Tato prezentace je součástí wiki-prezentace Metoda GUHA, LISp-Miner a typové úlohy
Je dostupná z této adresy
Verse 17. 8. 2019

Typ úlohy: Podíl podmínky na kategoriích
Data: Hotel
Problém: Existují podmínky a den týdne takové, že alespoň 50% ze všech hostů přijíždějících v tomto dni vyhovuje podmínkám?

Jan Rauch
Katedra informačního a znalostního inženýrství
Vysoká škola ekonomická v Praze

© Jan Rauch

Podíl podmínky na kategoriích

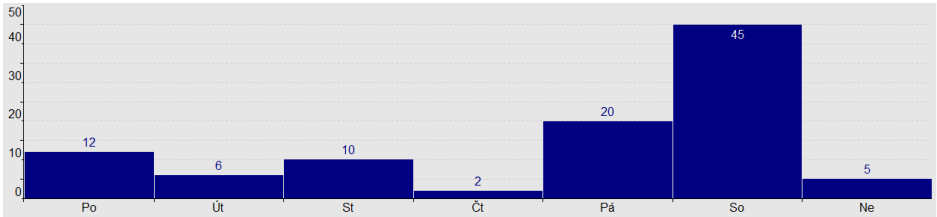
- Podíl kategorií na podmínce X Podíl podmínky na kategoriích
- Podíl podmínky na kategoriích - příklad

Podíl kategorií na podmínce X Podíl podmínky na kategoriích

Počet začínajících pobytů

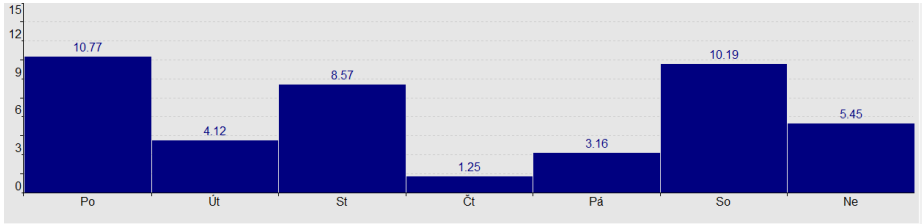
	Po	Út	St	Čt	Pá	So	Ne	
Drážďany	14	7	12	2	24	54	6	119
Celkem	130	170	140	160	760	530	110	2 000

Podíl kategorií (Po, …) na podmínce (Drážďany)



	Po	Út	St	Čt	Pá	So	Ne
dne	12	6	10	2	20	45	5

Podíl podmínky (Drážďany) na kategoriích (Po…)



	Po	Út	St	Čt	Pá	So	Ne
dne	10.77	4.12	8.57	1.25	3.16	10.19	5.45

12 = 100*14/**119**

45 = 100*54/**119**

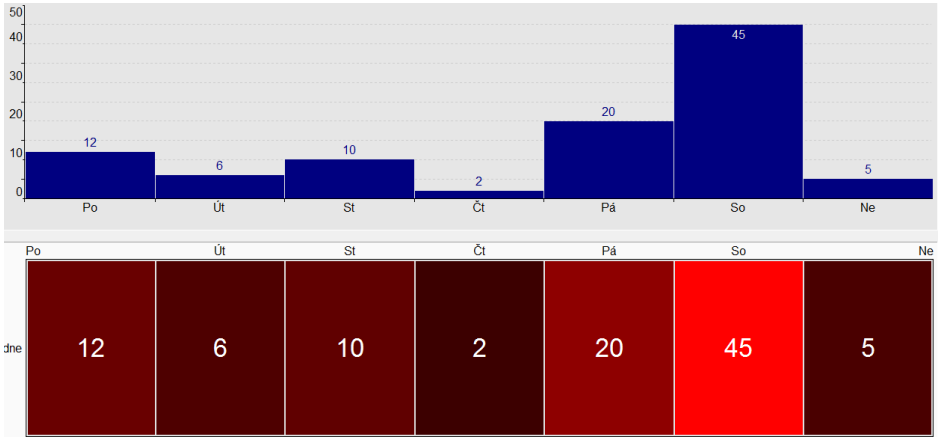
10.77 = 100*14/**130**

10.19 = 100*54/**530**

Podíl kategorií na podmínce

Počet začínajících pobytů

	Po	Út	St	Čt	Pá	So	Ne	
Drážďany	14	7	12	2	24	54	6	119
Celkem	130	170	140	160	760	530	110	2 000



12 = 100*14/**119**

45 = 100*54/**119**

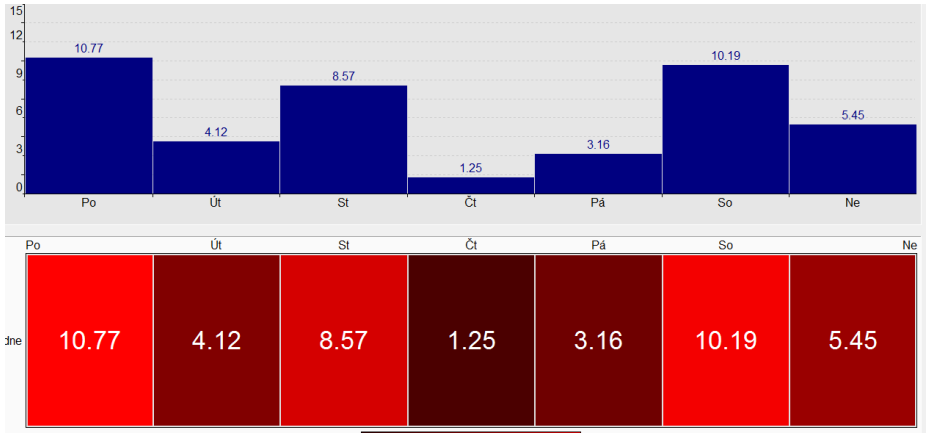
- 14 ze všech 119 pobytů hostů z Drážďan začíná v pondělí.
- 12 % všech pobytů hostů z Drážďan začíná v pondělí.

- 54 ze všech pobytů 119 pobytů hostů z Drážďan začíná v sobotu.
- 45 % všech pobytů hostů z Drážďan začíná v sobotu.

Podíl podmínky na kategoriích

Počet začínajících pobytů

	Po	Út	St	Čt	Pá	So	Ne	
Drážďany	14	7	12	2	24	54	6	119
Celkem	130	170	140	160	760	530	110	2 000



- 14 ze všech 130 pobytů začínajících v pondělí se týká hostů z Drážďan.
- 10.77 % všech pobytů začínajících v pondělí se týká hostů z Drážďan.

- 54 ze všech pobytů 530 začínajících v sobotu se týká hostů z Drážďan.
- 10.19 % všech pobytů začínajících v sobotu se týká hostů z Drážďan.

10.77 = 100*14/**130**

10.19 = 100*54/**530**

Podíl kategorií na podmínce X podíl podmínky na kategoriích – shrnutí

Předpoklady:

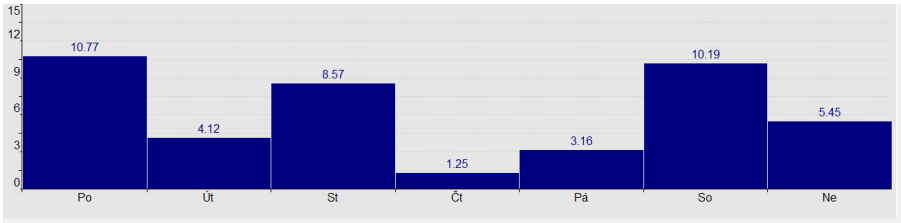
- Atribut A má kategorie $a_1,\ldots, a_K$
- Matice dat M má m řádků, podmínku χ splňuje n řádků

Kategorie atributu A	a_1	…	a_K	Σ
absolutní četnosti kategorií v matici dat M	m_1	…	m_K	m
absolutní četnosti kategorií v matici dat M/χ	g_1	…	g_K	n
podíl kategorie na podmínce	$100*g_1/n$	…	$100*g_K/n$	100
podíl podmínky na kategorii	$100*g_1/m_1$	…	$100*g_K/m_K$	-

Podíl podmínky na kategoriích - příklad

Inspirace: Podíl hostů z Drážďan na počtu všech přijíždějících hostů kolísá v jednotlivých dnech týdne mezi 1.25 a 10.77 procent.

	Po	Út	St	Čt	Pá	So	Ne	
Drážďany	14	7	12	2	24	54	6	119
Celkem	130	170	140	160	760	530	110	2 000



	Po	Út	St	Čt	Pá	So	Ne
dne	10.77	4.12	8.57	1.25	3.16	10.19	5.45

MAX ≥ 50 %

Existuje den týdne a kombinace údajů o bydlišti hosta, o typu pobytu a o počasí tak, že hostů vyhovujících dané kombinaci je alespoň 50% ze všech hostů přijíždějících v daném dni?

Podíl podmínky na kategoriích - příklad

Hotel.txt

HVek	HPohlavi	HMesto	HMesto_X	HMesto_Y	HStat	PPobytOd	PNoci	POsob	PTypPobytu	PCenaUbytovani	PCenaStrava	PCenaSleva	PCenaCelkem	DHodnoceni			
21	žena	České Budějovice	14.4757883	48.9763169	ČR	31.5.2013	1	1	rekreační	1450.00	0.000	0.00	1450.00	spokojen	91	82	81	56
34	muž	Linec	14.2862742	48.3066489	Rakousko	2.8.2013	2	4	rekreační	11600.00	1440.000	200.00	12840.00	průměr	44	21	62	84
30				48.211563														
62			1283	50.25														
35			767	48.2115														
58	muž	České Budějovice	14.475788															
81	žena	Vídeň	16.3736767	48.2115														
22	žena	Drážďany	13.7397044	51.														
82	muž	Katovice	19.0241283	50.25														
55	muž	Praha	14.4212806	50.087496														
75	žena	Berlín	13.3908886	52.5176189	Německo	26.5.2012	1	4	rekreační	4840.00	0.000	200.00	4640.00	nespokojen	30	39	32	3
66	žena	Linec	14.2862742	48.3066489	Rakousko	23.2.2012	1	2	služební	2420.00	0.000	0.00	2420.00	spokojen	95	93	75	21
64	žena	Linec	14.2862742	48.3066489	Rakousko	16.3.2013	7	2	rekreační	20300.00	0.000	600.00	19700.00	průměr	52	30	59	52
35	muž	Košice	21.2543528	48.7160408	Slovens													
32	muž	Mnichov	11.5836375	48.1364669	Německo													
65	muž	Plzeň	13.3771556	49.7490406	ČR	13.3.												
79	muž	Brno	16.6153758	49.1921808	ČR	3.5.2012	1	3	rekreační	3630.00	0.000	0.00	3630.00	průměr	59	43	48	59
28	žena	Drážďany	13.7397044	51.0497456	Německo	24.12.2012	1	2	služební	2420.00	0.000	0.00	2420.00	průměr	51	66	58	48
35	žena	Hamburg	10.0043528	53.5498325	Německo	5.1.2013	14	2	rekreační	46600.00	5040.000	200.00	45440.00	nespokojen	27	15	30	12
22	muž	Plzeň	13.3771556	49.7490406	ČR	9.11.2013	4	1	rekreační	5800.00	0.000	200.00	5600.00	spokojen	84	79	86	48
25	muž	Karlovy Vary	12.8690381	50.2311075	ČR	9.11.2013	7	2	rekreační	20300.00	0.000	600.00	19700.00	spokojen	88	74	94	54
20	žena	Hamburg	10.0043528	53.5498325	Německo	19.1.2013	14	4	rekreační	81200.00	10080.000	200.00	91080.00	průměr	59	23	46	96
30	žena	Linec				29.11.2013	2	4	rekreační	11600.00	0.000	200.00	11400.00	spokojen	88	82	84	61
45	žena	Karlov			R	28.12.2012	2	2	rekreační	4840.00	600.000	200.00	5240.00	spokojen	81	99	87	23

Existuje den týdne a kombinace údajů o bydlišti hosta, o typu pobytu a o počasí tak, že hostů vyhovujících dané kombinaci je alespoň 50% ze všech hostů přijíždějících v daném dni?

? CF: $_{\text{MAX} \geq 50\ \%\ \text{Cat}}$ PDenTydne / Bydliště, Typ, Meteo

Meteo.txt

MDatum	MTeplota	MObloha
4.1.2012	-6.3	slunečno
5.1.2012	-6.6	zataženo
6.1.2012	6.1	srážky
7.1.2012	1.6	srážky
8.1.2012	-1.3	srážky
	7	zataženo
		srážky
		zataženo
12.1.2012	-3.1	srážky
13.1.2012	-8.1	zataženo
14.1.2012	-10.7	srážky
15.1.2012	-5.5	zataženo
16.1.2012	2.3	zataženo
17.1.2012	-1.9	zataženo
18.1.2012	-8.6	zataženo

? CF: $_{\text{MAX} \geq 50\ \%\ \text{Cat}}$ PDenTydne / Bydliště, Typ, Meteo zadání pro CF-Miner

LM Hotel_Slidy MB - LISp-Miner Workspace module - 27.08.00

File Data Introduction Preprocessing Interactive Analysis Data-mining Tasks Domain Knowledge Window Help

Attributes | Tables | Tasks Overview | 01: PDenTydne / Bydl.. | 01: PDenTydne / Bydl.. | KL analysis (56) | CF-Miner - zadání ob.. | 02: Dotaznik x Dotaz.. | 02: Dotaznik x Dotaz..

Data-mining Task basic parameters

Name: 01: PDenTydne / Bydliště, Meteo - Podíl podmínky na kategoriích — ID: 51

Comment: SUM alespon 50

Taskgroup: 01: CF-Miner

Task type: CF-Miner — Data matrix: HotelPlusExterni — Edit

ATTRIBUTES FOR HISTOGRAM

Attributes
» PDenTydne(7 categories)

PDenTydne

QUANTIFIERS

Type	Source	Range	Rel.	Value	Units
SUM	Abs	all	>=	50	Abs
MAX	%Cat	all	>=	50	Abs

SUM ≥ 50

MAX ≥ 50 % Cat

CONDITION

Host/Bydliště	Con, 0 - 1
» HMesto (subset), 1 - 1	B, pos
» HStat (subset), 1 - 1	B, pos
» HCizinec_b (subset), 1 - 1	B, pos
Typ pobytu	Con, 0 - 3
» PPresSobotniNoc (subset), 1 - 1	B, pos
» PTypPobytu (subset), 1 - 1	B, pos
Meteo	Con, 0 - 1
» MObloha (subset), 1 - 1	B, pos

Host/Bydliště(*) ∧
∧ Typ pobytu(*) ∧ Meteo(*)

Total length: 1 - 5 {1 - 4}

Generation information

Status: Solved, 26 run(s)

Mode: Standard

Aggregate function

Type: Count(*)

Attribute: -

Task parameters

Include 'worse' extensions of condition: Yes — Maximal

Include extensions of coefficients with no change in the histogram: Yes

Include extensions of cedents with no change in the histogram: Yes — Extensio

Jedno z více možných zadání pro řešení dané analytické otázky.

Params | Validate | Task Clone

Run | Bkgrnd Run | Grid Run | Show Results

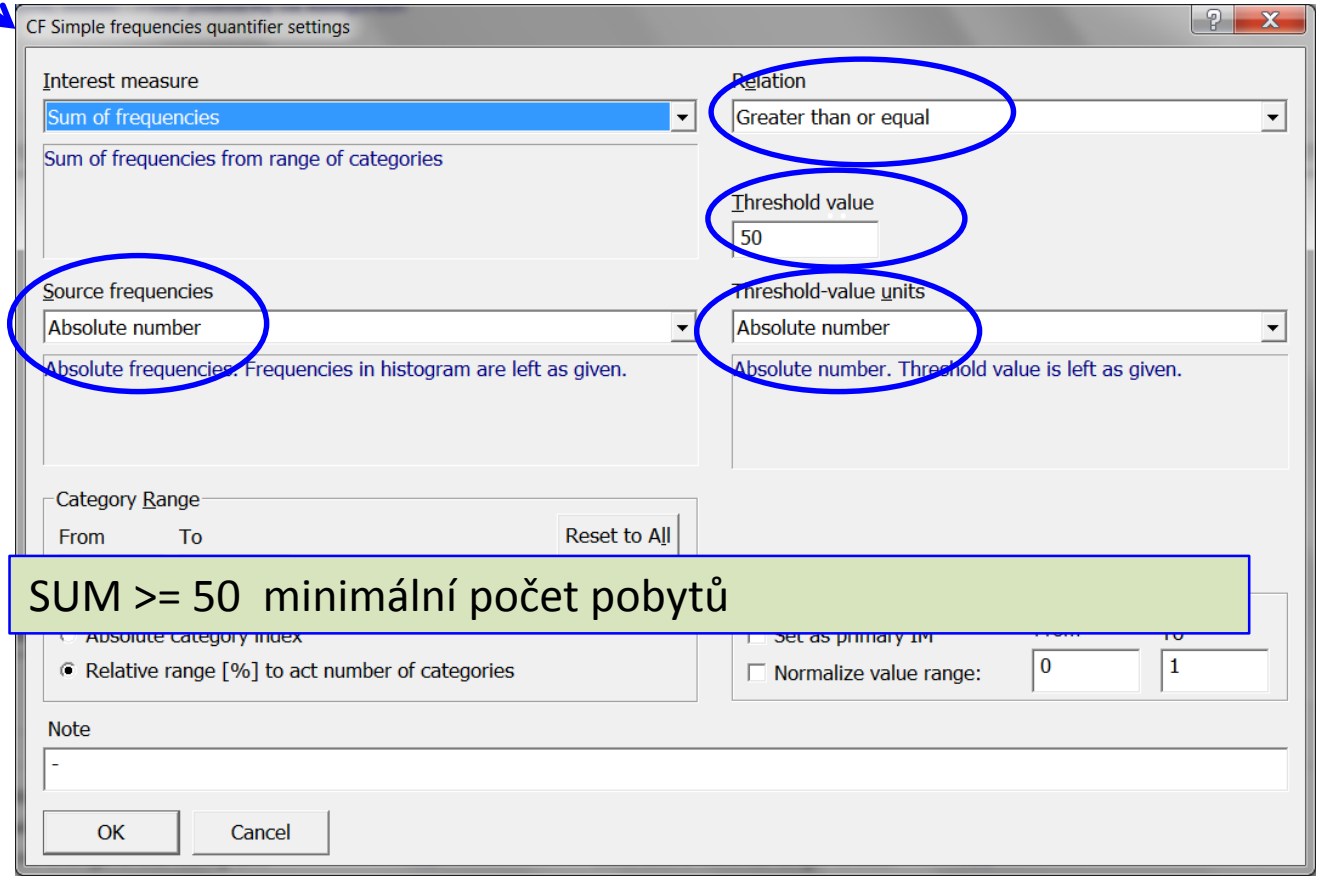
Ready — NUM

CS 100% 21:37 26.3.2017

CF-kvantifikátor SUM

QUANTIFIERS

Type	Source	Range	Rel.	Value	Units
SUM	Abs	all	>=	50	Abs
MAX	%Cat	all	>=	50	Abs

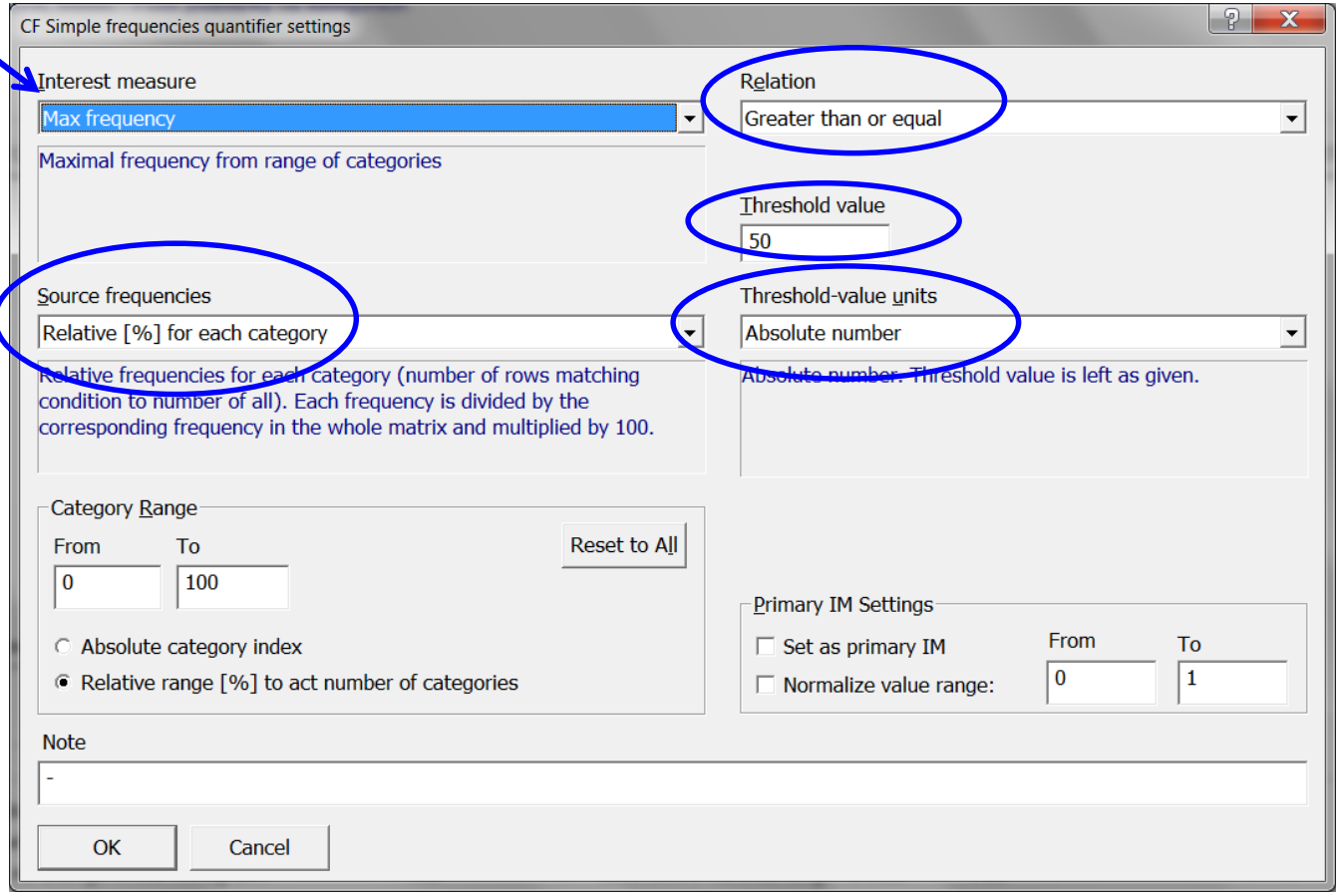


SUM >= 50 minimální počet pobytů

CF-kvantifikátor MAX ≥ 50 % Cat

QUANTIFIERS

Type	Source	Range	Rel.	Value	Units
SUM	Abs	all	>=	50	Abs
MAX	%Cat	all	>=	50	Abs



? CF: $_{\text{MAX} \geq 50\ \%\ \text{Cat}}$ PDenTydne / Bydliště, Typ, Meteo
výstup procedury CF-Miner

Task run

Start: 27.3.2017 08:45:17 Total time: 0h 0m 0s

Number of verifications: 110

Number of hypotheses: 8 Mode: Standard

Add group De

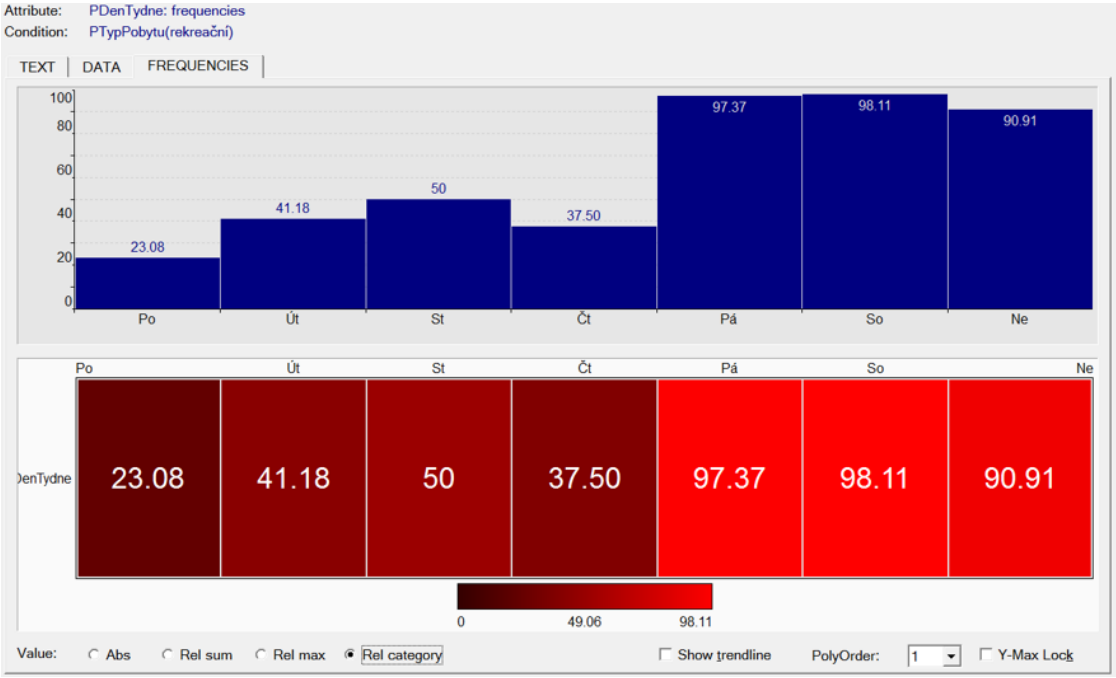
Actual group of hypotheses: All hypotheses

Hypotheses in group: 8 Shown hypotheses: 8 Highlighted: 0

Nr.	Id	MAX	Hypothesis
1	2	98.113	**PDenTydne / PTypPobytu**(*rekreační*)
2	4	76.923	**PDenTydne / PTypPobytu**(*služební*)
3	7	58.235	**PDenTydne / HCizinec**(*ano*)
4	8	56.415	**PDenTydne / HCizinec**(*ano*) & **PTypPobytu**(*rekreační*)
5	5	54.615	**PDenTydne / HStat**(*ČR*)
6	6	54.615	**PDenTydne / HCizinec**(*ne*)
7	1	52.500	**PDenTydne / MObloha**(*slunečno*)
8	3	51.184	**PDenTydne / PTypPobytu**(*rekreační*) & **MObloha**(*slunečno*)

Detailní výstup prvního histogramu (1)

Attribute: PDenTydne: frequencies
Condition: PTypPobytu(rekreační)



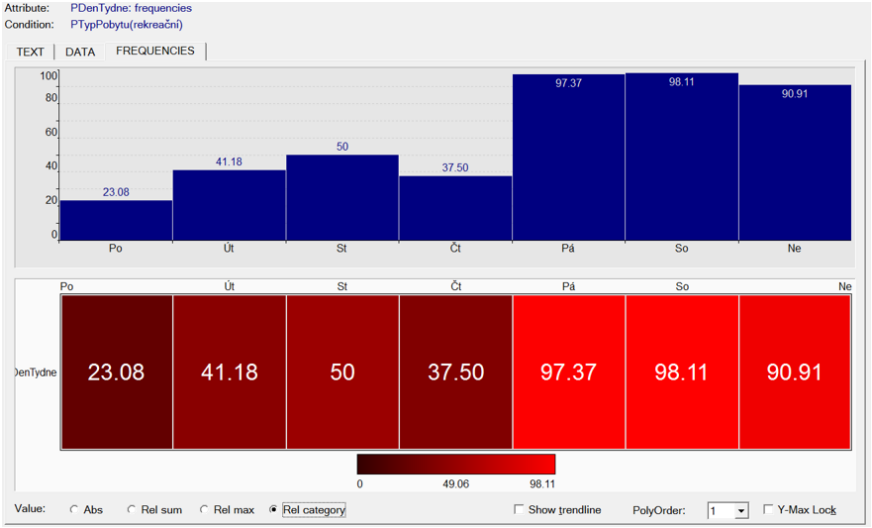
Histogram podílu podmínky PTypPobytu(rekreační) na kategoriích – dnech týdne říká, že

- z pobytů zahajovaných v pátek bylo 97.37 % pobytů rekreačních
- z pobytů zahajovaných v sobotu bylo 98.11 % pobytů rekreačních
- … atd.

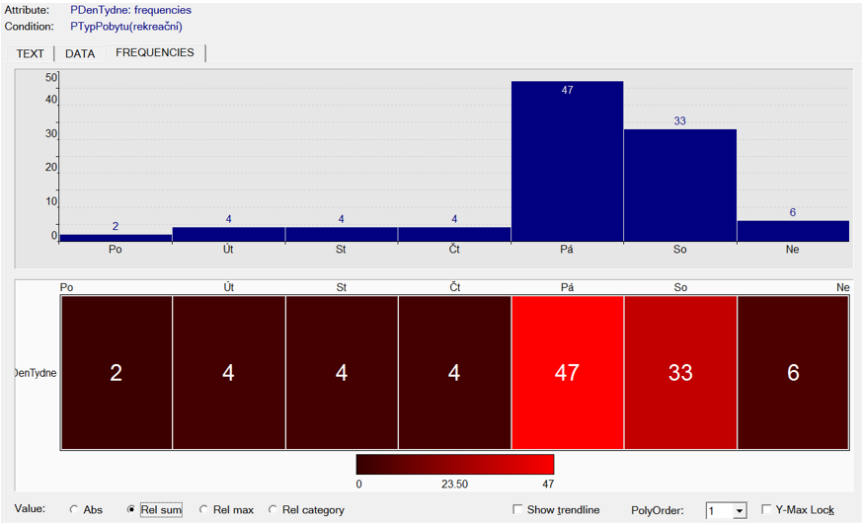
Detailní výstup prvního histogramu (2)

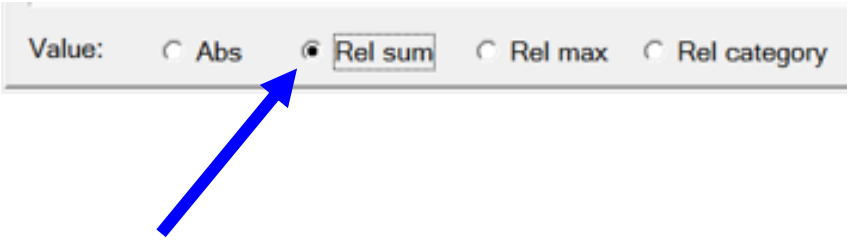
Attribute: PDenTydne: frequencies
Condition: PTypPobytu(rekreační)

Podíl podmínky na kategoriích
(odpovídá zadání)



Podíl kategorií na podmínce
(pro informaci)





Detailní výstup prvního histogramu (3)

Attribute: PDenTydne: frequencies
Condition: PTypPobytu(rekreační)

Attribute: PDenTydne: frequencies
Condition: PTypPobytu(rekreační)

TEXT | DATA | FREQUENCIES

MedianIA	5	5	Index (absolute <1;K>) of the median category
MedianIR	0.57	0.5714285714	Index (relative <0;1>) of the median category
StpsUp	2	2	Number of increases in frequency between adjectant categories in histogram
StpsDn	3	3	Number of decreases in frequency between adjectant categories in histogram
StpsUpC	1	1	Maximal number of consecutive increases in frequency between adjectant categories in histogr
StpsDnC	2	2	Maximal number of consecutive decreases in frequency between adjectant categories in histogr
PDSum	-1	-1	Sum of absolute values of differences of frequencies in the histogram and a given pattern
PDMin	-1	-1	The minimal of absolute values of differences of frequencies in the histogram and a given patter
PDMax	-1	-1	The maximal of absolute values of differences of frequencies in the histogram and a given patte

Values from underlying data:

AvgA	5.12	5.1194968553	Arithmetic average of cardinal values
AvgG	4.9	4.8969402816	Geometric average of cardinal values
Var	1.48	1.4762865393	Variance of cardinal values
StDev	1.22	1.2150253246	Standard deviation of cardinal values
Skew	-1.39	-1.3864529167	Skewness of cardinal values
Asym	-0.22	-0.2201257862	Asymetry coeficient of cardinal values

Frequency table

Category	Frequency	% Cummulative	% of Sum	% of Max	% of Category
Po	30	1.9	1.9	4.1	23.1
Út	70	6.3	4.4	9.5	41.2
St	70	10.7	4.4	9.5	50.0
Čt	60	14.5	3.8	8.1	37.5
Pá	740	61.0	46.5	100.0	97.4
So	520	93.7	32.7	70.3	98.1
Ne	100	100.0	6.3	13.5	90.9

Most-frequent category (or categories)

Category	Frequency	% of Sum	% of Max	% of Category
Pá	740	46.5	100.0	97.4