Tato prezentace je součástí wiki-prezentace Metoda GUHA, LISp-Miner a typové úlohy
Je dostupná z této adresy
Verse 19. 8. 2019

Typ úlohy: Asociační pravidla - výjimky z histogramu

Data: Insolvency

Problém: Hledáme pravidla $\varphi \rightarrow$ Days_to_start(a_i) taková, že jejich konfidence je alespoň čtyřikrát nižší, než odpovídá výšce sloupce a_i v histogramu na celé matici dat

Jan Rauch

Katedra informačního a znalostního inženýrství

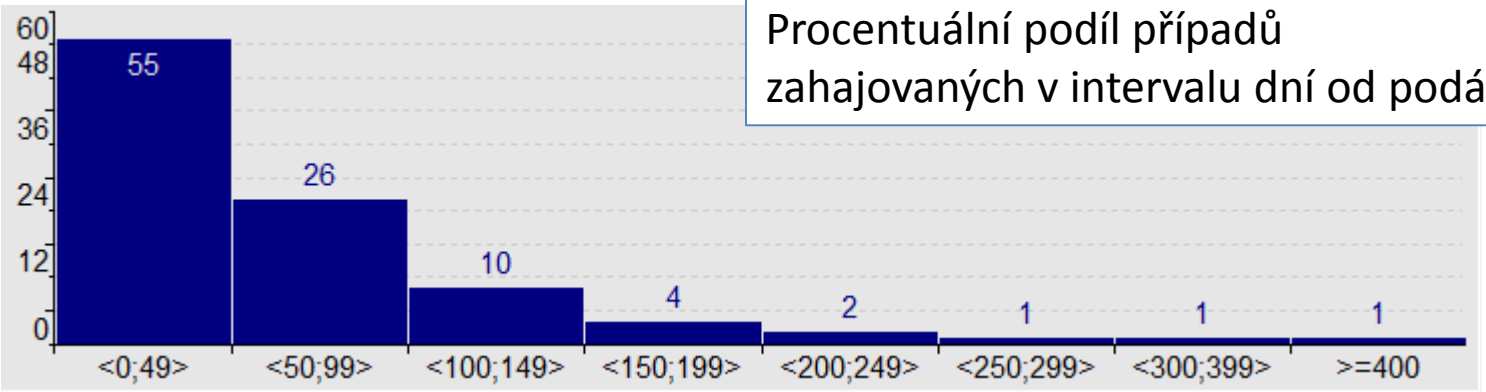
Vysoká škola ekonomická v Praze

© Jan Rauch

Data o insolvencích a exception rules

- Příklady
- Asociační pravidlo jako výjimka k histogramu
- Asociační pravidla - výjimky z obecného histogramu

Days_to_start - počet dní do zahájení řízení



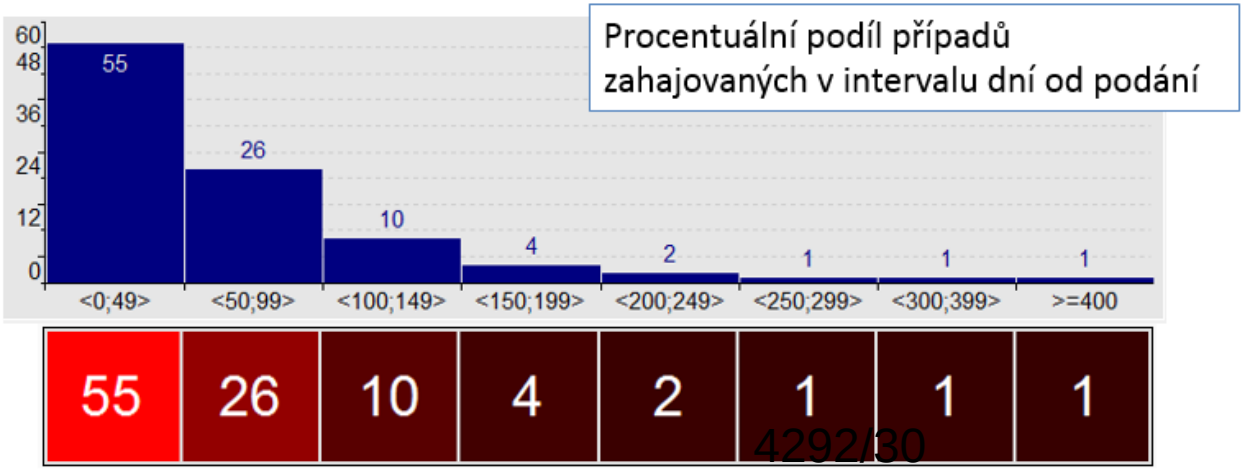
Procentuální podíl případů zahajovaných v intervalu dní od podání

55	26	10	4	2	1	1	1

Za jakých okolností se výrazně liší procentuální podíl případů pro některý z intervalů?

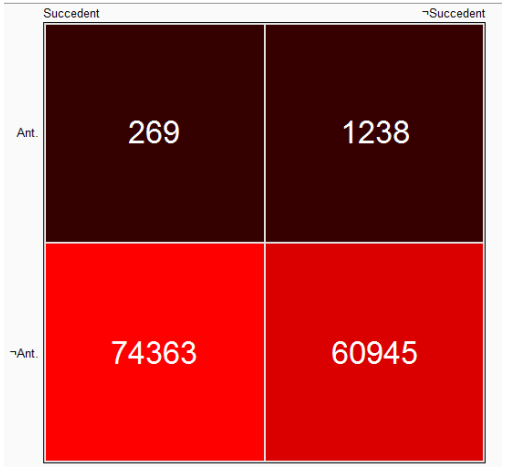
Je alespoň x-krát menší/větší ?

Days_to_start - počet dní do zahájení řízení



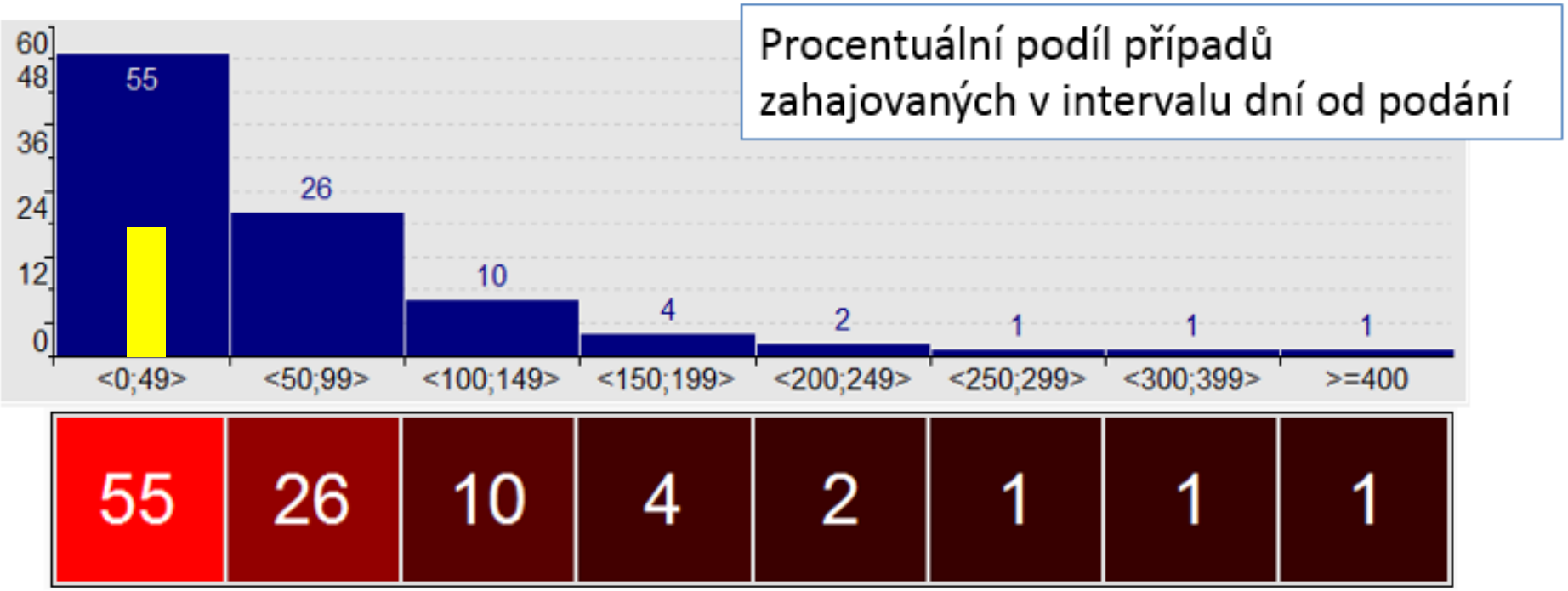
Region(Jihomoravský) $\wedge$ proposal.Year(2012) $\rightarrow$ Days_to_start$\langle 0;49 \rangle$

Konfidence = 269/(269+1238) = 0.18 t.j.
3.06 x méně, než odpovídá sloupci histogramu kde je 55%



Region(Jihomoravský) $\wedge$ proposal.Year(2012) $\rightarrow$ Days_to_start$\langle 0;49 \rangle$
je exception rule vzhledem k histogramu

Days_to_start - počet dní do zahájení řízení

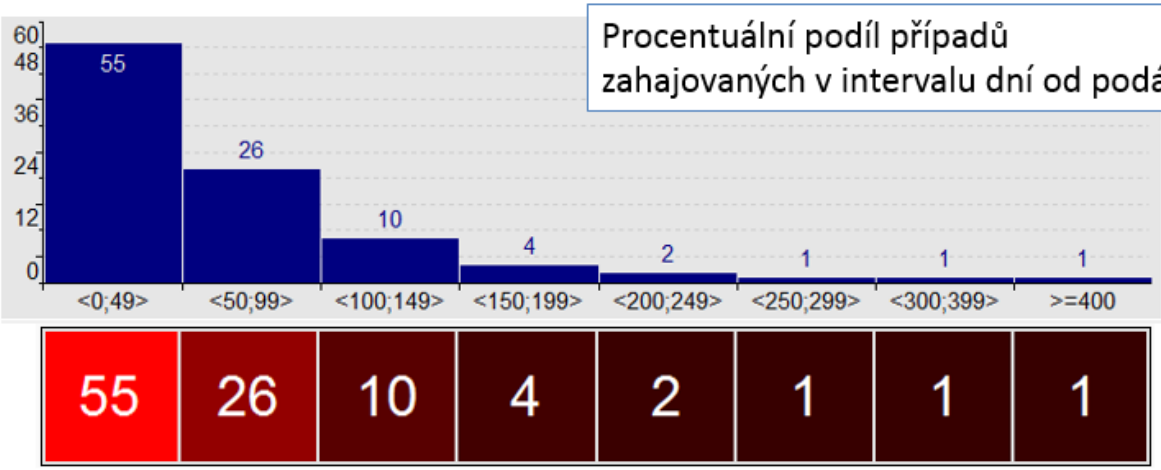


Region(Jihomoravský) $\wedge$ proposal.Year(2012) $\rightarrow$ Days_to_start$\langle 0;49 \rangle$

Konfidence = 269/(269+1238) = 0.18 t.j.

3.06 x méně, než odpovídá sloupci histogramu kde je 55%

Days_to_start - počet dní do zahájení řízení



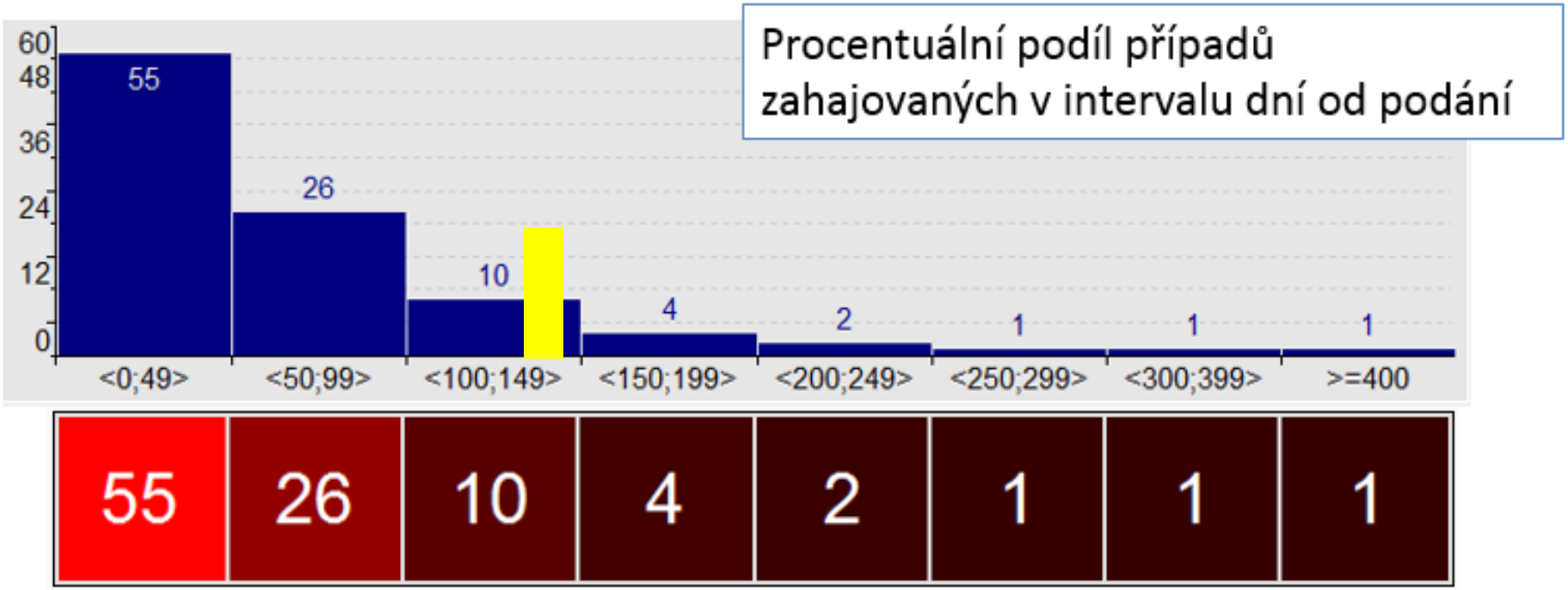
Administrator_name(IREKON, v.o.s.) $\wedge$ proposal.Year(2012) $\rightarrow$ Days_to_start$\langle 100;149 \rangle$

Konfidence = 71/(71+239) = 0.23 t.j.
2.27 x více, než odpovídá sloupci histogramu kde je 10%

	Succedent	¬Succedent
Ant.	71	239
¬Ant.	13868	122637

Administrator_name(IREKON, v.o.s.) $\wedge$ proposal.Year(2012) $\rightarrow$ Days_to_start$\langle 100;149 \rangle$
je exception rule vzhledem k histogramu

Days_to_start - počet dní do zahájení řízení



Administrator_name(IREKON, v.o.s.) $\wedge$ proposal.Year(2012) $\rightarrow$ Days_to_start$\langle$100;149$\rangle$

Konfidence = 71/(71+239) = 0.23 t.j.
2.27 x více, než odpovídá sloupci histogramu kde je 10%

Asociační pravidlo jako výjimka k histogramu

Užitečné pojmy zavedené pro vyjádření více faktů než je třeba pro příklady.

- Histogram jako vektor
- Míry odlišnosti histogramu a asociačního pravidla (Exception measures)
- Asociační pravidlo je výjimkou k histogramu

Histogram jako vektor (1)

Histogram H atributu A typu *ftype* je výraz H(A, χ, *ftype*) kde

- A je atribut s kategoriemi a_1,…, a_t
- χ je výraz *All* nebo booleovský atribut
- *ftype* je jeden z výrazů *Abs*, *Rel*, *Cat*

H(A, *All*, *ftype*) je obecný histogram typu *ftype pro* atribut A

H(A, χ, *ftype*) je podmíněný histogram typu *ftype pro* atribut A a podmínku χ

CF-tabulka CF(A, χ, $\mathcal{M}$) pro atribut A, podmínku χ a matici dat $\mathcal{M}$ (stručná):

Data matrix	a_1	$\dots$	a_t	Σ
$\mathcal{M}/\chi$	m_1	$\dots$	m_t	m
$\mathcal{M}$	n_1	$\dots$	n_t	n

- m_i je počet řádků matice dat $\mathcal{M}$ splňujících χ a A(a_i) pro i = 1, …, t
- n_i je počet řádků matice dat $\mathcal{M}$ splňujících χ a A(a_i) pro i = 1, …, t

Histogram jako vektor (2)

Realizace H($\mathcal{M}$, A, χ, *ftype*) histogramu H(A, χ, *ftype*) v matici dat $\mathcal{M}$ je vektor $\langle w_1, \ldots, w_t \rangle$ pro který platí:

- $w_i = m_i$ pro i = 1, …, t pokud *ftype = Abs*
- $w_i = 100 * m_i / m$ pro i = 1, …, t pokud *ftype = Rel*
- $w_i = 100 * m_i / n_i$ pro i = 1, …, t pokud *ftype = Cat.*

Data matrix	a_1	$\ldots$	a_t	Σ
$\mathcal{M}/\chi$	m_1	$\ldots$	m_t	m
$\mathcal{M}$	n_1	$\ldots$	n_t	n

Jestliže χ = All, pak $m_i = n_i$ pro i = 1, …, t.

Histogram jako vektor - příklad

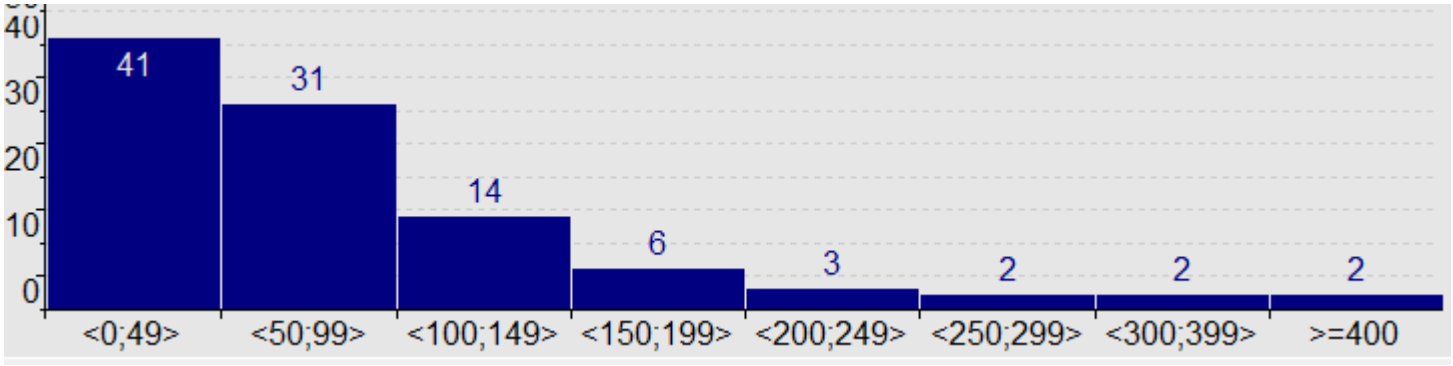
CF-tabulka CF(Days_to_start, Region(Praha), Insolvency)

	<0;49>	<50;99>	<100;149>	<150;199>	<200;249>	<250;299>	<300;399>	>=400	Σ
Region(Praha)	3535	2691	1225	516	262	152	168	155	8 704

H(Insolvency, Days_to_start, Region(Praha), *Rel*)

41	31	14	6	3	2	2	2

41 = 100 * 3535/8704, …



Porovnání histogramu a asociačního pravidla

Data matrix	a_1	$\dots$	a_t	Σ
$\mathcal{M}/\chi$	m_1	$\dots$	m_t	m
$\mathcal{M}$	n_1	$\dots$	n_t	n

CF-table $CF(A, \chi, \mathcal{M})$

$\mathcal{M}/\chi$	$A(a_i)$	$\neg A(a_i)$
φ	a	b
$\neg\varphi$	c	d

4ft-table $4ft(\varphi, A(a_i), \mathcal{M}/\chi)$

Porovnáváme realizaci podmíněného histogramu H($\mathcal{M}$, A, χ, *Rel*) atributu A v matici dat $\mathcal{M}$ a podmíněné asociační pravidlo $\varphi \rightarrow A(a_i)/\chi$ v matici dat $\mathcal{M}$

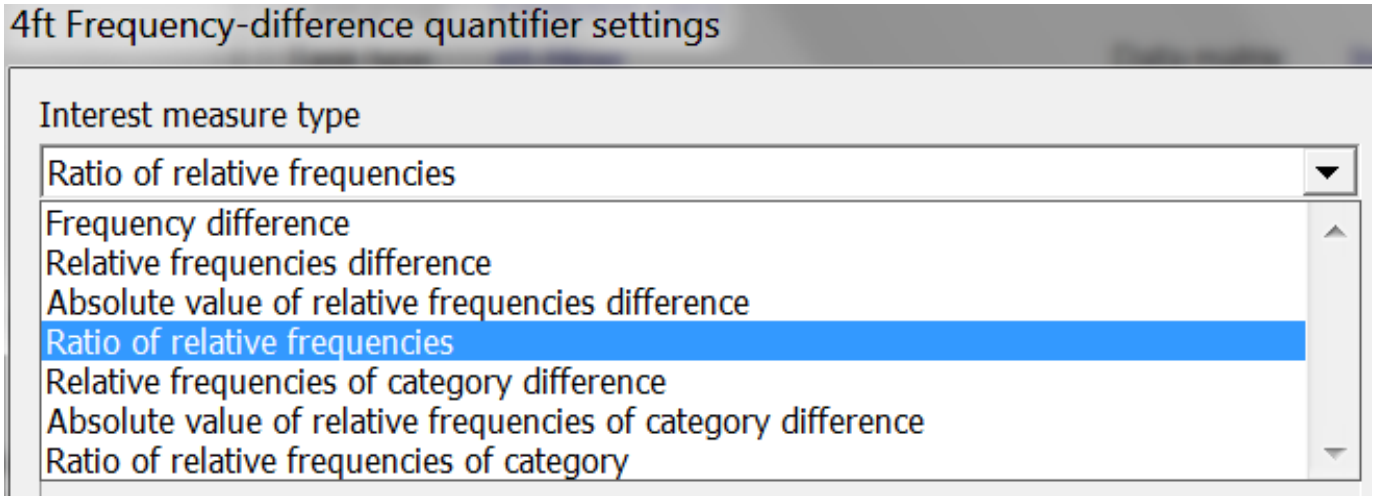
$\varphi \rightarrow A(a_i)/\chi$ je výjimkou vzhledem k H($\mathcal{M}$, A, χ, *Rel*) pokud se významně liší vhodné charakteristiky i-tého sloupce a asociačního pravidla.

Výška sloupce odpovídajícího kategorii a_i je $100 * m_i / m$

Výška sloupce odpovídajícího kategorii a_i pokud krom χ navíc platí i φ odpovídá konfidenci podmíněného asociačního pravidla $\varphi \rightarrow A(a_i)/\chi$, tedy

Míry odlišnosti histogramu a asociačního pravidla

#	Name	symbol	definition
1	Frequency difference	$FDVal$	$m_i - a$
2	Relative frequencies difference	$FDRel$	$\frac{m_i}{m} - \frac{a}{a+b}$
3	Absolute value of relative frequencies difference	$FDRelA$	$\lvert\frac{m_i}{m} - \frac{a}{a+b}\rvert$
4	Ratio of relative frequencies	$FDRelR$	$\frac{\frac{m_i}{m}}{\frac{a}{a+b}}$
5	Relative frequencies of category difference	$FDCat$	$\frac{m_i}{n_i} - \frac{a}{a+c}$
6	Absolute value of relative frequencies of category difference	$FDCatA$	$\lvert\frac{m_i}{n_i} - \frac{a}{a+c}\rvert$
7	Ratio of relative frequencies of category	$FDCatR$	$\frac{\frac{m_i}{n_i}}{\frac{a}{a+c}}$



Poměr relativních frekvencí *FDRelR* = $\frac{\frac{m_i}{m}}{\frac{a}{a+b}}$

Data matrix	a_1	...	a_t	Σ
$\mathcal{M}/\chi$	m_1	...	m_t	m
$\mathcal{M}$	n_1	...	n_t	n

CF-table $CF(A, \chi, \mathcal{M})$

$\mathcal{M}/\chi$	$A(a_i)$	$\neg A(a_i)$
φ	a	b
$\neg\varphi$	c	d

4ft-table $4ft(\varphi, A(a_i), \mathcal{M}/\chi)$

Pokud $\frac{\frac{m_i}{m}}{\frac{a}{a+b}} = X > 1$, pak:

- výška sloupce v histogramu je X-krát větší než odpovídá konfidenci pravidla $\varphi \rightarrow A(a_i)/\chi$
- konfidence pravidla $\varphi \rightarrow A(a_i)/\chi$ je X-krát nižší, než odpovídá výšce sloupce v histogramu

$$\frac{\frac{m_i}{m}}{\frac{a}{a+b}} = 3.06$$

$$\frac{a}{a+b} = 0.18$$

Poměr relativních frekvencí *FDRelR* = $\dfrac{\frac{m_i}{m}}{\frac{a}{a+b}}$

Data matrix	a_1	...	a_t	Σ
$\mathcal{M}/\chi$	m_1	...	m_t	m
$\mathcal{M}$	n_1	...	n_t	n

CF-table $CF(A, \chi, \mathcal{M})$

$\mathcal{M}/\chi$	$A(a_i)$	$\neg A(a_i)$
φ	a	b
$\neg\varphi$	c	d

4ft-table $4ft(\varphi, A(a_i), \mathcal{M}/\chi)$

Pokud $\dfrac{\frac{m_i}{m}}{\frac{a}{a+b}} = X < 1$, pak vhodnější $\dfrac{\frac{a}{a+b}}{\frac{m_i}{m}} = Y > 1$:

- výška sloupce v histogramu je Y-krát menší, než odpovídá konfidenci pravidla $\varphi \rightarrow A(a_i)/\chi$
- konfidence pravidla $\varphi \rightarrow A(a_i)/\chi$ je Y-krát vyšší, než odpovídá výšce sloupce v histogramu

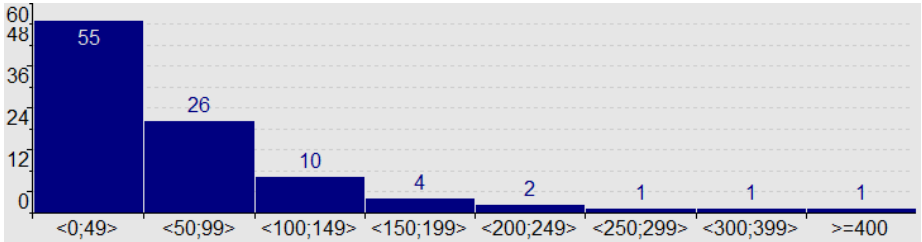
$$\frac{\frac{a}{a+b}}{\frac{m_i}{m}} = 2.27$$

$$\frac{a}{a+b} = 0.23$$

$$\frac{\frac{m_i}{m}}{\frac{a}{a+b}} = 0.44$$

Asociační pravidla - výjimky z obecného histogramu (1)

H(Insolvency, Days_to_start, All, *Rel*)



55	26	10	4	2	1	1	1

Hledáme asociační pravidla
- výjimky k obecnému histogramu

Data matrix	a_1	...	a_t	Σ
$\mathcal{M}/\chi$	m_1	...	m_t	m
$\mathcal{M}$	n_1	...	n_t	n

CF-table $CF(A, \chi, \mathcal{M})$

$\mathcal{M}/\chi$	$A(a_i)$	$\neg A(a_i)$
φ	a	b
$\neg\varphi$	c	d

4ft-table $4ft(\varphi, A(a_i), \mathcal{M}/\chi)$

Hledáme pravidlo φ → Days_to_start(a_i) takové, že jeho konfidence je alespoň čtyřikrát nižší, než odpovídá výšce sloupce v histogramu.

$$FDRelR = \frac{\frac{m_i}{m}}{\frac{a}{a+b}} \geq 4$$

Asociační pravidla - výjimky z obecného histogramu (2)

$\varphi \rightarrow$ Days_to_start(a_i)

ANTECEDENT

Administrator	Con, 0 - 1
» Administrator_name (subset), 1 - 1	B, pos
» Administrator_type_id (subset), 1 - 1	B, pos
Debtor	Con, 0 - 4
» Age_10 (subset), 1 - 1	B, pos
» Gender (subset), 1 - 1	B, pos
» Has_ico (subset), 1 - 1	B, pos
» Person_type (subset), 1 - 1	B, pos
» Region (subset), 1 - 1	B, pos
Proposal	Con, 0 - 2
» proposal.Quarter (subset), 1 - 1	R, pos
» proposal.Year (subset), 1 - 1	B, pos

φ

Total length: 1 - 7

QUANTIFIERS

BASE p= 90 Abs.
FDRε >= 4.00 Abs

Generation information

Status: Solved, 6 run(s)
Mode: Standard

SUCCEDENT

Default Partial Cedent	Con, 1 - 5
» Days_to_start_ed50m (subset), 1 - 1	B, pos

Days_to_start

Total length: 1

Task parameters

Handling of missing values: Ignore X-categories
Prime rule test for implications enabled: No
Include succedent extensions of 100% implications: Yes
Include extensions of coefficients with no change in the four-fold table: Yes
Include extensions of cedents with no change in the four-fold table: Yes
Include 'worse' antecedent extensions (for implications and AAD/BAD): Yes
Include both symetric hypotheses: Yes Extensions minimal length check: Yes
Maximal number of hypotheses: 1000

Params | Switch | Validate | Task Clone
Run | Bkgrnd Run | Grid Run | Show Results

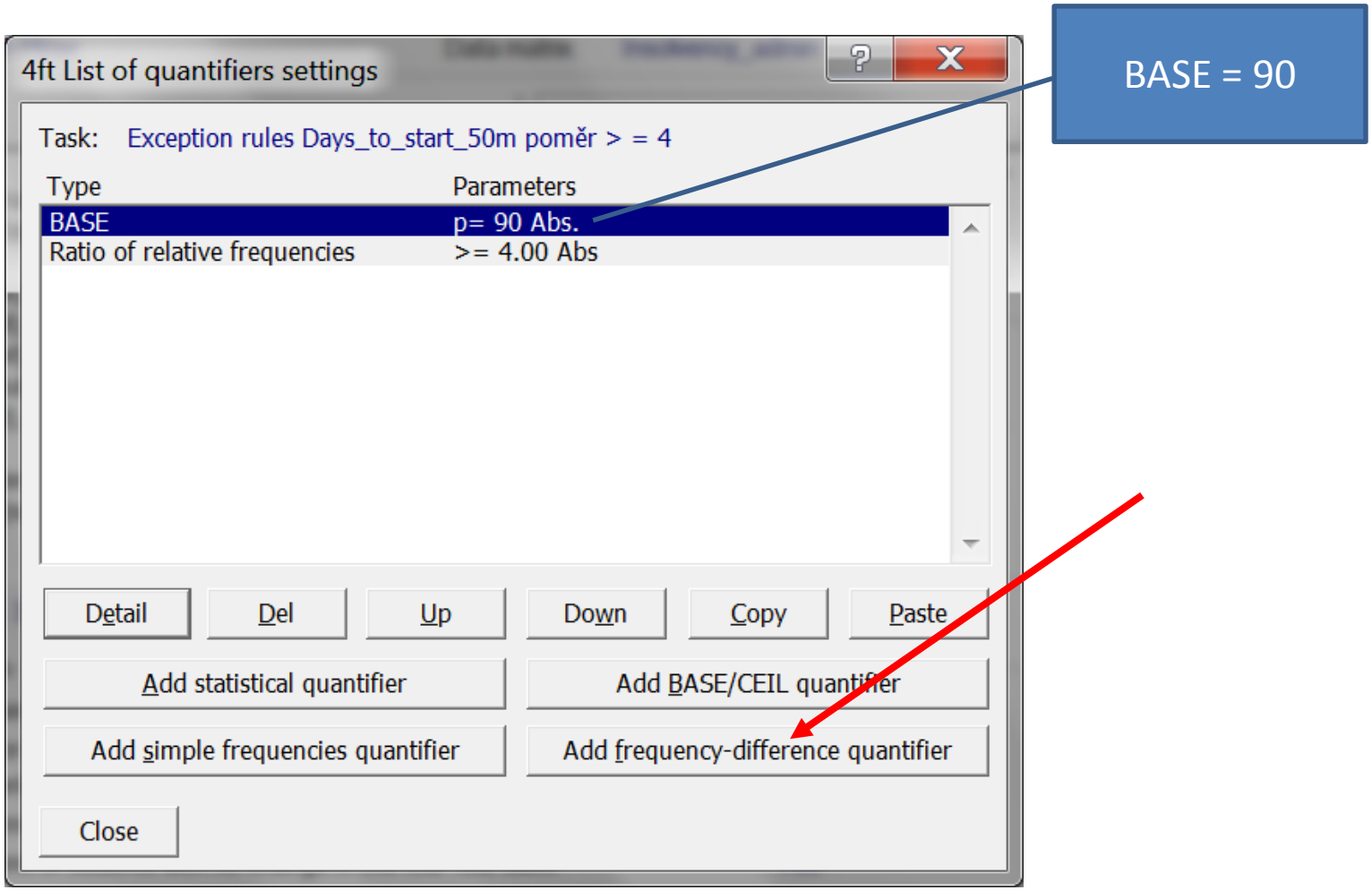
CONDITION

Default Partial Cedent	Con, 0 - 5

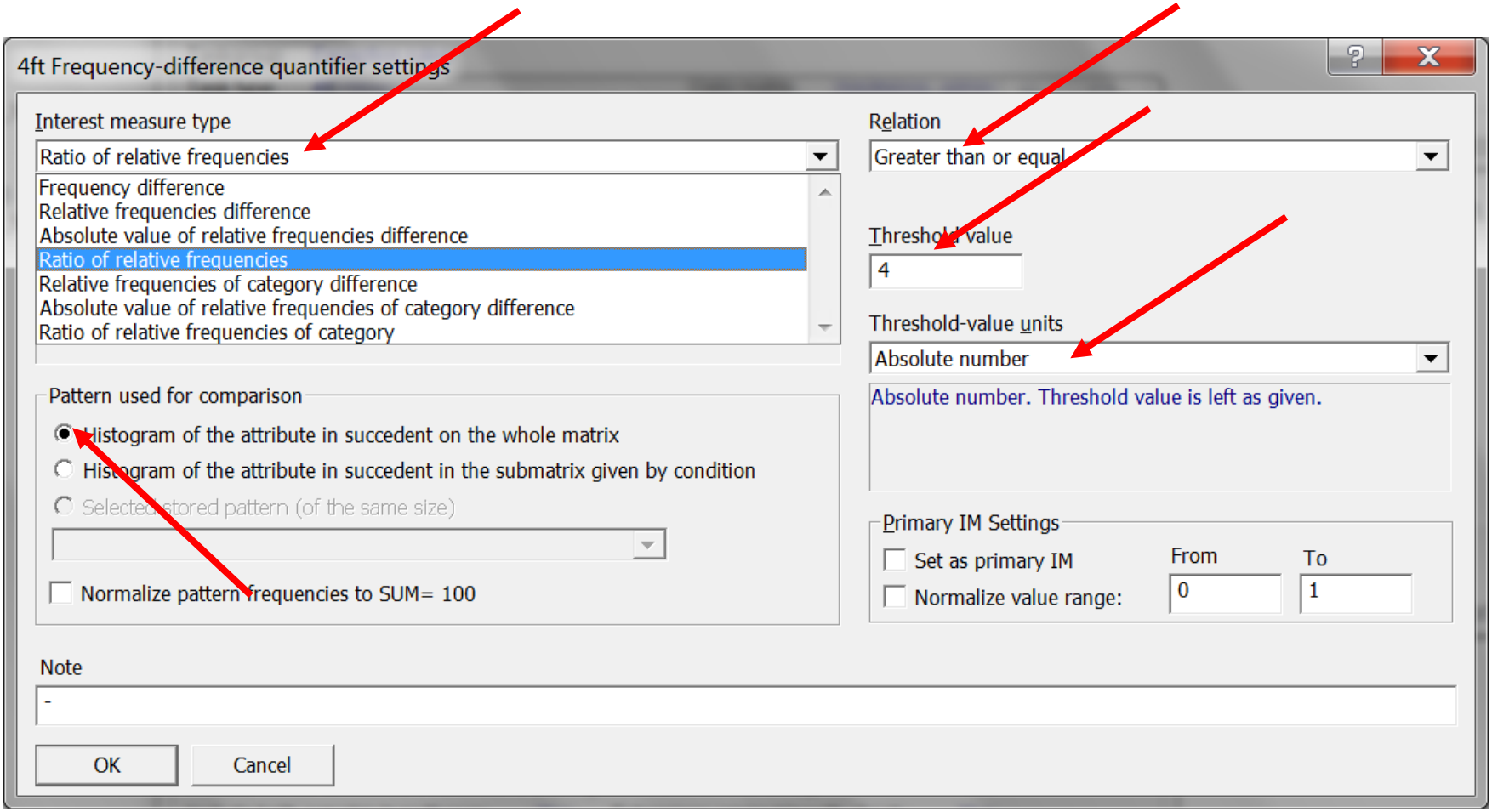
Podmínka nezadána

Total length: 0

Asociační pravidla - výjimky z obecného histogramu (3)



Asociační pravidla - výjimky z obecného histogramu (4)



Asociační pravidla - výjimky z obecného histogramu (5)

Task run

Start: 11.11.2018 12:00:17 Total time: 0h 0m 25s

Number of verifications: 381584

Number of hypotheses: 16 Mode: Standard

Add group | Del group | Edit group

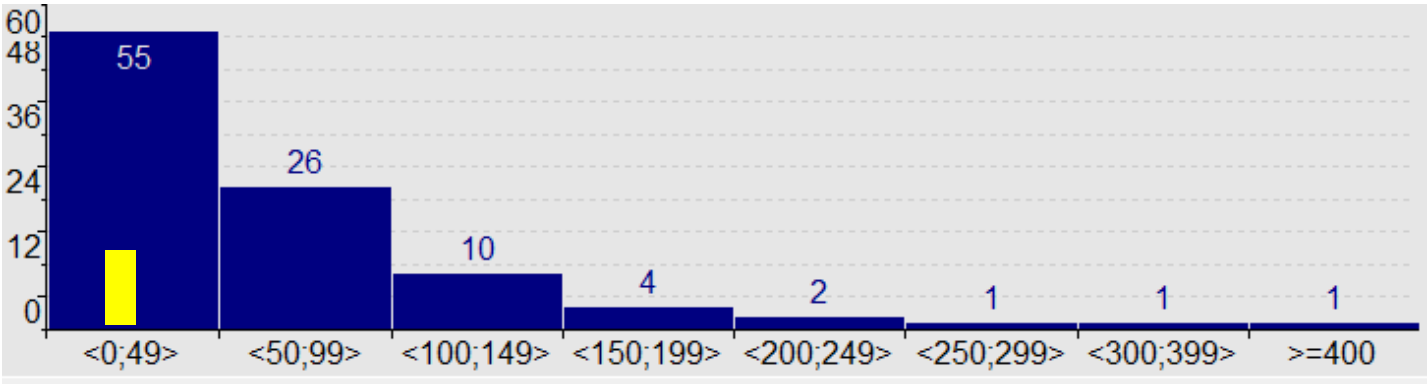
Actual group of hypotheses: All hypotheses

Hypotheses in group: 16 Shown hypotheses: 16 Highlighted: 0

Delete hypo

Nr.	Id	FDRelRatio	Hypothesis
1	15	5.839	Administrator_type_id(*1*) & Region(*Pardubický*) & proposal.Year(*2012*) >÷< Days_to_start(<=*49*)
2	7	5.839	Region(*Pardubický*) & proposal.Year(*2012*) >÷< Days_to_start(<=*49*)
3	16	5.686	Administrator_type_id(*1*) & Person_type(*N*) & Region(*Pardubický*) & proposal.Year(*2012*) >÷< Days_to_start(<=*49*)
4	8	5.686	Person_type(*N*) & Region(*Pardubický*) & proposal.Year(*2012*) >÷< Days_to_start(<=*49*)
5	11	4.328	Administrator_type_id(*1*) & Has_ico(*No*) & Person_type(*N*) & proposal.Quarter(*4*) & proposal.Year(*2016*) >÷< Days_to_start(*<50;99>*)
6	3	4.328	Has_ico(*No*) & Person_type(*N*) & proposal.Quarter(*4*) & proposal.Year(*2016*) >÷< Days_to_start(*<50;99>*)
7	9	4.313	Administrator_type_id(*1*) & Has_ico(*No*) & proposal.Quarter(*4*) & proposal.Year(*2016*) >÷< Days_to_start(*<50;99>*)
8	1	4.313	Has_ico(*No*) & proposal.Quarter(*4*) & proposal.Year(*2016*) >÷< Days_to_start(*<50;99>*)
9	2	4.235	Has_ico(*No*) & proposal.Year(*2016*) >÷< Days_to_start(*<150;199>*)
10	10	4.235	Administrator_type_id(*1*) & Has_ico(*No*) & proposal.Year(*2016*) >÷< Days_to_start(*<150;199>*)
11	4	4.222	Has_ico(*No*) & Person_type(*N*) & proposal.Year(*2016*) >÷< Days_to_start(*<150;199>*)
12	12	4.222	Administrator_type_id(*1*) & Has_ico(*No*) & Person_type(*N*) & proposal.Year(*2016*) >÷< Days_to_start(*<150;199>*)
13	14	4.098	Administrator_type_id(*1*) & Gender(*M*) & Person_type(*N*) & proposal.Quarter(*4*) & proposal.Year(*2016*) >÷< Days_to_start(*<50;99>*)
14	6	4.098	Gender(*M*) & Person_type(*N*) & proposal.Quarter(*4*) & proposal.Year(*2016*) >÷< Days_to_start(*<50;99>*)
15	13	4.053	Administrator_type_id(*1*) & Gender(*M*) & proposal.Quarter(*4*) & proposal.Year(*2016*) >÷< Days_to_start(*<50;99>*)
16	5	4.053	Gender(*M*) & proposal.Quarter(*4*) & proposal.Year(*2016*) >÷< Days_to_start(*<50;99>*)

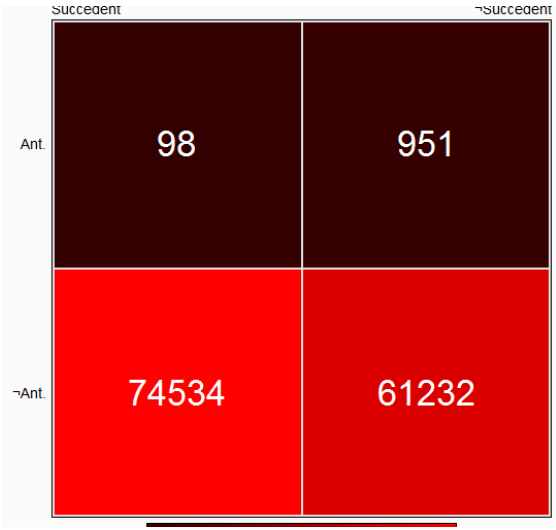
Asociační pravidla - výjimky z obecného histogramu (6)



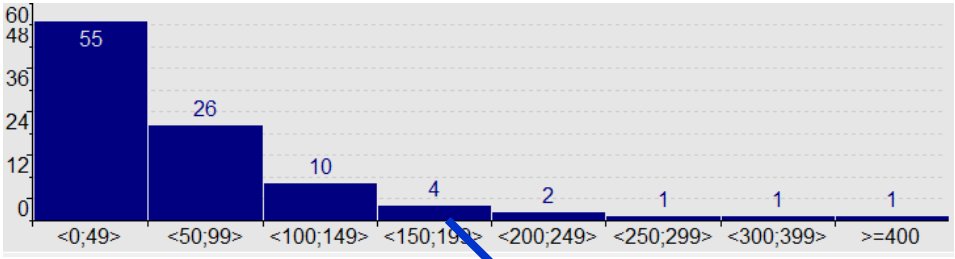
Region(Pardubický) $\wedge$ proposal.Year(2012) $\rightarrow$ Days_to_start$\langle 0;49 \rangle$

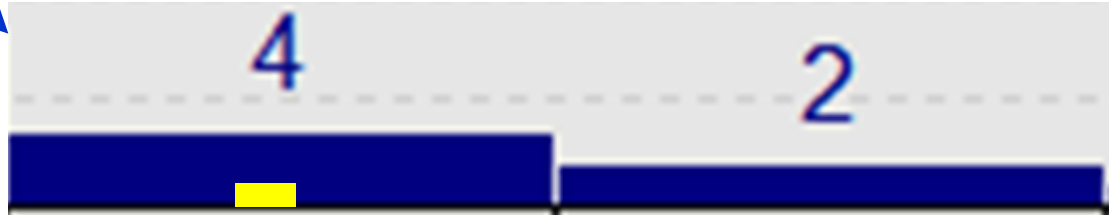
Konfidence = 98/(98+951) = 0.09 t.j.
5.84 x méně, než odpovídá sloupci histogramu kde je 55%

$$FDRelR = \frac{\frac{m_i}{m}}{\frac{a}{a+b}} = 5.84$$



Asociační pravidla - výjimky z obecného histogramu (7)





Has_ico(No) $\wedge$ proposal.Year(2016) $\rightarrow$ Days_to_start$\langle$150;199$\rangle$

Konfidence = 132/(132+13608) = 0.01 t.j.

4.24 x méně, než odpovídá sloupci histogramu kde jsou 4%

$$FDRelR = \frac{\frac{m_i}{m}}{\frac{a}{a+b}} = 4.24$$

	Succedent	¬Succedent
Ant.	132	13608
¬Ant.	5435	117640