Y) - \text{Support}(Y)$$



Step 1: Compute Support and Confidence

- **Support(Bread) =** $\frac{4}{5} = 0.8$
- **Support(Butter) =** $\frac{4}{5} = 0.8$
- **Support(Bread, Butter) =** $\frac{3}{5} = 0.6$
- **Confidence(Bread → Butter) =** $\frac{0.6}{0.8} = 0.75$

Step 2: Compute Interest

$$\text{Interest}(Bread \to Butter) = \text{Confidence}(Bread \to Butter) - \text{Support}(Butter)$$

$$= 0.75 - 0.8 = -0.05$$

- **If Interest > 0**, the rule is stronger than expected.
- **If Interest < 0**, the rule is weaker than expected (meaning Butter appears in transactions independently of Bread).
- **If Interest = 0**, the rule is no better than random occurrence.

- Counting Pairs
    - 10^5 items
    - Number of pairs of items = 10^5(10^5-1)/2=5*10^9
    - Triangular matrix approach
        - to find pair {I,j} position
            - (I-1)*(n-i/2)+(j-i)
        - Total pair
            - n(n-1)/2
        - Total bytes = 2n(n-1)   known 4bytes per pair
- A-priori algorithm
    - A two pass approach called a-priori limits the need for main memory
    - Monotonicity
        - If a set of items I appears s times then so does every subset j of I

    - Pass1: read baskets and count in main memory the occurrences of each single item
    - Pass2: read baskets again and count in main memory only those pairs of items where both were found in pass 1 to be frequent
- PCY Algorithm
    - Generate all possible pairs for each basket
    - Hashes them to buckets
    - Keeps a count for each hash bucket
    - Identifies frequent Buckets where count>=s
- Random Sampling
    - Read a sample that represent entire data set
- Savasere Omiecinki and Navathe (Son algorithm)
    - Pass one: In-memory, read all small subsets and let itemset become candidate
    - Pass two: count all candidate itemsets and determine which are frequent in the entire set
    - Map Reduce
        - Phase 1: find local candidate
        - Phase 2: find true frequent itemsets
- Toivonen's Algorithm
    - Negative border
    - First path: find negative boarder