Evaluation of Association Patterns

Evaluation of Association Patterns

- Association analysis algorithms have the potential to generate a large number of patterns.
- In real commercial databases we could easily end up with thousands or even millions of patterns, many of which might not be interesting.
- Very important to establish a set of well-accepted criteria for evaluating the quality of association patterns.
- Measures of Interestingness.

Measures of Interestingness

Objective Interestingness Measures

- Used to rank patterns
- Some provide statistical information.
- Eg: Support, Confidence, Lift, Correlation

Subjective measures of Interestingness

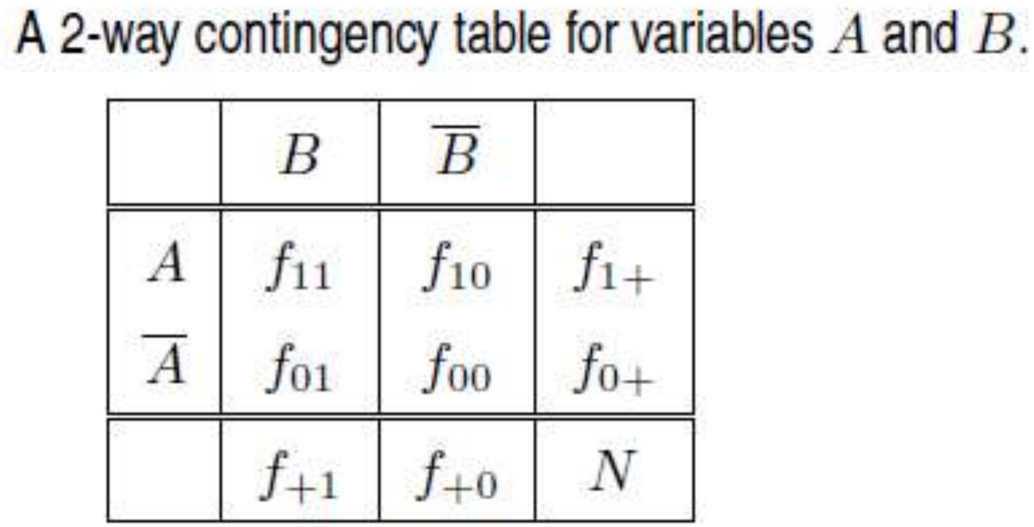
- Pattern is interesting if it is unexpected.

  For eg: {bread}-> {butter} may not be interesting even with high support and confidence values, because it is rather obvious.

  But, {Diapers}-> {Beer} is interesting because it is unexpected and suggest a new cross selling opportunity for retailers.
- Incorporating subjective knowledge for evaluation is a difficult task because it requires a considerable amount of prior information from domain experts.

Objective Measures of Interestingness

- An objective measure is a data-driven approach for evaluating the quality of association patterns.
- It is domain-independent and requires only that the user specifies a threshold for filtering low-quality patterns.
- An objective measure is usually computed based on the frequency counts tabulated in a *contingency table.*



A 2-way contingency table for variables A and B.

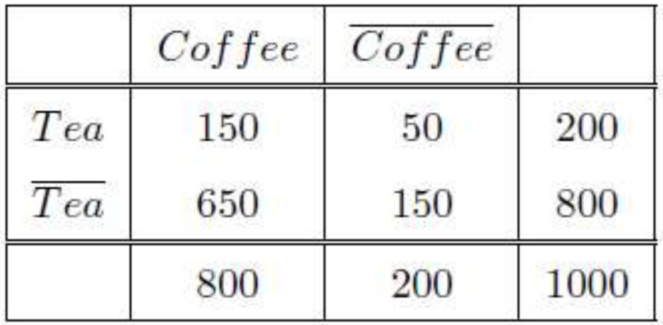
	B	$\overline{B}$	
A	f_{11}	f_{10}	f_{1+}
$\overline{A}$	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	N

Limitations of the Support-Confidence Framework

- The classical association rule mining formulation relies on the ***support*** and ***confidence*** measures to eliminate uninteresting patterns.
- The drawback of support is that many potentially interesting patterns involving low support items might be eliminated by the support threshold.
- The drawback of confidence is that despite having high confidence value, the rule could be misleading.

- Consider the association rule *{Tea} –→{Coffee}*
- *Support = f(T ^C)/N = 150/1000=15%*
- *Confidence= f(T ^C)/f(T) = 150/200 = 75%*

. Beverage preferences among a group of 1000 people.

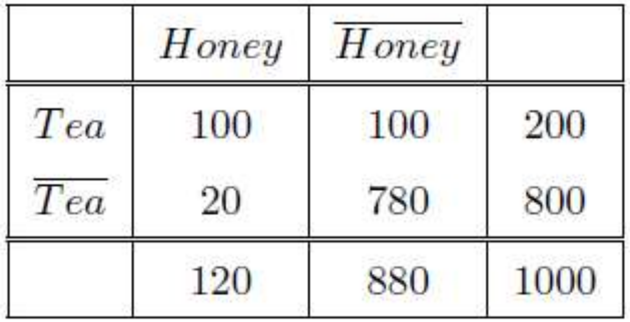


	$Coffee$	$\overline{Coffee}$	
Tea	150	50	200
$\overline{Tea}$	650	150	800
	800	200	1000

- *Fraction of people who drink coffee = 800/1000= 80%*
- *Therefore, knowing that a person is a tea drinker reduces the possibility of being a coffee drinker from 80% to 75%.*
- *This misleading.*
- *Evaluate {Tea}->{Honey}*
- *Confidence as a measure*

can falsely accept or reject a rule.

Information about people who drink tea and people who use honey in their beverage.



	$Honey$	$\overline{Honey}$	
Tea	100	100	200
$\overline{Tea}$	20	780	800
	120	880	1000

- The support $s(A,B)$ of a pair of a variables A and B measures the probability of the two variables occurring together.
- Hence, the joint probability $P(A,B)$ can be written as

$$P(A,B) = s(A,B) = \frac{f_{11}}{N}.$$

- If we assume A and B are statistically independent, i.e. there is no relationship between the occurrences of A and B, then $P(A,B) = P(A) \times P(B)$.
- Hence, under the assumption of statistical independence between A and B, the support $s_{\text{indep}}(A,B)$ of A and B can be written as

$$s_{\text{indep}}(A,B) = s(A) \times s(B) \quad \text{or equivalently,} \quad s_{\text{indep}}(A,B) = \frac{f_{1+}}{N} \times \frac{f_{+1}}{N}.$$

- If the support between two variables, $s(A,B)$ is equal to $s_{\text{indep}}(A,B)$, then A and B can be considered to be unrelated to each other.
- If $s(A,B)$ is widely different from $s_{\text{indep}}(A,B)$, then A and B are most likely dependent.
- Hence, any deviation of $s(A,B)$ from $s(A)\times s(B)$ can be seen as an indication of a statistical relationship between A and B.

Interest Factor

- The interest factor is also called as the "lift".

$$I(A,B) = \frac{s(A,B)}{s(A) \times s(B)} = \frac{N f_{11}}{f_{1+} f_{+1}}.$$

- $s(A) \times s(B) = s_{\text{indep}}(A,B)$.
- Hence, the interest factor measures the ratio of the support of a pattern $s(A,B)$ against its baseline support $s_{\text{indep}}(A,B)$ computed under the statistical independence assumption.
- we can interpret the measure as follows:

$$I(A,B) \begin{cases} = 1, & \text{if } A \text{ and } B \text{ are independent;} \\ > 1, & \text{if } A \text{ and } B \text{ are positively related;} \\ < 1, & \text{if } A \text{ and } B \text{ are negatively related.} \end{cases}$$

- For the tea-coffee example, $I = \frac{0.15}{0.2 \times 0.8} = 0.9375$ -> slightly negative relationship
- For the tea-honey example, $I = \frac{0.1}{0.12 \times 0.2} = 4.1667$; -> strong positive relationship

- The interest factor has a number of statistical advantages over the confidence measure that make it a suitable measure for analyzing statistical independence between variables.

Piatesky-Shapiro (PS) Measure

- Instead of computing the ratio between $s(A,B)$ and $s_{\text{indep}}(A,B) = s(A) \times s(B)$, the *PS* measure considers the difference between $s(A,B)$ and $s(A) \times s(B)$ as follows.

$$PS = s(A,B) - s(A) \times s(B) = \frac{f_{11}}{N} - \frac{f_{1+}f_{+1}}{N^2}$$

PS value

0 when *A* and *B* are mutually independent of each other.

>0 when there is a positive relationship between the two variables

< 0 when there is a negative relationship.

Correlation Analysis

- One of the most popular techniques for analyzing relationships between a pair of variables.
- For continuous variables, correlation is defined using Pearson's correlation coefficient.
- For binary variables, correlation can be measured using the *φ*-coefficient, which is defined as

$$\phi = \frac{f_{11}f_{00} - f_{01}f_{10}}{\sqrt{f_{1+}f_{+1}f_{0+}f_{+0}}}.$$

- The *φ*-coefficient can be rewritten in terms of the support measures of *A*, *B*, and *{A,B}* as follows:

$$\phi = \frac{s(A,B) - s(A) \times s(B)}{\sqrt{s(A) \times (1 - s(A)) \times s(B) \times (1 - s(B))}}$$

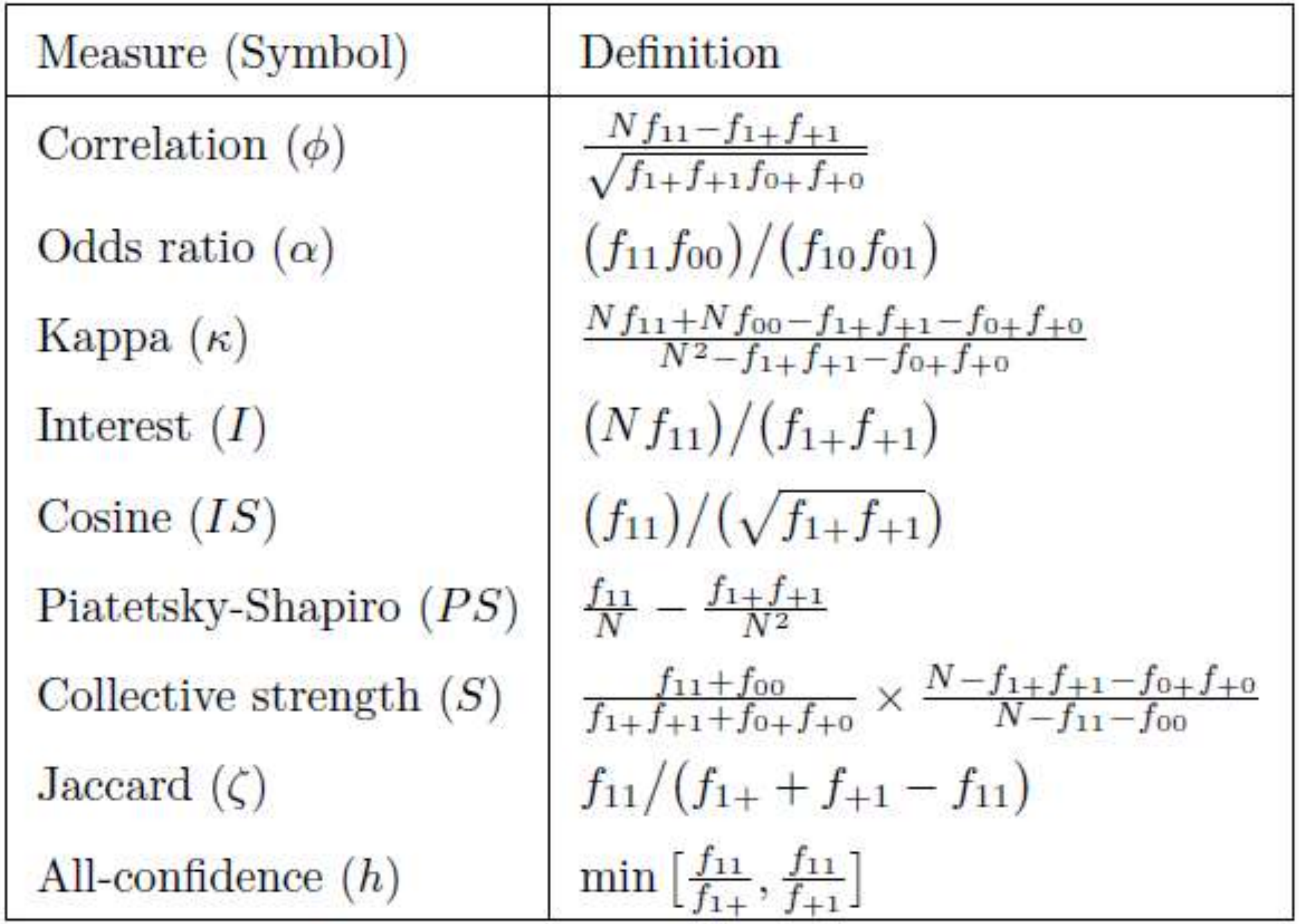
Correlation Analysis

- The numerator in the above equation is identical to the *PS* measure.
- Hence, the φ-coefficient can be understood as a normalized version of the *PS* measure, where that the value of the φ-coefficient ranges from −1 to +1.
- From a statistical viewpoint, the correlation captures the normalized difference between $s(A,B)$ and $s_{\text{indep}}(A,B)$.
- A correlation value of 0 means no relationship.
- A value of +1 suggests a perfect positive relationship.
- A value of −1 suggests a perfect negative relationship.

IS Measure

- An alternative measure for capturing the relationship between *s*(*A*,*B*) and *s*(*A*) × *s*(*B*).
- The IS measure is defined as follows: $IS(A,B) = \sqrt{I(A,B) \times s(A,B)} = \frac{s(A,B)}{\sqrt{s(A)s(B)}} = \frac{f_{11}}{\sqrt{f_{1+}f_{+1}}}$
- *IS* is the geometric mean between the interest factor and the support of a pattern, *IS* is large when both the interest factor and support are large.
- If the interest factor of two patterns are identical, the *IS* has a preference of selecting the pattern with higher support.
- *IS* is mathematically equivalent to the cosine measure for binary .
- The value of *IS* thus varies from 0 to 1, where an *IS* value of 0 corresponds to no co-occurrence of the two variables, while an *IS* value of 1 denotes perfect relationship, since they occur in exactly the same transactions.

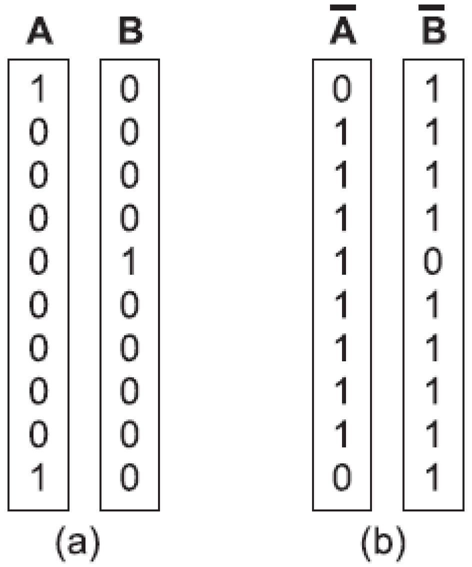
Examples of objective measures for the itemset $\{A, B\}$



Measure (Symbol)	Definition
Correlation (ϕ)	$\frac{N f_{11} - f_{1+} f_{+1}}{\sqrt{f_{1+} f_{+1} f_{0+} f_{+0}}}$
Odds ratio (α)	$(f_{11} f_{00}) / (f_{10} f_{01})$
Kappa (κ)	$\frac{N f_{11} + N f_{00} - f_{1+} f_{+1} - f_{0+} f_{+0}}{N^2 - f_{1+} f_{+1} - f_{0+} f_{+0}}$
Interest (I)	$(N f_{11}) / (f_{1+} f_{+1})$
Cosine (IS)	$(f_{11}) / (\sqrt{f_{1+} f_{+1}})$
Piatetsky-Shapiro (PS)	$\frac{f_{11}}{N} - \frac{f_{1+} f_{+1}}{N^2}$
Collective strength (S)	$\frac{f_{11} + f_{00}}{f_{1+} f_{+1} + f_{0+} f_{+0}} \times \frac{N - f_{1+} f_{+1} - f_{0+} f_{+0}}{N - f_{11} - f_{00}}$
Jaccard (ζ)	$f_{11} / (f_{1+} + f_{+1} - f_{11})$
All-confidence (h)	$\min \left[\frac{f_{11}}{f_{1+}}, \frac{f_{11}}{f_{+1}} \right]$

Properties of Objective Measures

- Inversion Property
- An objective measure *M* is invariant under the inversion operation if its value remains the same when exchanging the frequency counts f_{11} with f_{00} and f_{10} with f_{01}.
- Applying transformation to a binary vector is called **inversion**.
- If a measure is invariant under the inversion operation, then its value for the vector pair $\{A,B\}$ should be identical to its value for $\{\bar{A},\bar{B}\}$.
- Measures that are invariant to the inversion property include the correlation (φ-coefficient), odds ratio, κ, and collective strength.
- These measures are especially useful in scenarios where the presence (1's) of a variable is as important as its absence (0's).
- Measures that do not remain invariant under the inversion operation include the interest factor and the *IS* measure.



A	B	$\bar{A}$	$\bar{B}$
1	0	0	1
0	0	1	1
0	0	1	1
0	0	1	1
0	1	1	0
0	0	1	1
0	0	1	1
0	0	1	1
0	0	1	1
1	0	0	1

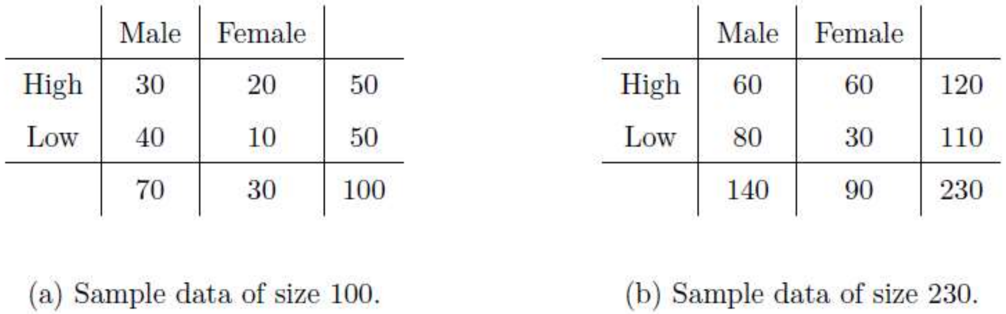
(a) (b)

Properties of Objective Measures

Scaling Invariance Property

- Let T be a contingency table with frequency counts $[f_{11}; f_{10}; f_{01}; f_{00}]$.
- Let T‘ be the transformed contingency table with scaled frequency counts $[k_1k_3f_{11}; k_2k_3f_{10}; k_1k_4f_{01}; k_2k_4f_{00}]$, where k_1, k_2, k_3, k_4 are positive constants used to scale the two rows and the two columns of T.
- An objective measure M is invariant under the row/column scaling operation if $M(T) = M(T’)$.
- A row or column scaling may be performed in the contingency tables.
- Only the odds ratio (α) is invariant to row and column scaling operations.
- φ-coefficient, κ, IS, interest factor, and collective strength (S) change their values
- Odds ratio is a preferred choice of measure in the medical domain, where it is important to find relationships that do not change with differences in the population sample chosen for a study.

Table 5.13. The grade-gender example.



	Male	Female	
High	30	20	50
Low	40	10	50
	70	30	100

(a) Sample data of size 100.

	Male	Female	
High	60	60	120
Low	80	30	110
	140	90	230

(b) Sample data of size 230.

Properties of Objective Measures

Null Addition Property

- An objective measure *M* is invariant under the null addition operation if it is not affected by increasing f_{00}, while all other frequencies in the contingency table stay the same.
- Measures that satisfy -cosine (*IS*) and Jaccard (ξ) measures
- Measures that violate this property include interest factor, *PS*, odds ratio, and the φ-coefficient.

Properties of Objective Measures

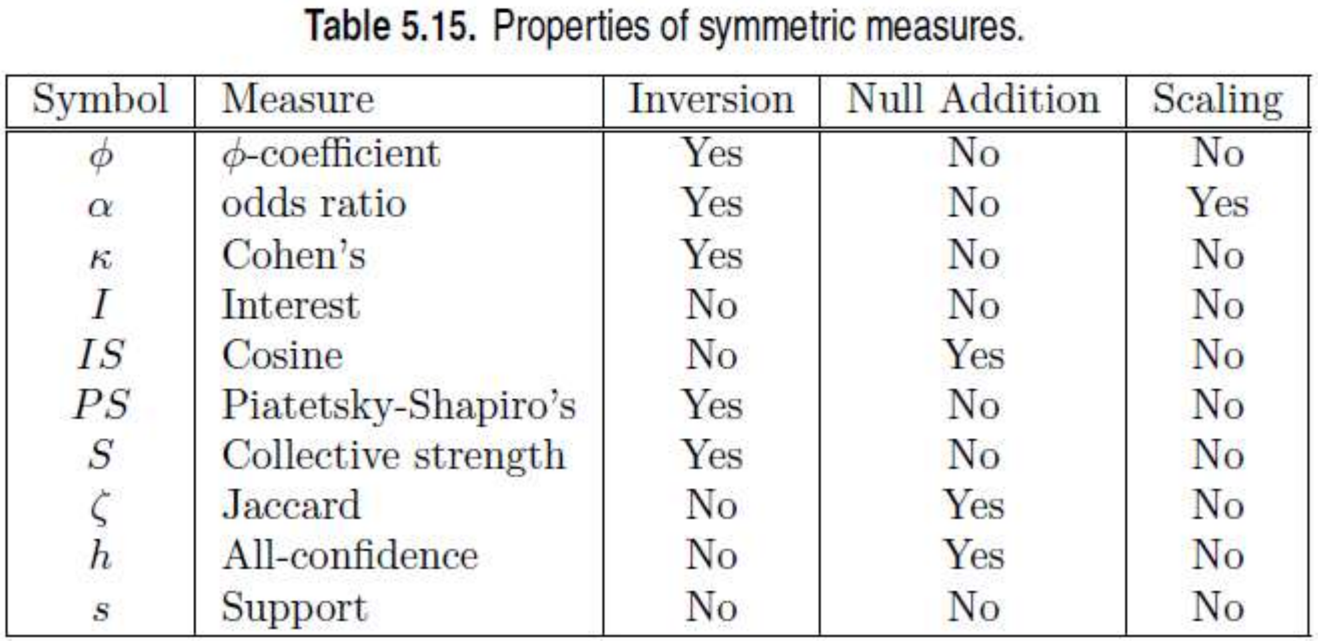


Table 5.15. Properties of symmetric measures.

Symbol	Measure	Inversion	Null Addition	Scaling
ϕ	ϕ-coefficient	Yes	No	No
α	odds ratio	Yes	No	Yes
κ	Cohen's	Yes	No	No
I	Interest	No	No	No
IS	Cosine	No	Yes	No
PS	Piatetsky-Shapiro's	Yes	No	No
S	Collective strength	Yes	No	No
ζ	Jaccard	No	Yes	No
h	All-confidence	No	Yes	No
s	Support	No	No	No

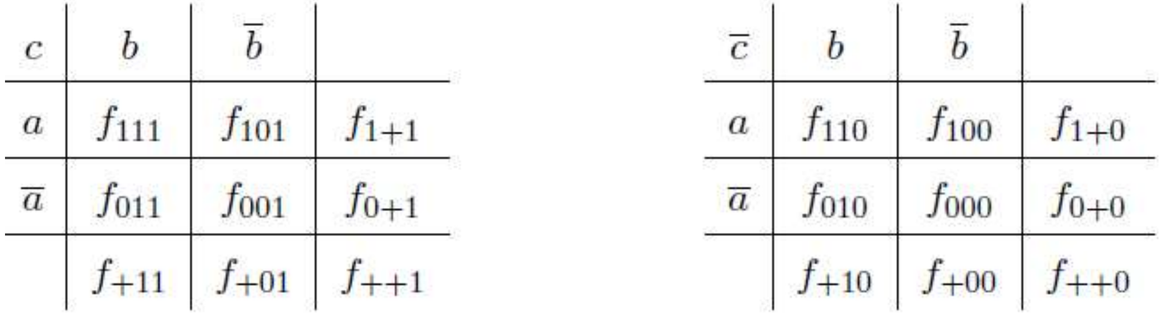
Asymmetric Interestingness Measures

- If M is a measure and A and B are two variables, then $M(A,B)$ is equal to $M(B,A)$ if the order of the variables does not matter. Such measures are called **symmetric**.
- Measures that depend on the order of variables ($M(A,B) \neq M(B,A)$) are called **asymmetric** measures.
- For example, the interest factor is a symmetric measure because its value is identical for the rules $A \longrightarrow B$ and $B \longrightarrow A$.
- In contrast, confidence is an asymmetric measure since the confidence for $A \longrightarrow B$ and $B \longrightarrow A$ may not be the same.
- Asymmetric measures are more suitable for analyzing association rules, since the items in a rule do have a specific order.

Measures beyond Pairs of Binary Variables

- Some Measures are defined for pairs of binary variables others like support and all-confidence, are also applicable to larger-sized itemsets.
- Measures such as interest factor, *IS*, *PS*, and Jaccard coefficient, can be extended to more than two variables using the frequency tables tabulated in a multidimensional contingency table.

Table 5.16. Example of a three-dimensional contingency table.



c	b	$\overline{b}$	
a	f_{111}	f_{101}	f_{1+1}
$\overline{a}$	f_{011}	f_{001}	f_{0+1}
	f_{+11}	f_{+01}	f_{++1}

$\overline{c}$	b	$\overline{b}$	
a	f_{110}	f_{100}	f_{1+0}
$\overline{a}$	f_{010}	f_{000}	f_{0+0}
	f_{+10}	f_{+00}	f_{++0}

- Given a k-itemset $\{i_1, i_2, \ldots, i_k\}$, the condition for statistical independence can be stated as follows:

$$f_{i_1 i_2 \ldots i_k} = \frac{f_{i_1+\ldots+} \times f_{+i_2\ldots+} \times \cdots \times f_{++\ldots i_k}}{N^{k-1}}$$

Measures beyond Pairs of Binary Variables

- Analysis of multidimensional contingency tables is more complicated because of the presence of partial associations in the data.
- For example, some associations may appear or disappear when conditioned upon the value of certain variables.
- This problem is known as **Simpson’s paradox**

Simpson’s paradox

- When interpreting the association between variables , the observed relationship may be influenced by the presence of other confounding factors, i.e., hidden variables that are not included in the analysis.
- In some cases, the hidden variables may cause the observed relationship between a pair of variables to disappear or reverse its direction.
- This phenomenon is known as Simpson’s paradox.

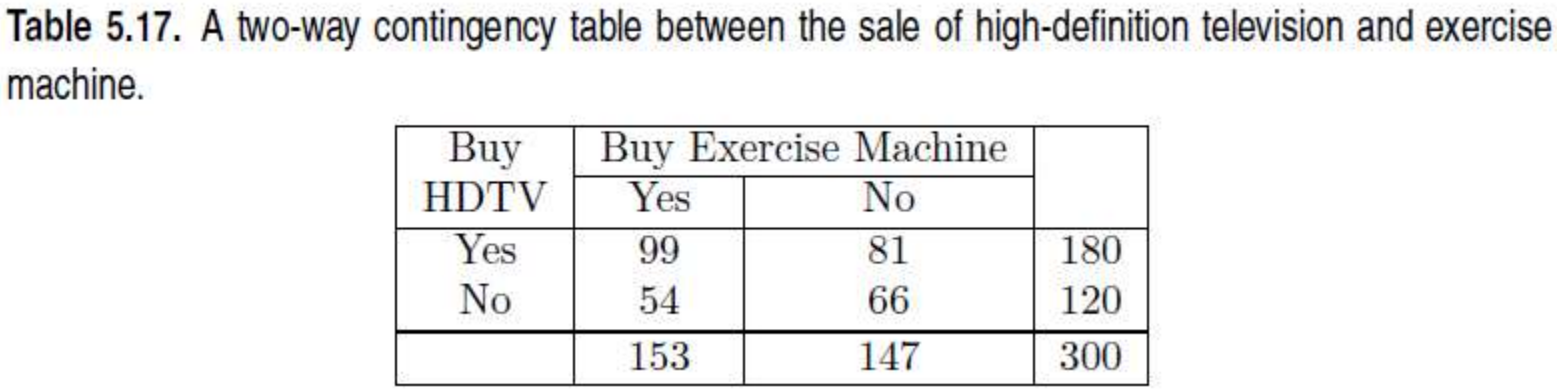
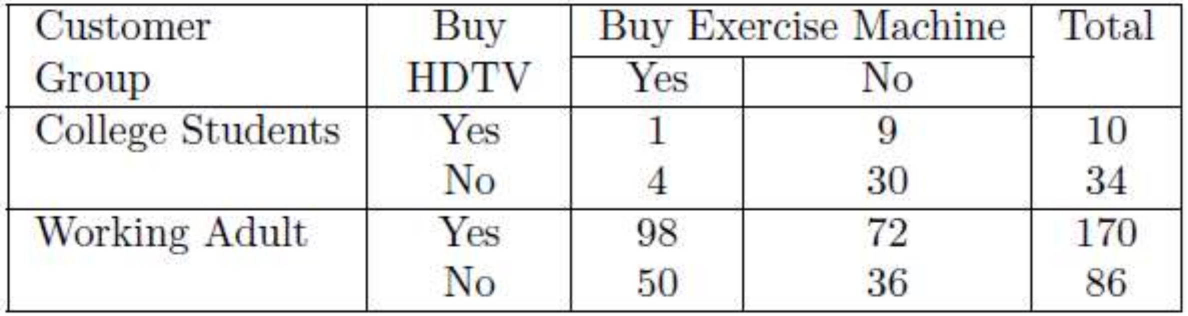


Table 5.17. A two-way contingency table between the sale of high-definition television and exercise machine.

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

- The rule {HDTV=Yes}→ {Exercise machine=Yes} has a confidence of 99/180 = 55% and the rule {HDTV=No} → {Exercise machine=Yes} has a confidence of 54/120 = 45%.
- 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.

Table 5.18. Example of a three-way contingency table.



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

- The total column adds upto 300.
- There are more working adults than college students who buy these items.
- For college students:
- *C* ({HDTV=Yes} → {Exercise machine=Yes}) = 1/10 = 10%,
- *C* ({HDTV=No} −→ {Exercise machine=Yes})= 4/34 = 11.8%,
- For working adults:
- *C* ({HDTV=Yes} −→ {Exercise machine=Yes})= 98/170 = 57.7%,
- *C* ({HDTV=No} −→ {Exercise machine=Yes})= 50/86 = 58.1%.
- Inference: 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.
- This reversal in the direction of association is known as **Simpson's paradox.**

Simpson’s paradox

- The Simpson’s paradox can also be illustrated mathematically as follows.
- Suppose : $a/b < c/d$ and $p/q < r/s$

where a/b and p/q may represent the confidence of the rule $A \rightarrow B$ in two different strata, while c/d and r/s may represent the confidence of the rule $\overline{A} \rightarrow B$ in the two strata.

- When the data is pooled together, the confidence values of the rules in the combined data are $(a+p)/(b+q)$ and $(c+r)/(d+s)$, respectively.
- Simpson’s paradox occurs when $a + p /b + q > c + r/d + s$,thus leading to the wrong conclusion about the relationship between the variables.
- The lesson here is that proper stratification is needed to avoid generating spurious patterns resulting from Simpson’s paradox.
- Eg: supermarket->stratified according to store location

  Medical records -> stratified according to age and gender

Association Using Orange Tool

- https://www.youtube.com/watch?v=WVS24l1zKZU