Simpson's Paradox

- Observed relationship in data may be influenced by the presence of other confounding factors (hidden variables)
  - Hidden variables may cause the observed relationship to disappear or reverse its direction!

- Proper measures are required to avoid generating spurious patterns

Simpson's Paradox

- Recovery rate from Covid
  - Hospital A: 80%
  - Hospital B: 90%
- Which hospital is better?

Simpson's Paradox

- Recovery rate from Covid
  - Hospital A: 80%
  - Hospital B: 90%
- Which hospital is better?
- Covid recovery rate on older population
  - Hospital A: 50%
  - Hospital B: 30%
- Covid recovery rate on younger population
  - Hospital A: 99%
  - Hospital B: 98%

Simpson's Paradox

Buy HDTV	Buy Exercise Machine		
	Yes	No	
Yes	99	81	180
No	54	66	120
	153	147	300

Simpson's Paradox

- Together, these rules suggest that customers who buy high-definition televisions are more likely to buy exercise machines than those who do not buy high-definition televisions

Simpson's Paradox-Continued

Customer Group	Buy HDTV	Buy Exercise Machine		Total
		Yes	No	
College Students	Yes	1	9	10
	No	4	30	34
Working Adult	Yes	98	72	170
	No	50	36	86

Simpson's Paradox

- The rules suggest that, for each group customers who do not buy high-definition televisions are more likely to buy exercise machines, which contradicts the previous conclusion when data from the two customer groups are pooled together.
- The reversal in the direction of association is known as Simpson's paradox

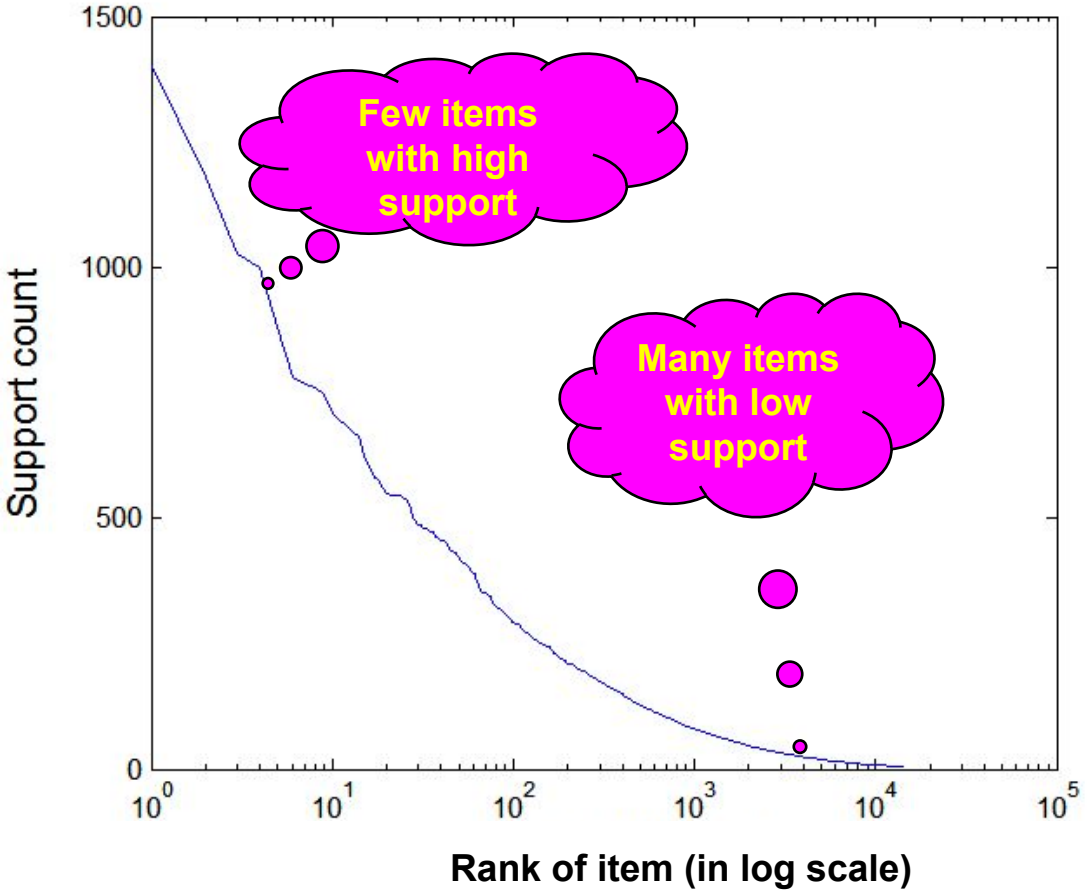
Simpson's Paradox

- Notice that most customers who buy HDTVs are working adults. Working adults are also the largest group of customers who buy exercise machines. Because nearly 85% of the customers are working adults, the observed relationship between HDTV and exercise machine turns out to be stronger in the combined data than what it would have been if the data is stratified.
- The reversal in the direction of association is known as Simpson's paradox

Effect of Support Distribution on Association Mining

- Many real data sets have skewed support distribution

Support distribution of a retail data set



Example

- Many real data sets have skewed support distribution

Group	G_1	G_2	G_3
Support	$< 1\%$	$1\% - 90\%$	$> 90\%$
Number of Items	1735	358	20

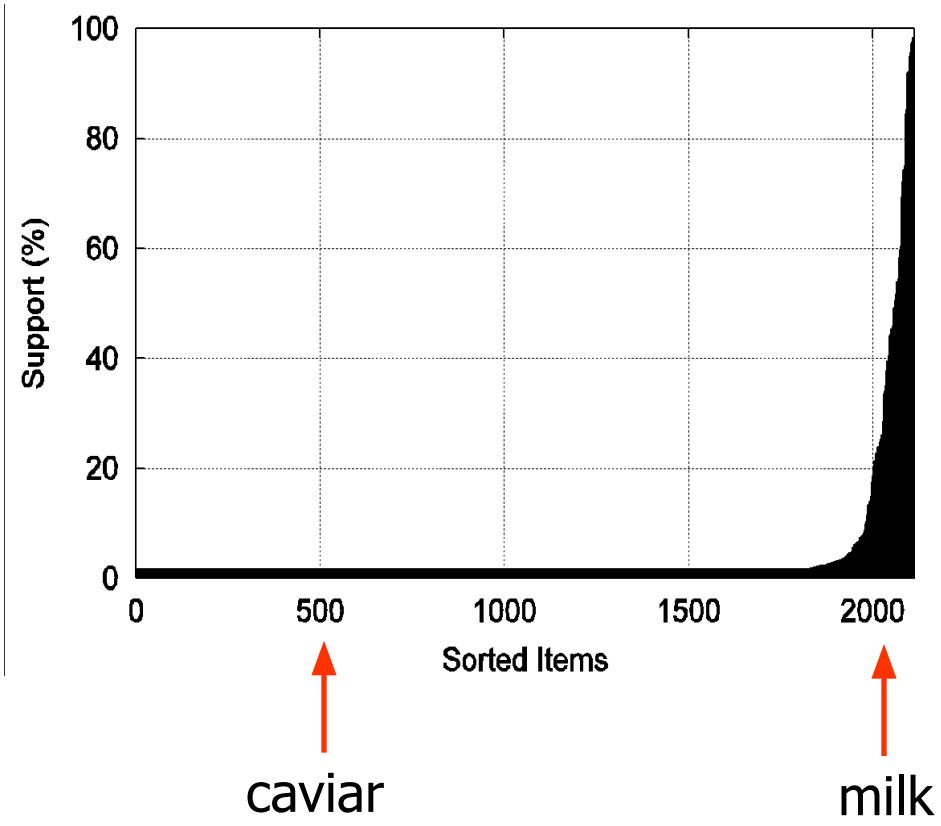
Effect of Support Distribution

- Difficult to set the appropriate *minsup* threshold
  - If *minsup* is too high, we could miss itemsets involving interesting rare items (e.g., {caviar, vodka})
  - If *minsup* is too low, it is computationally expensive and the number of itemsets is very large

Cross-Support Patterns

- First, the computational and memory requirements of existing association analysis algorithms increase considerably with low support thresholds.
- Second, the number of extracted patterns also increases substantially with low support thresholds.
- Third, we may extract many spurious patterns that relate a high-frequency item such as milk to a low-frequency item such as caviar.
- Such patterns, which are called cross-support patterns, are likely to be spurious because their correlations tend to be weak

Cross-Support Patterns



A cross-support pattern involves items with varying degree of support

- Example: {caviar,milk}

A Measure of Cross Support

- Given an itemset, $X = \{x_1, x_2, \ldots, x_d\}$, with d items, we can define a measure of cross support, r, for the itemset

$$r(X) = \frac{\mathbf{min}\{s(x_1), s(x_2), \ldots, s(x_d)\}}{\mathbf{max}\{s(x_1), s(x_2), \ldots, s(x_d)\}}$$

where $s(x_i)$ is the support of item x_i

  - Can use $r(X)$ to prune cross support patterns

Example

Example 6.4. Suppose the support for milk is 70%, while the support for sugar is 10% and caviar is 0.04%. Given $h_c = 0.01$, the frequent itemset {milk, sugar, caviar} is a cross-support pattern because its support ratio is

$$r = \frac{\min\left[0.7, 0.1, 0.0004\right]}{\max\left[0.7, 0.1, 0.0004\right]} = \frac{0.0004}{0.7} = 0.00058 < 0.01.$$