Statistika

Zpracování informací ze statistického šetření – Třídění statistického souboru

Roman Biskup

(zapálený) statistik ve výslužbě, aktuálně analytik v praxi ;-)
roman.biskup(at)email.cz

20. února 2012

Obsah

Třídění dle statistického znaku

- Důvody třídění:
  - zpřehlednění souboru,
  - zjištění empirického rozdělení statistického souboru,
  - snížení numerické náročnosti výpočtu statistických charakteristik.
- Dle počtu třídících znaků:
  - jednostupňové,
  - dvoustupňové (kontingenční tabulky),
  - vícestupňové.
- Dle typu třídění:
  - třídění prosté (malý počet různých hodnot znaku),
  - třídění intervalové (velký počet různých hodnot znaku, spojitý numerický znak).
- Základní zásady při třídění:
  - zásada úplnosti (každá jednotka musí někam patřit),
  - zásada jednoznačnosti (každá jednotka musí mít právě jedno místo při třídění).

Postup třídění I

- Prosté třídění – libovolný statistický znak
  1. stanovení počtu pozorování různých hodnot znaku (předpokládejme k různých hodnot)
- Intervalové třídění – numerický statistický znak
  1. stanovení počtu intervalů k, optimálně $8 \leq k \leq 20$
     - $k \approx 1 + 3{,}3 \cdot \log n$ (Sturgesovo pravidlo)
     - $k \approx \frac{8}{100}(\max x - minx)$
     - $k \approx \sqrt{n}$
  2. stanovení délky intervalu h

$$h = \frac{\max x - \min x}{k}$$

  3. rozdělení na intervaly $\langle \bullet; \bullet), \langle \bullet; \bullet), \ldots, \langle \bullet; \bullet \rangle$

$$\langle \min x + i \cdot h\,; \min x + (i+1) \cdot h)\,, \text{ pro } i = 0, \ldots, k-2 \text{ a}$$

$$\langle \min x + (k-1) \cdot h\,; \max x \rangle$$

  - Pro popis statistického znaku je vhodné jak délku intervalů, tak hranice intervalů „učesat“, tj. vhodně zaokrouhlit; je však třeba zajistit, aby takto upravené intervaly pokryly všechny hodnoty statistického znaku.

Postup třídění II

- Meze jednotlivých intervalů je třeba volit tak, aby nedocházelo k nejasnostem, tj. aby se každé pozorování jednoznačně „spadalo“ do určitého intervalu.

4. stanovení počtu pozorovaní s hodnotou znaku spadajícího do příslušného intervalu

Absolutní a relativní četnost I

Označme sledovaný statistický znak x, nechť má N pozorování, pak pro $i = 1, \ldots, k$:

n_i absolutní četnost
$\rightsquigarrow$ počet pozorování s hodnotou znaku rovnou x_i, respektive počet pozorování s hodnotou znaku spadající do i-tého intervalu,

- zřejmě platí:

$$\sum_{i=1}^{k} n_i = N.$$

p_i relativní četnost
$\rightsquigarrow$ poměr počtu pozorování s hodnotou znaku rovnou x_i vzhledem celkovému počtu pozorování, respektive poměr počtu pozorování s hodnotou znaku spadající do i-tého intervalu vzhledem celkovému počtu pozorování,

$$p_i = \frac{n_i}{N} \quad (p_i \cdot 100\,\%), \qquad i = 1, \ldots, k;$$

Absolutní a relativní četnost II

- i zde zřejmě:

$$\sum_{i=1}^{k} p_i = 1 \quad (100\,\%).$$

Kumulativní četnosti I

k_{n_i} kumulativní (absolutní) četnost
$\rightsquigarrow$ počet pozorování, u nichž je hodnota statistického znaku $x \leq x_i$, respektive počet pozorování zařazených díky hodnotě statistického znaku od prvního až do i-tého intervalu včetně, tj.

$$k_{n_i} = n_1 + n_2 + \cdots + n_i = \sum_{j=1}^{i} n_j.$$

k_{p_i} kumulativní relativní četnost
$\rightsquigarrow$ udává poměr počtu pozorování, u nichž je hodnota statistického znaku $x \leq x_i$, vzhledem k celkovému počtu pozorování respektive poměr počtu pozorování zařazených díky hodnotě statistického znaku od prvního až do i-tého intervalu včetně vzhledem k celkovému počtu pozorování, tj.

$$k_{p_i} = p_1 + p_2 + \cdots + p_i = \sum_{j=1}^{i} p_j.$$

- Je nutné uspořádání znaku x, tj. má smysl dělat minimálně pro ordinální znak.
- Nebo ne? Jakou by pak měla kumulativní četnost interpretaci?
- $k_{n_k} = N$, $k_{p_k} = 1 \quad (100\,\%)$

Datový soubor – Evidence studijních výsledků LS 2005

Obor	Počet ∗	Zameškáno	Zápočet	Body	Hodnocení
PUPN	4	0	Ano	4	1
VZ	0	3	Ano	1,5	4
OP	0	2	Rost	4	2
PP	0	0	Biskup	2	4
VZ	0	3	Ano	1	4
OP	0	1	Rost	2	4
ZOO	1	0	Ano	4	2
BT	13	1	Ano	4	2
OP	0	0	Rost	0,5	4
VZ	1	2	Ano	4	2
VZ	0	3	Ne	0	4
VZ	0	2	Ano	1,5	4
ZOO	2	1	Ano	1,5	4
⋮					

Přípravné práce – Evidence studijních výsledků LS 2005

Body – počet bodů získaných z písemné části zkoušky ze statistiky LS 2005 (řádný termín)

1. $N = 139$; $k = 13$ $(0; 0{,}5; \ldots; 6$ bodů) stanovení počtu pozorování jednotlivých hodnot znaku ...

Tabulka četností – Evidence studijních výsledků LS 2005

Počet bodů získaných z písemné části zkoušky ze statistiky LS 2005 (řádný termín)

x_i	n_i	p_i (%)	k_{n_i}	k_{p_i} (%)
0,0	27	19,42	27	19,42
0,5	11	7,91	38	27,34
1,0	20	14,39	58	41,73
1,5	15	10,79	73	52,52
2,0	14	10,07	87	62,59
2,5	11	7,91	98	70,50
3,0	22	15,83	120	86,33
3,5	8	5,76	128	92,09
4,0	7	5,04	135	97,12
4,5	3	2,16	138	99,28
5,0	1	0,72	139	100,00
5,5	0	0,00	139	100,00
6,0	0	0,00	139	100,00
Σ	**139**	**100,00**		

Datový soubor – Splátkový prodej (2004)

Věk	Pohlaví	Stav	Vzdělání	Zaměstnání	Příjem (Kč)	Úvěr (Kč)	Splátek
59	žena	ženatý	základní	důchodce	7 200	5 390	20
27	žena	ženatý	střední	dělník	7 000	7 542	20
50	muž	rozvedený	střední	kuchař	61 000	6 216	10
29	muž	svobodný	vyučený	dělník	10 000	7 002	20
31	muž	ženatý	vyučený	řidič	15 000	8 982	10
19	žena	druh	základní	mateř dovolená	5 500	6 696	10
22	muž	svobodný	vyučený	malíř, natěrač	10 000	4 621	20
34	muž	ženatý	střední	stát. zam.	15 159	7 624	30
45	žena	ženatý	vyučený	podnikatel	10 000	7 515	20
24	muž	rozvedený	vyučený	technik	12 000	6 680	20
30	muž	rozvedený	vyučený	pekař	12 500	3 228	20
25	muž	svobodný	střední	pol. inspektor	14 000	14 229	30
⋮							

Příprava intervalů – Splátkový prodej (2004)

Úvěr – cena zaplacená za celkový spotřebitelský úvěr;

1. $N = 737$; $k \approx 1 + 3{,}3 \cdot \log 737 = 10{,}463$, zvolme $k = 11$;
2. $\min x = 1\,584$ a $\max x = 25\,164$;
3. $h = \dfrac{25\,164 - 1\,584}{11} = 2\,151{,}273$, položme $h = 2\,200$ a dolní mez prvního intervalu rovnu 1 500 pak:

$$\begin{array}{rlll} 1. & \langle 1\,500 & ; & 3\,700) \\ 2. & \langle 3\,700 & ; & 5\,900) \\ 3. & \langle 5\,900 & ; & 8\,100) \\ & \vdots & & \\ 10. & \langle 21\,300 & ; & 23\,500) \\ 11. & \langle 23\,500 & ; & 25\,700\rangle \end{array}$$

4. stanovení počtu pozorovaní v jednotlivých intervalech ...

Tabulka četností – Splátkový prodej (2004)

Cena zaplacená za celkový spotřebitelský úvěr

	Tabulka četností:**Celková výše úvěru**			
OD DO	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
1 500 Kč<=x<3 700 Kč	69	69	9,36228	9,3623
3 700 Kč<=x<5 900 Kč	217	286	29,44369	38,8060
5 900 Kč<=x<8 100 Kč	218	504	29,57938	68,3853
8 100 Kč<=x<10 300 Kč	104	608	14,11126	82,4966
10 300 Kč<=x<12 500 Kč	55	663	7,46269	89,9593
12 500 Kč<=x<14 700 Kč	54	717	7,32700	97,2863
14 700 Kč<=x<16 900 Kč	15	732	2,03528	99,3216
16 900 Kč<=x<19 100 Kč	3	735	0,40706	99,7286
19 100 Kč<=x<21 300 Kč	0	735	0,00000	99,7286
21 300 Kč<=x<23 500 Kč	1	736	0,13569	99,8643
23 500 Kč<=x<25 700 Kč	1	737	0,13569	100,0000
CHD	0	737	0,00000	100,0000

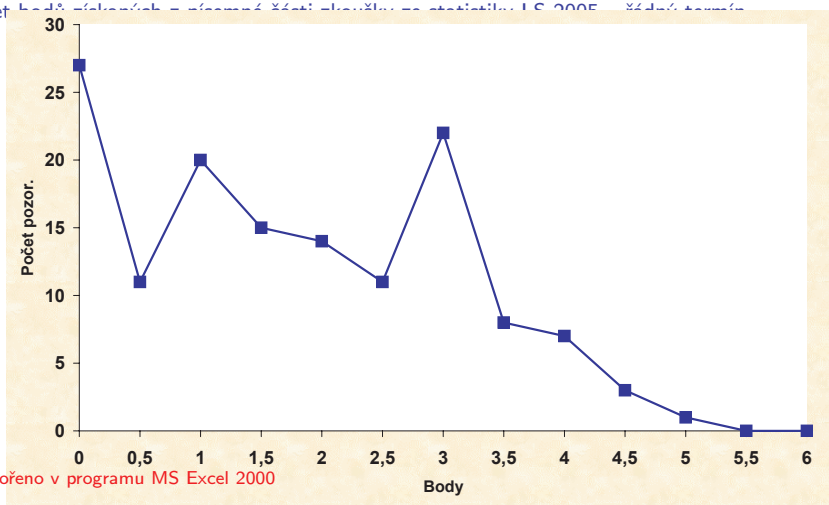
Vytvořeno v programu STATISTICA komplet 6.1 Cz

Polygon četností

- vizualizace absolutních četností – prosté třídění
  - na vodorovnou osu se vynáší hodnoty sledovaného znaku
  - na svislou osu se pak vynáší absolutní četnosti
  - nad jednotlivými hodnotami znaku jsou vynášeny hodnoty odpovídající příslušným absolutním četnostem
  - jednotlivé hodnoty jsou navíc spojeny lomenou čárou

Polygon četností

Počet bodů získaných z písemné části zkoušky ze statistiky LS 2005 – řádný termín



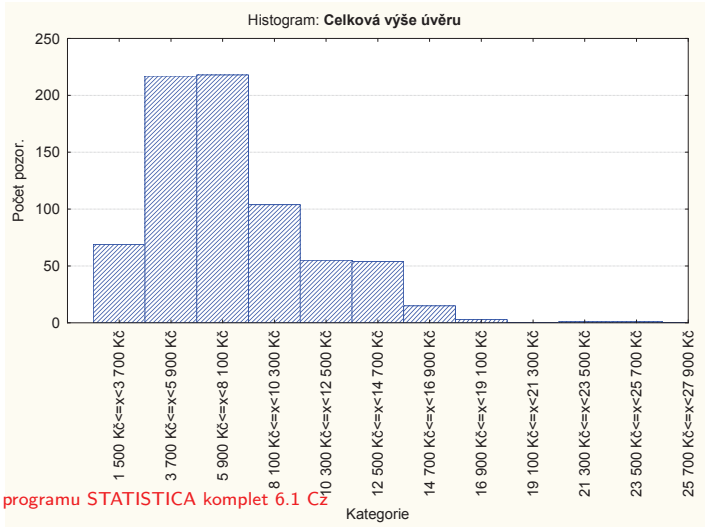
Vytvořeno v programu MS Excel 2000

Histogram četnosti

- vizualizace absolutních četností – intervalového třídění
  - na vodorovnou osu se vynáší meze intervalů
  - na svislou osu pak absolutní četnosti
  - nad jednotlivými intervaly jsou vykresleny sloupce s podstavou šířky intervalu a výškou absolutní četnosti
  - někdy jsou hodnoty vynášené na svislou osu modifikovány tak, aby celková plocha sloupců byla rovná jedné
- vše pochopitelně v měřítku ;-)

Histogram četnosti

Cena zaplacená za celkový spotřebitelský úvěr



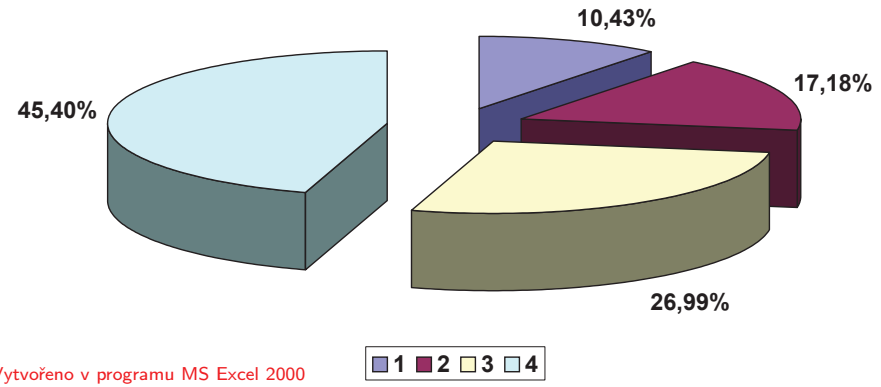
Vytvořeno v programu STATISTICA komplet 6.1 Cz

Výsečový (koláčový) graf

- vizualizace relativních četností
  - plocha grafu je dělena na kruhové výseče v poměru, který je dán relativní četnosti, tj.
  - $|\angle_i| = 360^\circ \cdot p_i$,
  - zřejmě platí: $\sum_{i=1}^{k} \angle_i = 360^\circ$.
- Graf je obvykle doplněn o legendu a relativní četnosti v procentech

Výsečový graf

Výsledné známky ze Statistiky 2004/05 – LS



Vytvořeno v programu MS Excel 2000