CS145 Discussion
Week 9

Junheng, Shengming, Yunsheng

11/30/2018

Today’s Discussion

- Announcement
- Roadmap
  - Closed Patterns and Max-Patterns
  - Apriori
  - FP-Growth
  - Association Rules & Pattern Evaluation
  - GSP
  - PrefixSpan
- Q & A

Announcement

- Homework 5 due today 23:59pm (Nov 30, 2018)
  - Submit on CCLE
  - Must include your report and Python code.
- Homework 6 is optional
  - We will drop the lowest HW score, i.e. take the best out of the 5 HW assignments

Closed Patterns and Max-Patterns

- For the transactional database on the right, let min-sup = 58, point out all the maximal frequent pattern(s) and closed frequent pattern(s). (hint: what if min-sup = 60)

T_1 = 2 3 4 ..... 79 (without 1)
T_2 = 1 3 4 ..... 79 (without 2)
:
:
:
:
T_{40}=1 2 3 4 ...... 79 (without 40)

T_{41}= 41 42 43 79
T_{42}= 41 42 43 79
:
:
T_{60}= 41 42 43 ... 79

Closed Patterns and Max-Patterns

- An itemset X is closed if X is *frequent* and there exists *no super-pattern* Y ⊃ X, *with the same support* as X (proposed by Pasquier, et al. @ ICDT'99)
- An itemset X is a max-pattern if X is frequent and there exists no frequent super-pattern Y ⊃ X (proposed by Bayardo @ SIGMOD'98)
- Closed pattern is a lossless compression of freq. patterns
  - Reducing the # of patterns and rules

Apriori

https://edumine.wordpress.com/2013/09/11/apriori-algorithm-simplified-with-an-example/

Apriori

https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html/2

Apriori: Example 1

1

Tid	Items Bought
1	Milk, Tea, Cake
2	Eggs, Tea, Cold Drink
3	Milk, Eggs, Tea, Cold Drink
4	Eggs, Cold Drink
5	Juice

https://t4tutorials.com/apriori-algorithm-in-data-mining-with-examples/

min_support=2

1

Tid	Items Bought
1	Milk, Tea, Cake
2	Eggs, Tea, Cold Drink
3	Milk, Eggs, Tea, Cold Drink
4	Eggs, Cold Drink
5	Juice

2

Items Bought	Support
Milk	2
Eggs	3
Tea	3
Cold Drink	3
Juice	1
Cake	1

https://t4tutorials.com/apriori-algorithm-in-data-mining-with-examples/

min_support=2

1

Tid	Items Bought
1	Milk, Tea, Cake
2	Eggs, Tea, Cold Drink
3	Milk, Eggs, Tea, Cold Drink
4	Eggs, Cold Drink
5	Juice

2

Items Bought	Support
Milk	2
Eggs	3
Tea	3
Cold Drink	3
Juice	1
Cake	1

3

Items Bought	Support
Milk	2
Eggs	3
Tea	3
Cold Drink	3

https://t4tutorials.com/apriori-algorithm-in-data-mining-with-examples/

min_support=2

1

Tid	Items Bought
1	Milk, Tea, Cake
2	Eggs, Tea, Cold Drink
3	Milk, Eggs, Tea, Cold Drink
4	Eggs, Cold Drink
5	Juice

2

Items Bought	Support
Milk	2
Eggs	3
Tea	3
Cold Drink	3
Juice	1
Cake	1

3

Items Bought	Support
Milk	2
Eggs	3
Tea	3
Cold Drink	3

4

Items Bought
Milk, Eggs
Milk, Tea
Milk, Cold Drink
Eggs, Tea
Eggs, Cold Drink
Tea, Cold Drink

https://t4tutorials.com/apriori-algorithm-in-data-mining-with-examples/

min_support=2

1

Tid	Items Bought
1	Milk, Tea, Cake
2	Eggs, Tea, Cold Drink
3	Milk, Eggs, Tea, Cold Drink
4	Eggs, Cold Drink
5	Juice

2

Items Bought	Support
Milk	2
Eggs	3
Tea	3
Cold Drink	3
Juice	1
Cake	1

3

Items Bought	Support
Milk	2
Eggs	3
Tea	3
Cold Drink	3

4

Items Bought
Milk, Eggs
Milk, Tea
Milk, Cold Drink
Eggs, Tea
Eggs, Cold Drink
Tea, Cold Drink

5

Items Bought	Support
Milk, Eggs	1
Milk, Tea	2
Milk, Cold Drink	1
Eggs, Tea	2
Eggs, Cold Drink	3
Tea, Cold Drink	2

https://t4tutorials.com/apriori-algorithm-in-data-mining-with-examples/

min_support=2

1

Tid	Items Bought
1	Milk, Tea, Cake
2	Eggs, Tea, Cold Drink
3	Milk, Eggs, Tea, Cold Drink
4	Eggs, Cold Drink
5	Juice

2

Items Bought	Support
Milk	2
Eggs	3
Tea	3
Cold Drink	3
Juice	1
Cake	1

3

Items Bought	Support
Milk	2
Eggs	3
Tea	3
Cold Drink	3

4

Items Bought
Milk, Eggs
Milk, Tea
Milk, Cold Drink
Eggs, Tea
Eggs, Cold Drink
Tea, Cold Drink

5

Items Bought	Support
Milk, Eggs	1
Milk, Tea	2
Milk, Cold Drink	1
Eggs, Tea	2
Eggs, Cold Drink	3
Tea, Cold Drink	2

6

Items Bought	Support
Milk, Tea	2
Eggs, Tea	2
Eggs, Cold Drink	3
Tea, Cold Drink	2

https://t4tutorials.com/apriori-algorithm-in-data-mining-with-examples/

min_support=2

1

Tid	Items Bought
1	Milk, Tea, Cake
2	Eggs, Tea, Cold Drink
3	Milk, Eggs, Tea, Cold Drink
4	Eggs, Cold Drink
5	Juice

2

Items Bought	Support
Milk	2
Eggs	3
Tea	3
Cold Drink	3
Juice	1
Cake	1

3

Items Bought	Support
Milk	2
Eggs	3
Tea	3
Cold Drink	3

4

Items Bought
Milk, Eggs
Milk, Tea
Milk, Cold Drink
Eggs, Tea
Eggs, Cold Drink
Tea, Cold Drink

5

Items Bought	Support
Milk, Eggs	1
Milk, Tea	2
Milk, Cold Drink	1
Eggs, Tea	2
Eggs, Cold Drink	3
Tea, Cold Drink	2

6

Items Bought	Support
Milk, Tea	2
Eggs, Tea	2
Eggs, Cold Drink	3
Tea, Cold Drink	2

7

Items Bought	Support
Eggs, Tea, Cold Drink	2

There is only one itemset with minimum support 2.
So only one itemset is frequent

https://t4tutorials.com/apriori-algorithm-in-data-mining-with-examples/

min_support=2

Apriori: Example 2

1

Tid	Items Bought
1	Milk, Tea, Cake
2	Eggs, Tea, Cold Drink
3	Milk, Eggs, Tea, Cold Drink
4	Eggs, Cold Drink
5	Juice

https://t4tutorials.com/apriori-algorithm-in-data-mining-with-examples/

min_support=3

1

Tid	Items Bought
1	Milk, Tea, Cake
2	Eggs, Tea, Cold Drink
3	Milk, Eggs, Tea, Cold Drink
4	Eggs, Cold Drink
5	Juice

2

Items Bought	Support
Milk	2
Eggs	3
Tea	3
Cold Drink	3
Juice	1
Cake	1

https://t4tutorials.com/apriori-algorithm-in-data-mining-with-examples/

min_support=3

1

Tid	Items Bought
1	Milk, Tea, Cake
2	Eggs, Tea, Cold Drink
3	Milk, Eggs, Tea, Cold Drink
4	Eggs, Cold Drink
5	Juice

2

Items Bought	Support
Milk	2
Eggs	3
Tea	3
Cold Drink	3
Juice	1
Cake	1

3

Items Bought	Support
Eggs	3
Tea	3
Cold Drink	3

https://t4tutorials.com/apriori-algorithm-in-data-mining-with-examples/

min_support=3

1

Tid	Items Bought
1	Milk, Tea, Cake
2	Eggs, Tea, Cold Drink
3	Milk, Eggs, Tea, Cold Drink
4	Eggs, Cold Drink
5	Juice

2

Items Bought	Support
Milk	2
Eggs	3
Tea	3
Cold Drink	3
Juice	1
Cake	1

3

Items Bought	Support
Eggs	3
Tea	3
Cold Drink	3

4

Items Bought
Eggs, Tea
Eggs, Cold Drink
Tea, Cold Drink

https://t4tutorials.com/apriori-algorithm-in-data-mining-with-examples/

min_support=3

1

Tid	Items Bought
1	Milk, Tea, Cake
2	Eggs, Tea, Cold Drink
3	Milk, Eggs, Tea, Cold Drink
4	Eggs, Cold Drink
5	Juice

2

Items Bought	Support
Milk	2
Eggs	3
Tea	3
Cold Drink	3
Juice	1
Cake	1

3

Items Bought	Support
Eggs	3
Tea	3
Cold Drink	3

4

Items Bought
Eggs, Tea
Eggs, Cold Drink
Tea, Cold Drink

5

Items Bought	Support
Eggs, Tea	2
Eggs, Cold Drink	3
Tea, Cold Drink	2

https://t4tutorials.com/apriori-algorithm-in-data-mining-with-examples/

min_support=3

1

Tid	Items Bought
1	Milk, Tea, Cake
2	Eggs, Tea, Cold Drink
3	Milk, Eggs, Tea, Cold Drink
4	Eggs, Cold Drink
5	Juice

2

Items Bought	Support
Milk	2
Eggs	3
Tea	3
Cold Drink	3
Juice	1
Cake	1

3

Items Bought	Support
Eggs	3
Tea	3
Cold Drink	3

4

Items Bought
Eggs, Tea
Eggs, Cold Drink
Tea, Cold Drink

5

Items Bought	Support
Eggs, Tea	2
Eggs, Cold Drink	3
Tea, Cold Drink	2

6

Items Bought	Support
Eggs, Cold Drink	3

https://t4tutorials.com/apriori-algorithm-in-data-mining-with-examples/

min_support=3

1

Tid	Items Bought
1	Milk, Tea, Cake
2	Eggs, Tea, Cold Drink
3	Milk, Eggs, Tea, Cold Drink
4	Eggs, Cold Drink
5	Juice

2

Items Bought	Support
Milk	2
Eggs	3
Tea	3
Cold Drink	3
Juice	1
Cake	1

3

Items Bought	Support
Eggs	3
Tea	3
Cold Drink	3

4

Items Bought
Eggs, Tea
Eggs, Cold Drink
Tea, Cold Drink

5

Items Bought	Support
Eggs, Tea	2
Eggs, Cold Drink	3
Tea, Cold Drink	2

6

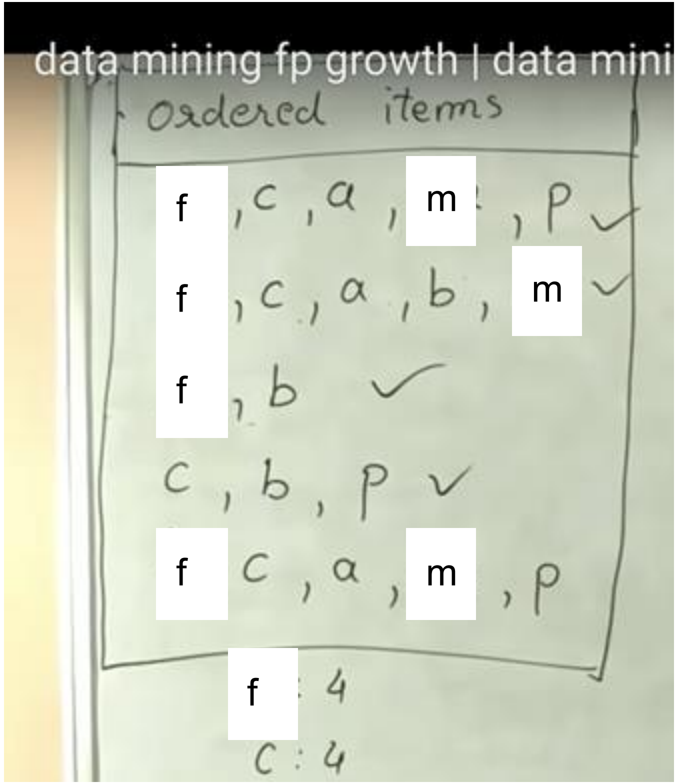
Items Bought	Support
Eggs, Cold Drink	3

There is no itemset with minimum support 3, so there is no frequent itemset because there are 0 itemset that have minimum support 3

https://t4tutorials.com/apriori-algorithm-in-data-mining-with-examples/

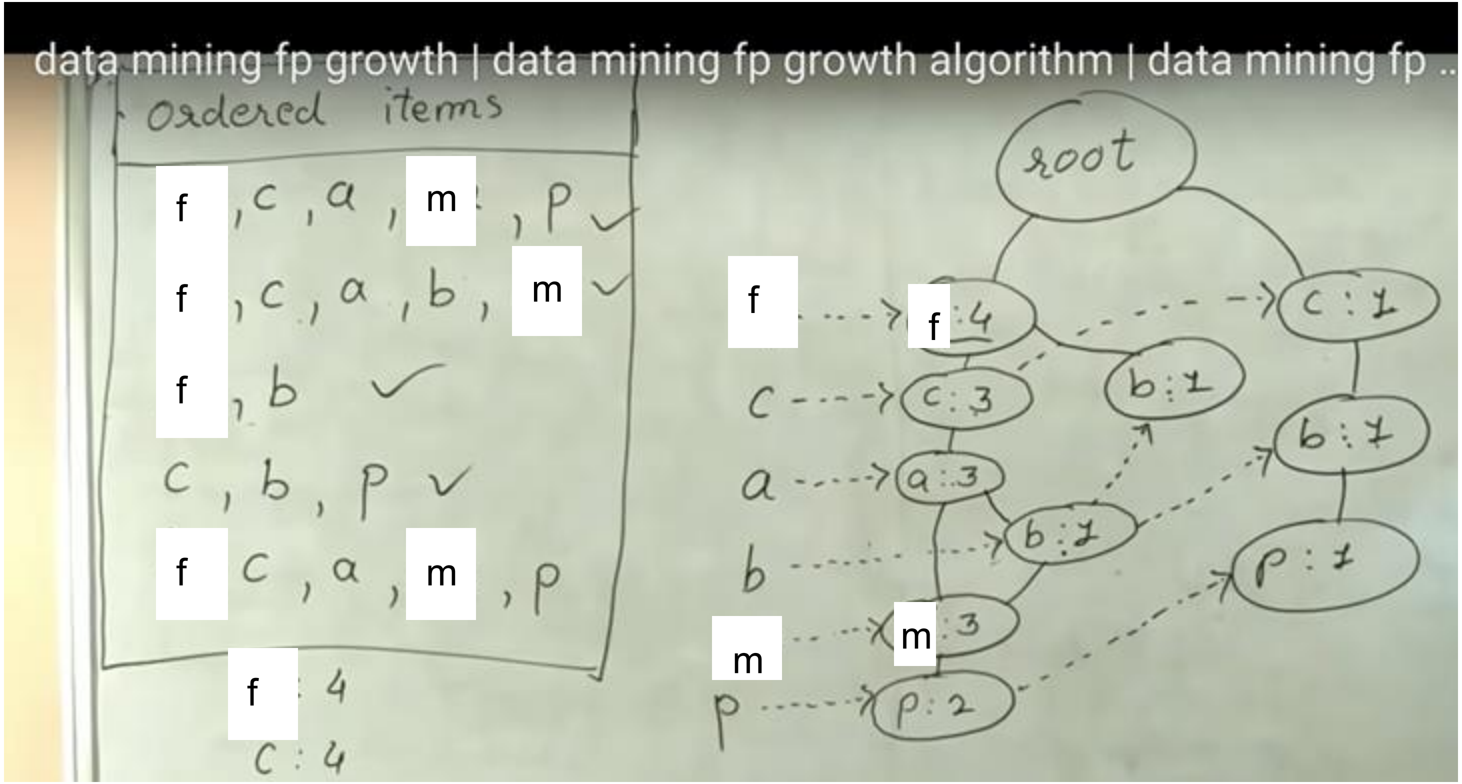
min_support=3

FP-Growth



https://www.youtube.com/watch?v=gq6nKbye648

FP-Growth



https://www.youtube.com/watch?v=gq6nKbye648

Association Rules & Pattern Evaluation

- Support

  Support {🍎} = $\frac{4}{8}$

- Confidence

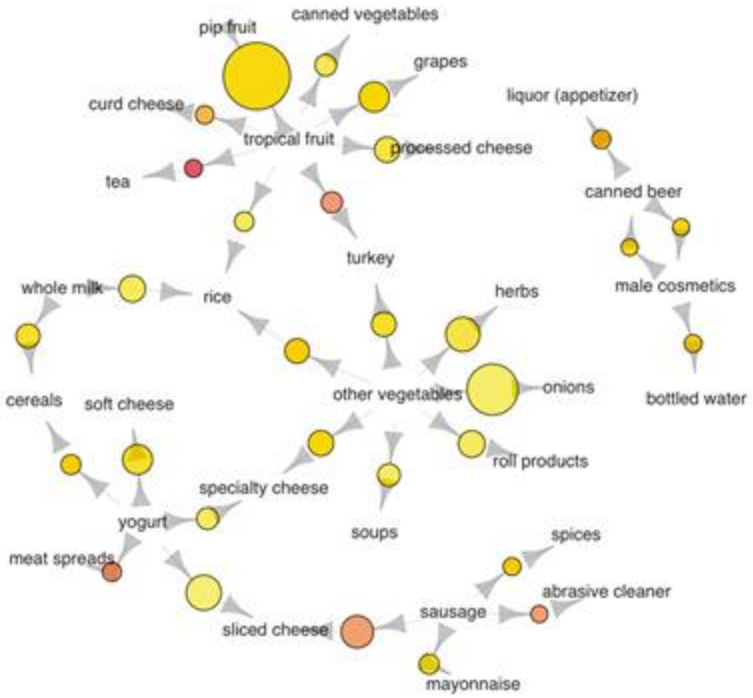
  Confidence {🍎 → 🍺} = $\frac{\text{Support }\{🍎, 🍺\}}{\text{Support }\{🍎\}}$

- Recall: $Confidence(A \Rightarrow B) = P(B|A) = \frac{support(A \cup B)}{support(A)}$
  - What is the misleading problem of it?

Transaction 1	🍎 🍺 🍚 🍗
Transaction 2	🍎 🍺 🍚
Transaction 3	🍎 🍺
Transaction 4	🍎 🍐
Transaction 5	🍼 🍺 🍚 🍗
Transaction 6	🍼 🍺 🍚
Transaction 7	🍼 🍺
Transaction 8	🍼 🍐

https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html

Drawback of the Confidence Measure



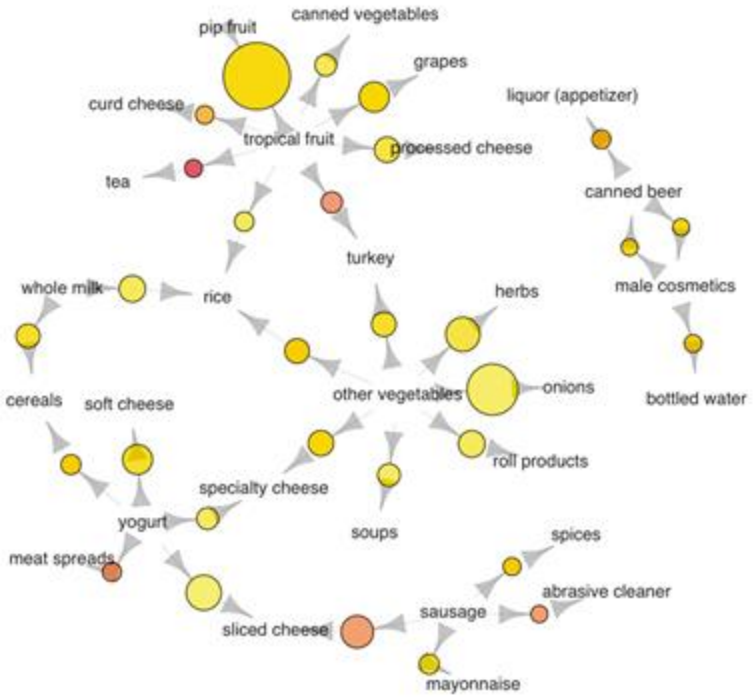
Transaction	Support
Canned Beer	10%
Soda	20%
Berries	3%
Male Cosmetics	0.5%

Transaction	Support	Confidence	Lift
Canned Beer → Soda	1%	20%	1.0
Canned Beer → Berries	0.1%	1%	0.3
Canned Beer → Male Cosmetics	0.1%	1%	2.6

The {beer -> soda} rule has the highest confidence at 20%. However, both beer and soda appear frequently across all transactions, so their association could simply be a fluke. This is confirmed by the lift value of {beer -> soda}, which is 1, implying no association between beer and soda.

https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html

Drawback of the Confidence Measure



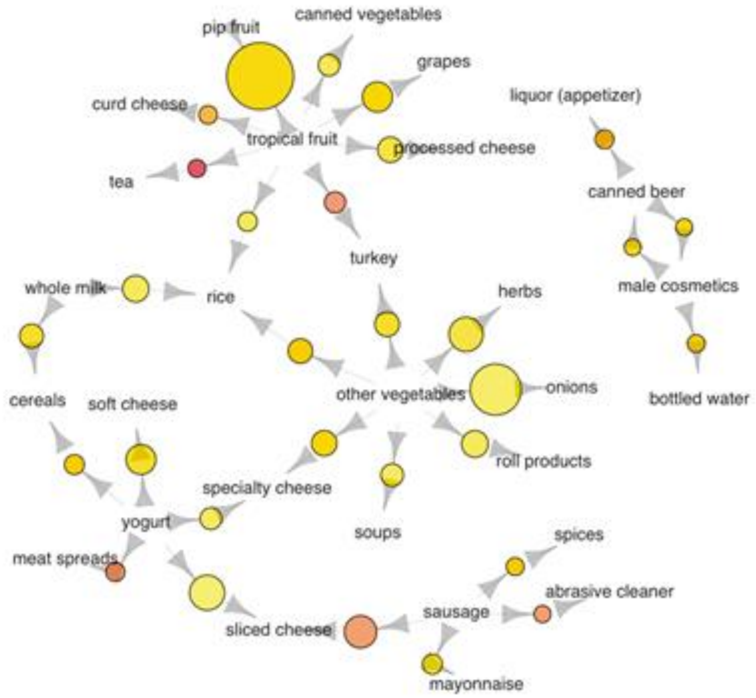
Transaction	Support
Canned Beer	10%
Soda	20%
Berries	3%
Male Cosmetics	0.5%

On the other hand, the {beer -> male cosmetics} rule has a low confidence, due to few purchases of male cosmetics in general.

Transaction	Support	Confidence	Lift
Canned Beer → Soda	1%	20%	1.0
Canned Beer → Berries	0.1%	1%	0.3
Canned Beer → Male Cosmetics	0.1%	1%	2.6

https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html

Drawback of the Confidence Measure



Transaction	Support
Canned Beer	10%
Soda	20%
Berries	3%
Male Cosmetics	0.5%

However, whenever someone does buy male cosmetics, he is very likely to buy beer as well, as inferred from a high lift value of 2.6. The converse is true for {beer -> berries}. With a lift value below 1, we may conclude that if someone buys berries, he would likely be averse to beer.

Transaction	Support	Confidence	Lift
Canned Beer → Soda	1%	20%	1.0
Canned Beer → Berries	0.1%	1%	0.3
Canned Beer → Male Cosmetics	0.1%	1%	2.6

https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html

Association Rules & Pattern Evaluation

- Lift

Lift {🍎 → 🍺} = Support {🍎, 🍺} / (Support {🍎} x Support {🍺})

$$lift = \frac{P(A \cup B)}{P(A)P(B)}$$

1: independent
>1: positively correlated
<1: negatively correlated

Transaction 1	🍎 🍺 🍚 🍗
Transaction 2	🍎 🍺 🍚
Transaction 3	🍎 🍺
Transaction 4	🍎 🍐
Transaction 5	🍼 🍺 🍚 🍗
Transaction 6	🍼 🍺 🍚
Transaction 7	🍼 🍺
Transaction 8	🍼 🍐

https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html

Association Rules & Pattern Evaluation

Support {🍎} = $\frac{4}{8}$

Confidence {🍎 → 🍺} =

$\frac{\text{Support \{🍎, 🍺\}}}{\text{Support \{🍎\}}}$

Lift {🍎 → 🍺} =

$\frac{\text{Support \{🍎, 🍺\}}}{\text{Support \{🍎\} x Support \{🍺\}}}$

Rule	Support	Confidence	Lift
$A \Rightarrow D$			
$C \Rightarrow A$			
$A \Rightarrow C$			
$B\&C \Rightarrow D$			

https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html

UCLA

Association Rules & Pattern Evaluation

Support {🍎} = $\frac{4}{8}$

Confidence {🍎 → 🍺} = $\frac{\text{Support } \{🍎, 🍺\}}{\text{Support } \{🍎\}}$

Lift {🍎 → 🍺} = $\frac{\text{Support } \{🍎, 🍺\}}{\text{Support } \{🍎\} \times \text{Support } \{🍺\}}$

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html

Chi-Square Test

- Question: Are education level and marital status related?

	name	marit	educ
1	Cameron	Never married	PhD or higher
2	Benjamin	Married	Middle school or lower
3	Camden	Divorced	Bachelor's
4	Brody	Widowed	PhD or higher
5	Connor	Married	PhD or higher

https://www.spss-tutorials.com/chi-square-independence-test/

Chi-Square Test

- Check the contingency table of marital status by education:

Marital Status by Education | n = 300

	Middle school or lower	High school	Bachelor's	Master's	PhD or higher	Total
Never married	18	36	21	9	6	90
Married	12	36	45	36	21	150
Divorced	6	9	9	3	3	30
Widowed	3	9	9	6	3	30
Total	39	90	84	54	33	300

https://www.spss-tutorials.com/chi-square-independence-test/

Chi-Square Test

- Is marital status related to education level and -if so- how?

Marital Status by Education | n = 300

	Middle school or lower	High school	Bachelor's	Master's	PhD or higher	Total
Never married	46%	40%	25%	17%	18%	30%
Married	31%	40%	54%	67%	64%	50%
Divorced	15%	10%	11%	6%	9%	10%
Widowed	8%	10%	11%	11%	9%	10%
Total	100%	100%	100%	100%	100%	100%

https://www.spss-tutorials.com/chi-square-independence-test/

Chi-Square Test

- Highly educated respondents → marry more often than less educated

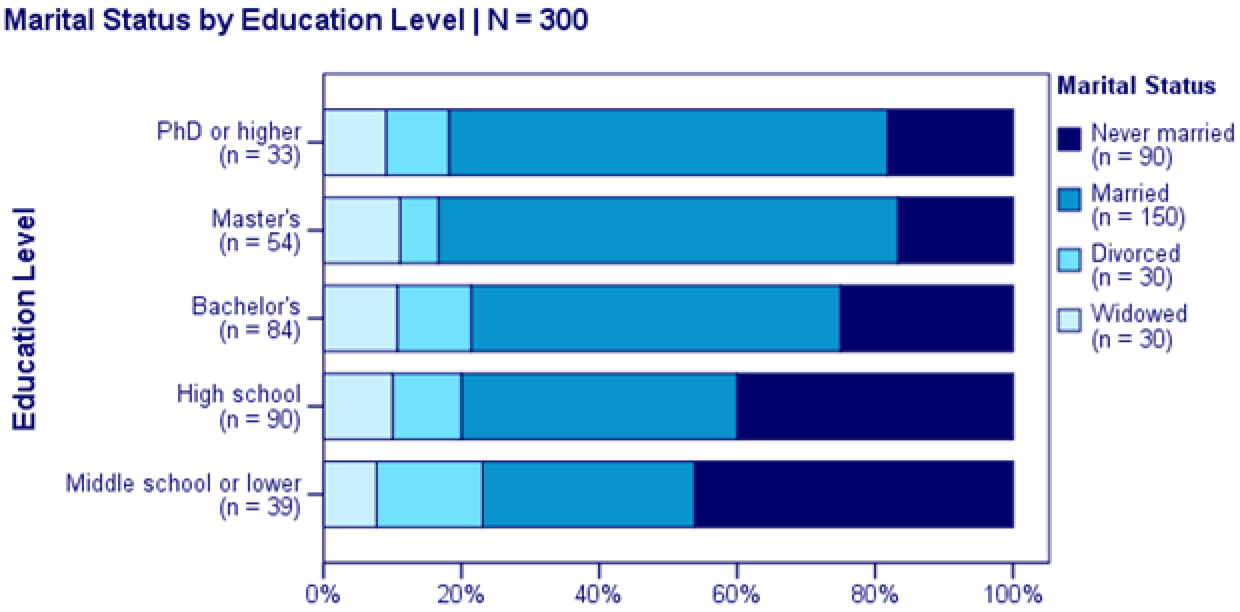
Marital Status by Education | n = 300

	Middle school or lower	High school	Bachelor's	Master's	PhD or higher	Total
Never married	46%	40%	25%	17%	18%	30%
Married	31%	40%	54%	67%	64%	50%
Divorced	15%	10%	11%	6%	9%	10%
Widowed	8%	10%	11%	11%	9%	10%
Total	100%	100%	100%	100%	100%	100%

https://www.spss-tutorials.com/chi-square-independence-test/

Chi-Square Test

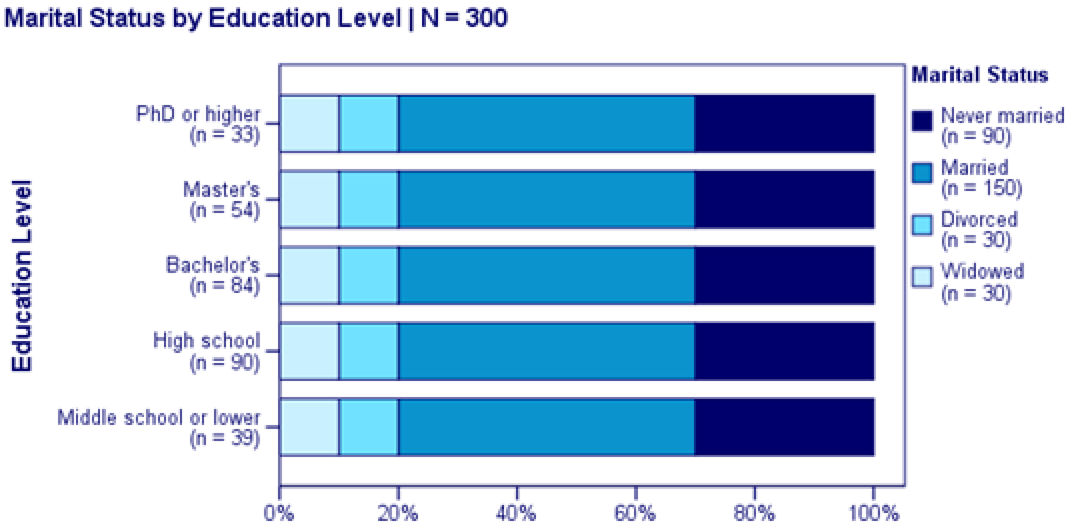
- Marital status is clearly associated with education level.



https://www.spss-tutorials.com/chi-square-independence-test/

Chi-Square Test

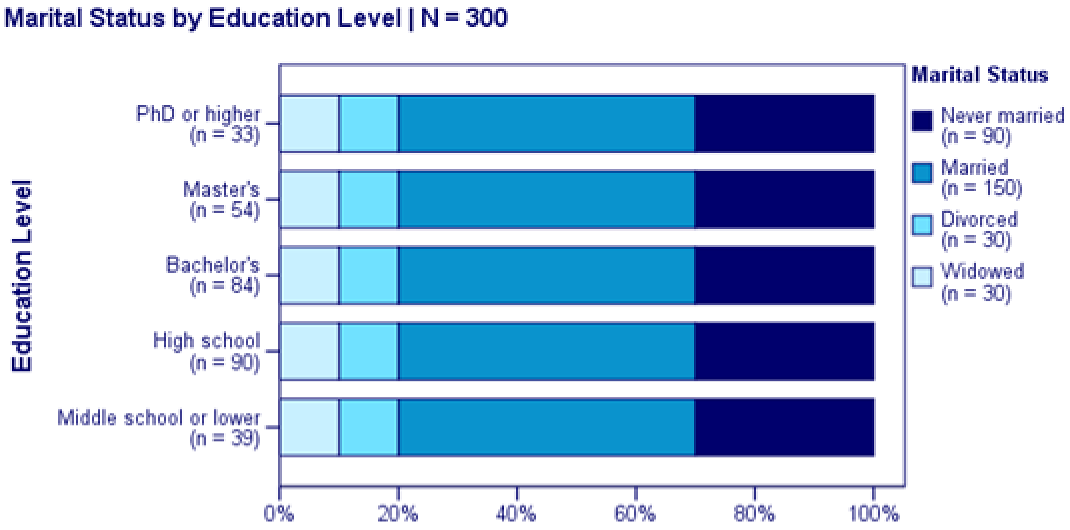
- The ***null hypothesis*** for a chi-square independence test is that
  - two categorical variables are ***independent*** in some population.



https://www.spss-tutorials.com/chi-square-independence-test/

Chi-Square Test

- ***Statistical independence*** means that
  - the frequency distribution of a variable is the same for all levels of some other variable.



https://www.spss-tutorials.com/chi-square-independence-test/

Chi-Square Test

- ***Expected frequencies*** are
  - the frequencies we expect in our sample if the ***null hypothesis*** holds.

Expected Frequencies for Perfectly Independent Variables

	Middle school or lower	High school	Bachelor's	Master's	PhD or higher	Total
Never married	11.7	27.0	25.2	16.2	9.9	90.0
Married	19.5	45.0	42.0	27.0	16.5	150.0
Divorced	3.9	9.0	8.4	5.4	3.3	30.0
Widowed	3.9	9.0	8.4	5.4	3.3	30.0
Total	39.0	90.0	84.0	54.0	33.0	300.0

https://www.spss-tutorials.com/chi-square-independence-test/

Chi-Square Test

- ***Assuming independence*** ***(null hypothesis)***

P(middle,never)=P(middle)P(never)=(39/300)*(90/300)
Expected # of (middle,never) =
300*P(middle,never)=39*90/300=11.7

Expected Frequencies for Perfectly Independent Variables

	Middle school or lower	High school	Bachelor's	Master's	PhD or higher	Total
Never married	11.7	27.0	25.2	16.2	9.9	90.0
Married	19.5	45.0	42.0	27.0	16.5	150.0
Divorced	3.9	9.0	8.4	5.4	3.3	30.0
Widowed	3.9	9.0	8.4	5.4	3.3	30.0
Total	39.0	90.0	84.0	54.0	33.0	300.0

https://www.spss-tutorials.com/chi-square-independence-test/

Chi-Square Test

- ***Assuming independence,***
- ***→ expected frequencies***

P(middle,never)=P(middle)P(never)=(39/300)*(90/300)
Expected # of (middle,never) =
300*P(middle,never)=39*90/300=11.7

Expected Frequencies for Perfectly Independent Variables

	Middle school or lower	High school	Bachelor's	Master's	PhD or higher	Total
Never married	11.7	27.0	25.2	16.2	9.9	90.0
Married	19.5	45.0	42.0	27.0	16.5	150.0
Divorced	3.9	9.0	8.4	5.4	3.3	30.0
Widowed	3.9	9.0	8.4	5.4	3.3	30.0
Total	39.0	90.0	84.0	54.0	33.0	300.0

https://www.spss-tutorials.com/chi-square-independence-test/

Chi-Square Test

- Real data
- → ***observed frequencies***:

Marital Status by Education | n = 300

	Middle school or lower	High school	Bachelor's	Master's	PhD or higher	Total
Never married	18	36	21	9	6	90
Married	12	36	45	36	21	150
Divorced	6	9	9	3	3	30
Widowed	3	9	9	6	3	30
Total	39	90	84	54	33	300

https://www.spss-tutorials.com/chi-square-independence-test/

Chi-Square Test

- Add up the differences for each of the 5*4=20 cells
    - → $\chi 2$

- $$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

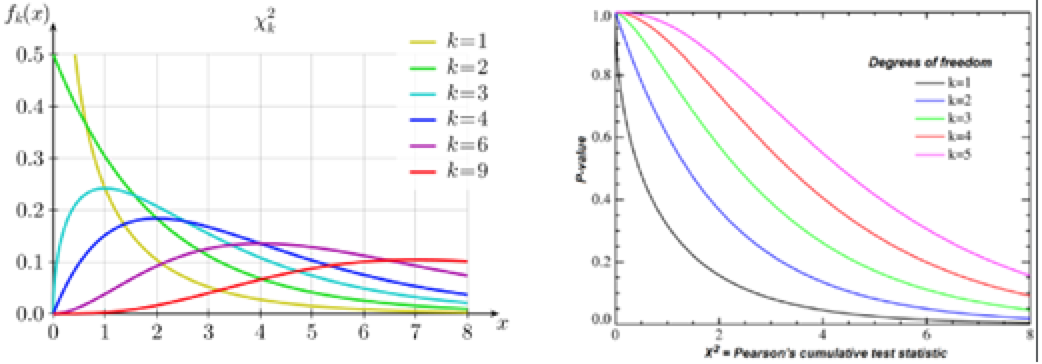
$$\chi^2 = \frac{(18-11.7)^2}{11.7} + \frac{(36-27)^2}{27} + \ldots + \frac{(6-5.4)^2}{5.4} = 23.57$$

https://www.spss-tutorials.com/chi-square-independence-test/

Chi-Square Test

- Is χ^2=23.57 a large value?
  - If yes, reject the null hypothesis → A and B are dependent
  - But how to tell if it is a large value?
    - χ^2 → Follows Chi-squared distribution with degree of freedom as $(r-1) \times (c-1)$

Pearson established it in 1900. [See more](#).



https://www.spss-tutorials.com/chi-square-independence-test/

Chi-Square Test

- What is a chi-squared distribution?

Definition [edit]

If $Z_1, \ldots, Z_k$ are independent, standard normal random variables, then the sum of their squares,

$$Q = \sum_{i=1}^{k} Z_i^2,$$

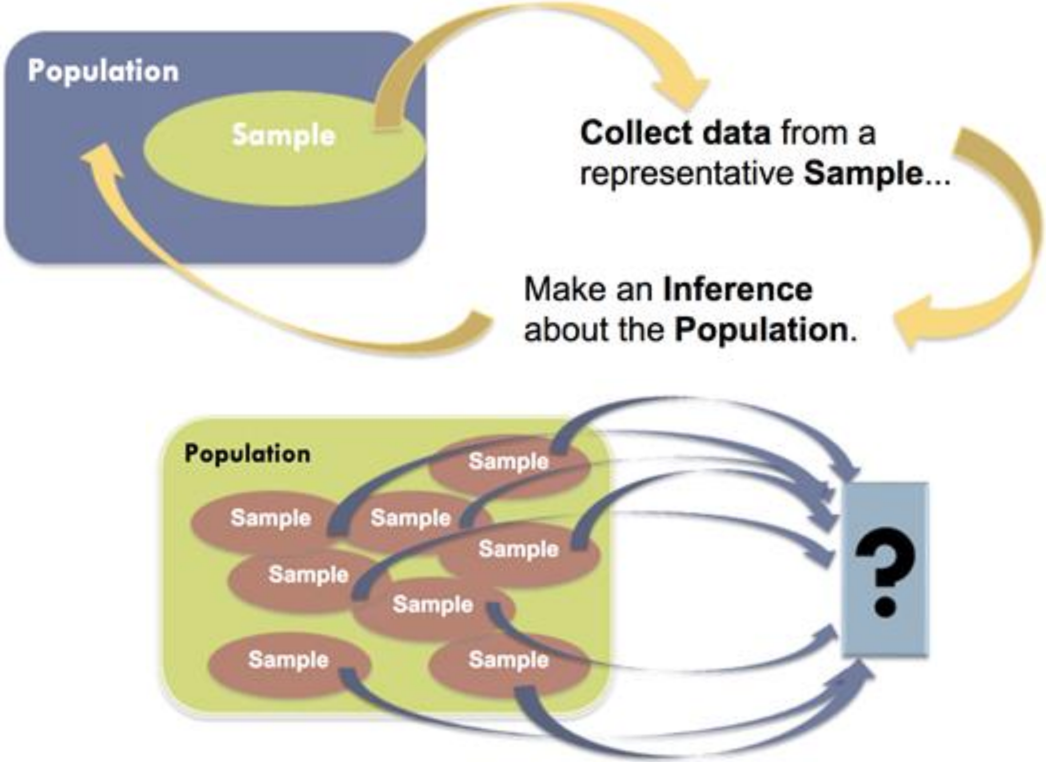
is distributed according to the chi-squared distribution with k degrees of freedom. This is usually denoted as

$$Q \sim \chi^2(k) \ \text{ or } \ Q \sim \chi_k^2.$$

The chi-squared distribution has one parameter: k, a positive integer that specifies the number of degrees of freedom (the number of Z_i's).

https://en.wikipedia.org/wiki/Chi-squared_distribution

Sampling Distribution vs Population Distribution



https://slideplayer.com/slide/5775066/

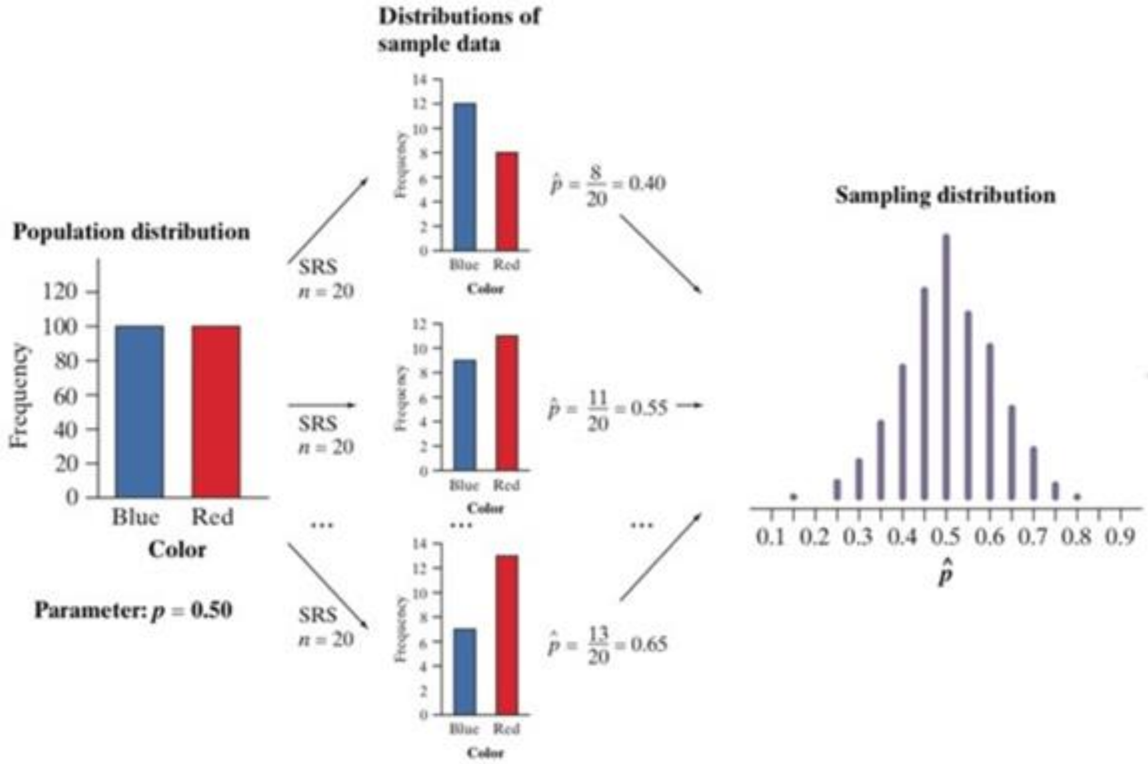
Sampling Distribution vs Population Distribution

Draw observations at random from any population with finite mean μ. The **law of large numbers** says that as the number of observations drawn increases, the sample mean of the observed values gets closer and closer to the mean μ of the population.

The **population distribution** of a variable is the distribution of values of the variable among all individuals in the population.

The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

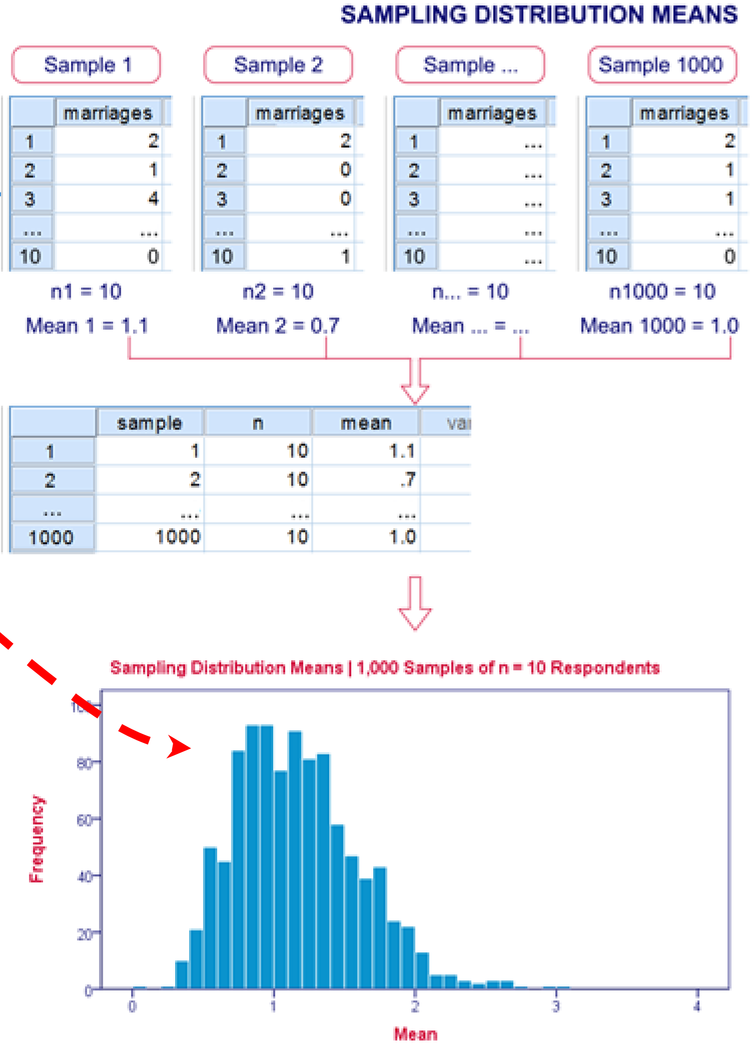
Sampling Distribution vs Population Distribution



https://slideplayer.com/slide/5775066/

Sample **means and sums** are **always normally distributed** (approximately) for reasonable sample sizes, say n > 30. This doesn't depend on whatever population distribution the data values may or may not follow.*

This phenomenon is known as the central limit theorem.



https://www.spss-tutorials.com/sampling-distribution-what-is-it/

Chi-Square Test

- In this example,
  - What is the sampling distribution of $\chi 2$?
  - Under what assumptions does the above hold?

https://www.spss-tutorials.com/chi-square-independence-test/

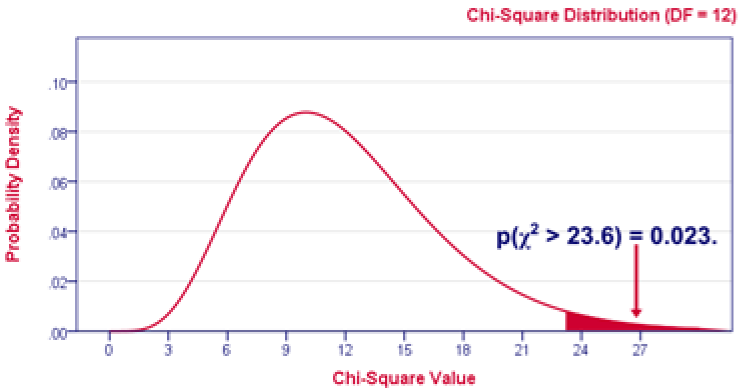
Chi-Square Test

- In this example,
  - What is the sampling distribution of $\chi 2$? $\chi 2$ follows a $\chi 2$ distribution.
  - Under what assumptions does the above hold? Independent observations, etc.

https://www.spss-tutorials.com/chi-square-independence-test/

Chi-Square Test

- In this example, $df = (5-1) \cdot (4-1) = 12.$
  - How to interpret P(χ2>23.6)=0.023?
    - The probability of _____ under _____ assumptions is very small, 2.3%.
  - A small p-value basically means that the data are unlikely under some null hypothesis. A somewhat arbitrary convention is to reject the null hypothesis if $p < 0.05$.
  - Should we reject the null hypothesis in this case? Yes!
    - ***"An association between education and marital status was observed, χ2(12) = 23.57, p = 0.023."***



- Lift and χ^2 are affected by null-transaction
  - E.g., number of transactions that do not contain milk nor coffee

Why?

https://www.spss-tutorials.com/chi-square-independence-test/
https://www.spss-tutorials.com/statistical-significance/

Summary: Conceptual Understanding

- What are ***hypothesis testing***, ***p-value***, and ***significance level***?
    - ***Hypothesis testing*** is the use of statistics to determine the probability that a given hypothesis is true. The usual process of hypothesis testing consists of four steps.
        - **Formulate the null hypothesis H_0** (commonly, that the observations are the result of pure chance) and the alternative hypothesis H_a (commonly, that the observations show a real effect combined with a component of chance variation).
        - **Identify a test statistic** that can be used to assess the truth of the null hypothesis.
        - **Compute the *p-value***, which is the probability of obtaining an effect *at least as extreme as the one in the sample* data assuming that the null hypothesis were true. The smaller the p-value, the stronger the evidence against the null hypothesis.
        - **Compare the p-value to an acceptable *significance value (level)*** α. If p<=α, that the observed effect is statistically significant, the null hypothesis is ruled out, and the alternative hypothesis is valid.

http://mathworld.wolfram.com/HypothesisTesting.html
http://blog.minitab.com/blog/adventures-in-statistics-2/understanding-hypothesis-tests-significance-levels-alpha-and-p-values-in-statistics
http://www.stat.yale.edu/Courses/1997-98/101/sigtest.htm

Summary: Conceptual Questions

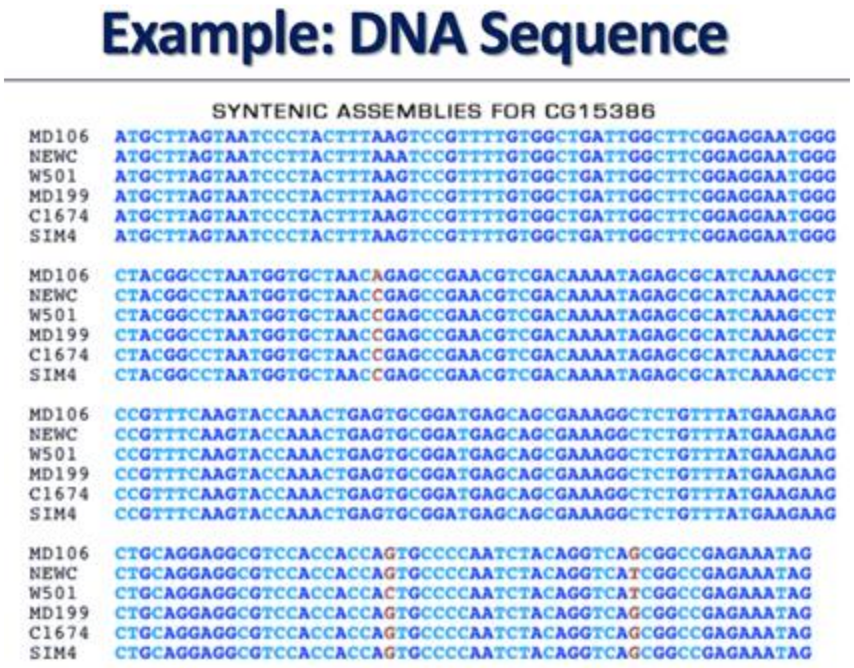
- In the previous example,
  - What is the null hypothesis?
  - What is the test statistics?
  - What is the p-value?
  - What is the significance level?
  - What is the conclusion?

UCLA

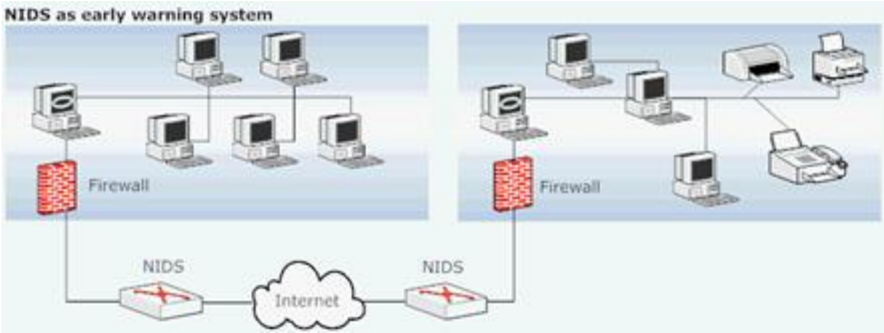
Summary: Conceptual Questions

- In the previous example,
  - What is the null hypothesis? There is ***no*** association between the two variables.
  - What is the test statistics? $\chi^2=23.6$
  - What is the p-value? $P(\chi^2>23.6)=0.023$
  - What is the significance level? 0.05
  - What is the conclusion? There is an association between the two variables.

Sequential Pattern Mining: Applications



- Music: midi files



https://www.cse.wustl.edu/~jain/cse571-07/ftp/ids/index.html#sec2

GSP

- Initial candidates: all singleton sequences
  - <a>, <b>, <c>, <d>, <e>, <f>, <g>, <h>
- Scan database once, count support for candidates

Seq. ID	Sequence
1	<(cd)(abc)(abf)(acdf)>
2	<(abf)(e)>
3	<(abf)>
4	<(dgh)(bf)(agh)>

Cand	Sup
<a>	
<b>	
<c>	
<d>	
<e>	
<f>	
<g>	
<h>	

https://www.slideshare.net/Krish_ver2/53-mining-sequential-patterns

GSP

- Initial candidates: all singleton sequences
  - <a>, <b>, <c>, <d>, <e>, <f>, <g>, <h>
- Scan database once, count support for candidates

Seq. ID	Sequence
1	<(cd)(abc)(abf)(acdf)>
2	<(abf)(e)>
3	<(abf)>
4	<(dgh)(bf)(agh)>

Cand	Sup
<a>	4
<b>	4
<c>	1
<d>	2
<e>	1
<f>	4
<g>	1
<h>	1

https://www.slideshare.net/Krish_ver2/53-mining-sequential-patterns

UCLA

GSP

Seq. ID	Sequence
1	<(cd)(abc)(abf)(acdf)>
2	<(abf)(e)>
3	<(abf)>
4	<(dgh)(bf)(agh)>

Cand	Sup
<a>	4
<b>	4
<d>	2
<f>	4

Length 2 Candidates generated by join

Length 2 Frequent Sequences

https://www.slideshare.net/Krish_ver2/53-mining-sequential-patterns

GSP

Seq. ID	Sequence
1	<(cd)(abc)(abf)(acdf)>
2	<(abf)(e)>
3	<(abf)>
4	<(dgh)(bf)(agh)>

Cand	Sup
<a>	4
<b>	4
<d>	2
<f>	4

Length 2 Candidates generated by join

<aa> <ab> <ad> <af> <ba> <bb> <bd> <bf>
<da> <db><dd> <df> <fa> <fb> <fd> <ff>
<(ab)> <(ad)> <(af)> <(bd)> <(bf)> <(df)>

Length 2 Frequent Sequences

https://www.slideshare.net/Krish_ver2/53-mining-sequential-patterns

GSP

Seq. ID	Sequence
1	<(cd)(abc)(abf)(acdf)>
2	<(abf)(e)>
3	<(abf)>
4	<(dgh)(bf)(agh)>

Cand	Sup
<a>	4
<b>	4
<d>	2
<f>	4

Length 2 Candidates generated by join

<aa> <ab> <ad> <af> <ba> <bb> <bd> <bf>
<da> <db><dd> <df> <fa> <fb> <fd> <ff>
<(ab)> <(ad)> <(af)> <(bd)> <(bf)> <(df)>

Length 2 Frequent Sequences

<ba> <da> <db> <df> <fa>
<(ab)> <(af)> <(bf)>

https://www.slideshare.net/Krish_ver2/53-mining-sequential-patterns

UCLA

GSP

Length 2 Frequent Sequences

<ba> <da> <db> <df> <fa>
<(ab)> <(af)> <(bf)>

Length 3 Candidates generated by join

Seq. ID	Sequence
1	<(cd)(abc)(abf)(acdf)>
2	<(abf)(e)>
3	<(abf)>
4	<(dgh)(bf)(agh)>

Length 3 Frequent Sequences

Length 4 Candidates generated by join

https://www.slideshare.net/Krish_ver2/53-mining-sequential-patterns

UCLA

GSP

Seq. ID	Sequence
1	<(cd)(abc)(abf)(acdf)>
2	<(abf)(e)>
3	<(abf)>
4	<(dgh)(bf)(agh)>

Length 2 Frequent Sequences

<ba> <da> <db> <df> <fa>
<(ab)> <(af)> <(bf)>

Length 3 Candidates generated by join

<ba> and <(ab)> - <b(ab)> {1}
<ba> and <(af)> - <b(af)> {1}
<da> and <(ab)> - <d(ab)> {1}
<da> and <(af)> - <d(af)> {1}
<db> and <(bf)> - <d(bf)> {1, 4}
<db> and <ba> - <dba> {1, 4}
<df> and <fa> - <dfa> {1, 4}
<fa> and <(ab)> - <f(ab)> -
<fa> and <(af)> - <f(af)> {1}
<(ab)> and <(bf)> - <(abf)> {1,2,3}
<(ab)> and <ba> - <(ab)a> {1}
<(af)> and <fa> - <(af)a) {1}
<(bf)> and <fa> - <(bf)a> {1, 4}

Length 3 Frequent Sequences

Length 4 Candidates generated by join

https://www.slideshare.net/Krish_ver2/53-mining-sequential-patterns

GSP

Seq. ID	Sequence
1	<(cd)(abc)(abf)(acdf)>
2	<(abf)(e)>
3	<(abf)>
4	<(dgh)(bf)(agh)>

Length 2 Frequent Sequences

<ba> <da> <db> <df> <fa>
<(ab)> <(af)> <(bf)>

Length 3 Candidates generated by join

<ba> and <(ab)> - <b(ab)> {1}
<ba> and <(af)> - <b(af)> {1}
<da> and <(ab)> - <d(ab)> {1}
<da> and <(af)> - <d(af)> {1}
<db> and <(bf)> - <d(bf)> {1, 4}
<db> and <ba> - <dba> {1, 4}
<df> and <fa> - <dfa> {1, 4}
<fa> and <(ab)> - <f(ab)> -
<fa> and <(af)> - <f(af)> {1}
<(ab)> and <(bf)> - <(abf)> {1,2,3}
<(ab)> and <ba> - <(ab)a> {1}
<(af)> and <fa> - <(af)a) {1}
<(bf)> and <fa> - <(bf)a> {1, 4}

Length 3 Frequent Sequences

<dba> <dfa> <(abf)> <(bf)a> <d(bf)>

Length 4 Candidates generated by join

https://www.slideshare.net/Krish_ver2/53-mining-sequential-patterns

GSP

Seq. ID	Sequence
1	<(cd)(abc)(abf)(acdf)>
2	<(abf)(e)>
3	<(abf)>
4	<(dgh)(bf)(agh)>

Length 2 Frequent Sequences

<ba> <da> <db> <df> <fa>
<(ab)> <(af)> <(bf)>

Length 3 Candidates generated by join

<ba> and <(ab)> - <b(ab)> {1}
<ba> and <(af)> - <b(af)> {1}
<da> and <(ab)> - <d(ab)> {1}
<da> and <(af)> - <d(af)> {1}
<db> and <(bf)> - <d(bf)> {1, 4}
<db> and <ba> - <dba> {1, 4}
<df> and <fa> - <dfa> {1, 4}
<fa> and <(ab)> - <f(ab)> -
<fa> and <(af)> - <f(af)> {1}
<(ab)> and <(bf)> - <(abf)> {1,2,3}
<(ab)> and <ba> - <(ab)a> {1}
<(af)> and <fa> - <(af)a) {1}
<(bf)> and <fa> - <(bf)a> {1, 4}

Length 3 Frequent Sequences

<dba> <dfa> <(abf)> <(bf)a> <d(bf)>

Length 4 Candidates generated by join

<d(bf)> and <(bf)a> - <d(bf)a> {1, 4}
<(abf)> and <(bf)a> - <(abf)a> {1}

https://www.slideshare.net/Krish_ver2/53-mining-sequential-patterns

PrefixSpan

- 1. Find length1sequential patterns
  - min_support = 2

id	Sequence
10	<a(abc)(ac)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(ab)(df)cb>
40	<eg(af)cbc>

Find length-1 sequential patterns

https://fenix.tecnico.ulisboa.pt/downloadFile/1407993358847907/licao_8.pdf

PrefixSpan

- 1. Find length1sequential patterns
  - min_support = 2

id	Sequence
10	<a(abc)(ac)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(ab)(df)cb>
40	<eg(af)cbc>

Find length-1 sequential patterns

<a>	<b>	<c>	<d>	<e>	<f>	~~<g>~~
4	4	4	3	3	3	~~1~~

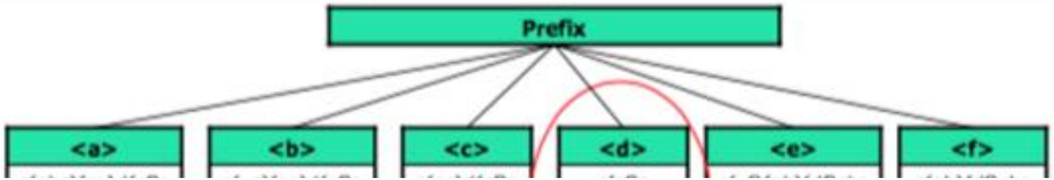
<a><b><c><d><e><f>

https://fenix.tecnico.ulisboa.pt/downloadFile/1407993358847907/licao_8.pdf

PrefixSpan

- 2. Divide search space

id	Sequence
10	<a(abc)(ac)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(ab)(df)cb>
40	<eg(af)cbc>

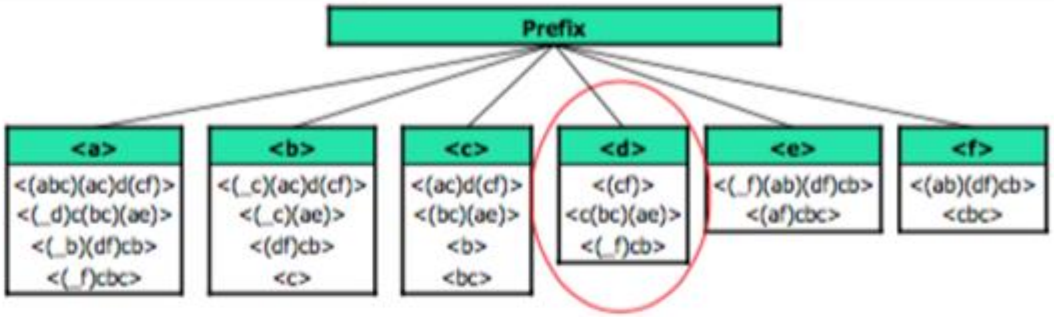


https://fenix.tecnico.ulisboa.pt/downloadFile/1407993358847907/licao_8.pdf

PrefixSpan

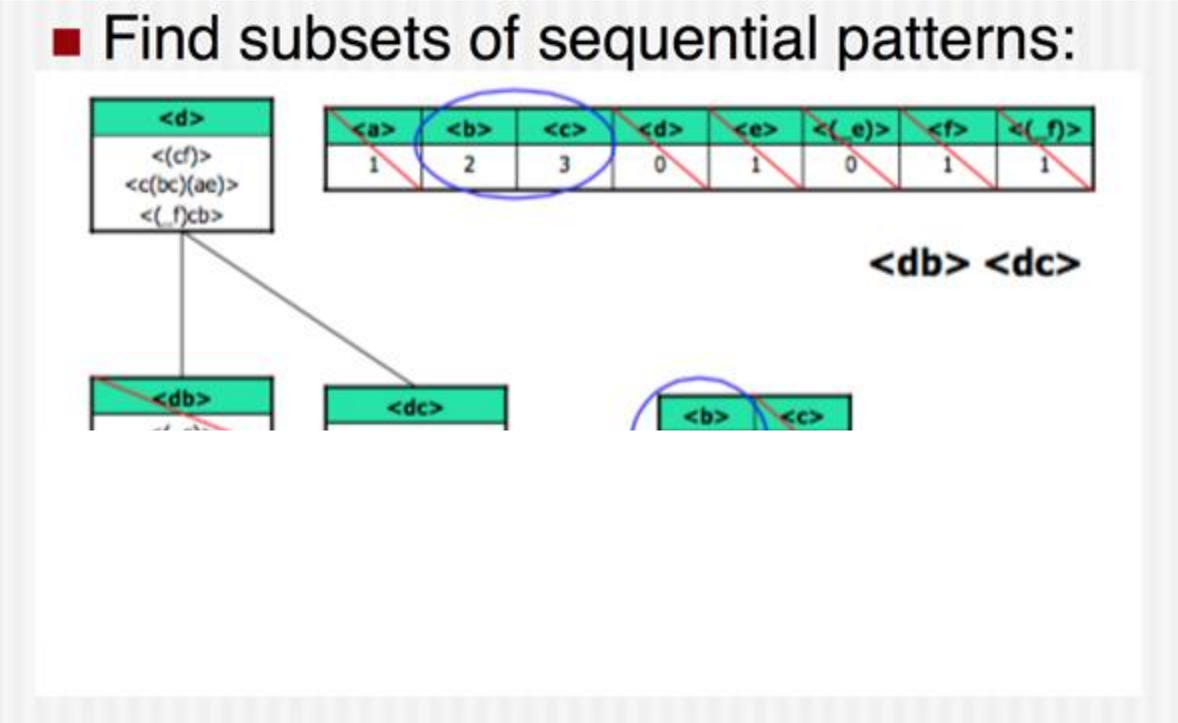
- 2. Divide search space

id	Sequence
10	<a(abc)(ac)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(ab)(df)cb>
40	<eg(af)cbc>



https://fenix.tecnico.ulisboa.pt/downloadFile/1407993358847907/licao_8.pdf

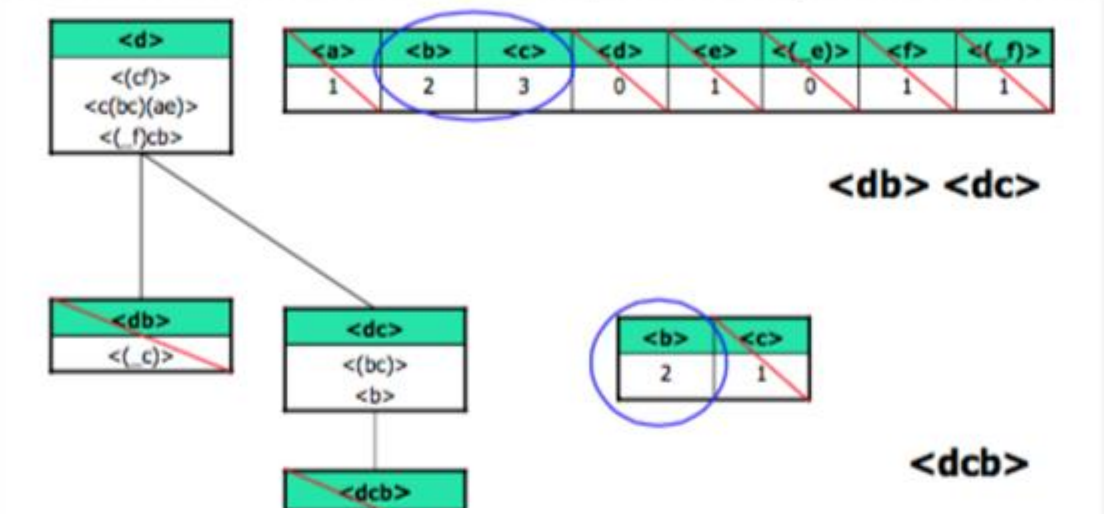
PrefixSpan



https://fenix.tecnico.ulisboa.pt/downloadFile/1407993358847907/licao_8.pdf

PrefixSpan

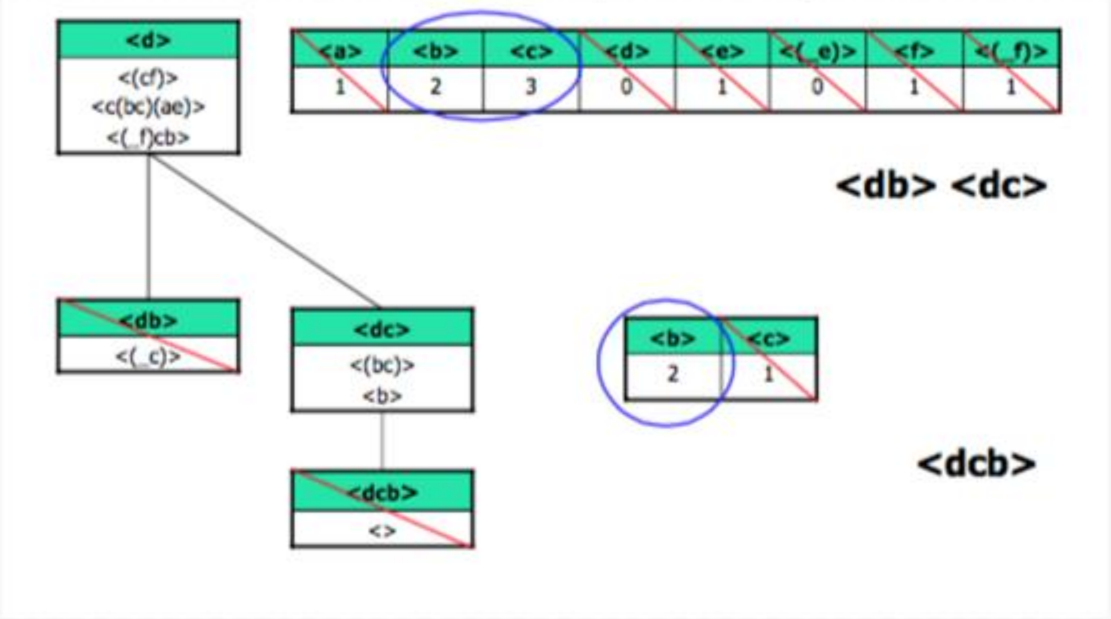
- Find subsets of sequential patterns:



https://fenix.tecnico.ulisboa.pt/downloadFile/1407993358847907/licao_8.pdf

PrefixSpan

- Find subsets of sequential patterns:



https://fenix.tecnico.ulisboa.pt/downloadFile/1407993358847907/licao_8.pdf

Thank you!

Q & A