PSY117/454
Statistická analýza dat v psychologii
Přednáška 2

ČETNOSTI A ROZLOŽENÍ ČETNOSTÍ

Je snadné lhát s pomocí statistiky. Je těžké říkat pravdu bez ní.

Andrejs Dunkels; wikiquote

Jaké hodnoty máme v datech?

- Jaké hodnoty proměnné/ých se v datech vyskytují?
  - Jaké různé odpovědi jsme získali na tu kterou otázku dotazníku?
  - Jaké různé počty sledovaných chování se při pozorování vyskytly?
- Kolik kterých hodnot máme? - četnosti
  - Je některých víc, jiných míň?
  - Zdá se být v četnostech jednotlivých hodnot nějaký řád?

Studium statistiky

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	mě asi bude bavit	11	12,2	12,9	12,9
	by mě možná mohlo bavit	28	31,1	32,9	45,9
	mě asi bavit nebude	36	40,0	42,4	88,2
	mě rozhodně bavit nebude	10	11,1	11,8	100,0
	Total	85	94,4	100,0	
Missing	nedokážu říci	5	5,6		
Total		90	100,0		

Kolik tak přečtete za měsíc knížek (včetně elektronických a sešitových komixů)?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,0	2	2,2	2,3	2,3
	1,0	12	13,3	13,8	16,1
	1,5	2	2,2	2,3	18,4
	2,0	29	32,2	33,3	51,7
	2,5	2	2,2	2,3	54,0
	3,0	14	15,6	16,1	70,1
	3,5	2	2,2	2,3	72,4
	4,0	7	7,8	8,0	80,5
	4,5	1	1,1	1,1	81,6
	5,0	6	6,7	6,9	88,5
	6,0	2	2,2	2,3	90,8
	7,0	2	2,2	2,3	93,1
	8,0	1	1,1	1,1	94,3
	10,0	3	3,3	3,4	97,7
	17,0	1	1,1	1,1	98,9
	17,5	1	1,1	1,1	100,0
	Total	87	96,7	100,0	
Missing	999,0	3	3,3		
Total		90	100,0		

Kolik tak přečtete za měsíc knížek (včetně elektronických a sešitových komixů)? (Binned)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<= 0,0	2	2,2	2,3	2,3
	0,1 - 2,0	43	47,8	49,4	51,7
	2,1 - 4,0	25	27,8	28,7	80,5
	4,1 - 6,0	9	10,0	10,3	90,8
	6,1 - 8,0	3	3,3	3,4	94,3
	8,1 - 10,0	3	3,3	3,4	97,7
	16,1+	2	2,2	2,3	100,0
	Total	87	96,7	100,0	
Missing	999	3	3,3		
Total		90	100,0		

SPSS intervalové četnosti samo nedělá. Je třeba rekódovat hodnoty do intervalů (nová proměnná). K tomu např. fce Transform - Visual Binning.

OD DO	Tabulka četností:P13: Kolik tak p?e?tete za m?s?c kn??ek (v?etn? e			
	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
0,000000<=x<2,000000	16	16	17,77778	17,7778
2,000000<=x<4,000000	47	63	52,22222	70,0000
4,000000<=x<6,000000	14	77	15,55556	85,5556
6,000000<=x<8,000000	4	81	4,44444	90,0000
8,000000<=x<10,00000	1	82	1,11111	91,1111
10,00000<=x<12,00000	3	85	3,33333	94,4444
12,00000<=x<14,00000	0	85	0,00000	94,4444
14,00000<=x<16,00000	0	85	0,00000	94,4444
16,00000<=x<18,00000	2	87	2,22222	96,6667
18,00000<=x<20,00000	0	87	0,00000	96,6667
ChD	3	90	3,33333	100,0000

Tabulka četností (frekvencí)

hodnota/ interval	**(absolutní) četnost**	**kumulativní četnost**	**relativní četn. (%)**	**kumulativní rel. č.**
Minimum / interval1				
Hodnota2 / interval2				
…				
Maximum / posl. interv.		*N*		100
Celkem	*N*		100	

☹: „počet" v Tab 3.2, hustota (jde o hustotu pravděpodobnosti), obr. 3.5 – ne frekvence, ale procenta
AJ: (absolute) frequencies, relative frequencies, percent, cumulative, value, interval (class), total, N=sample size
V Excelu funkce ČETNOSTI. Zadává se zrádně: vybrat buňky, které mají obsahovat absolutní četnosti; napsat funkci a !!ukončit Ctrl+Shift+Enter.

Tabulka četností - poznámky

- Od nejmenší hodnoty po nejvyšší
- v 1. a 2. sl. obvykle zahrnuty chybějící hodnoty
  - Pak se rozlišuje mezi platnými hodnotami a chybějícími hodnotami
- hodnoty – kategorické proměnné, málo hodnot u metrické
- intervaly(třídy) – metrické proměnné
  - volba šířky intervalu (stojí za to vyzkoušet více)
    - aby byl jejich počet přibližně $N/10$, <15, nebo $1+\log_2 N$ (Sturgisovo pravidlo)
    - stejná šířka všech intervalů
- Pojem **odlehlá hodnota** (outlier)
- Tabulka četností zobrazuje téměř všechna data
  - Použitím intervalů již data mírně redukujeme

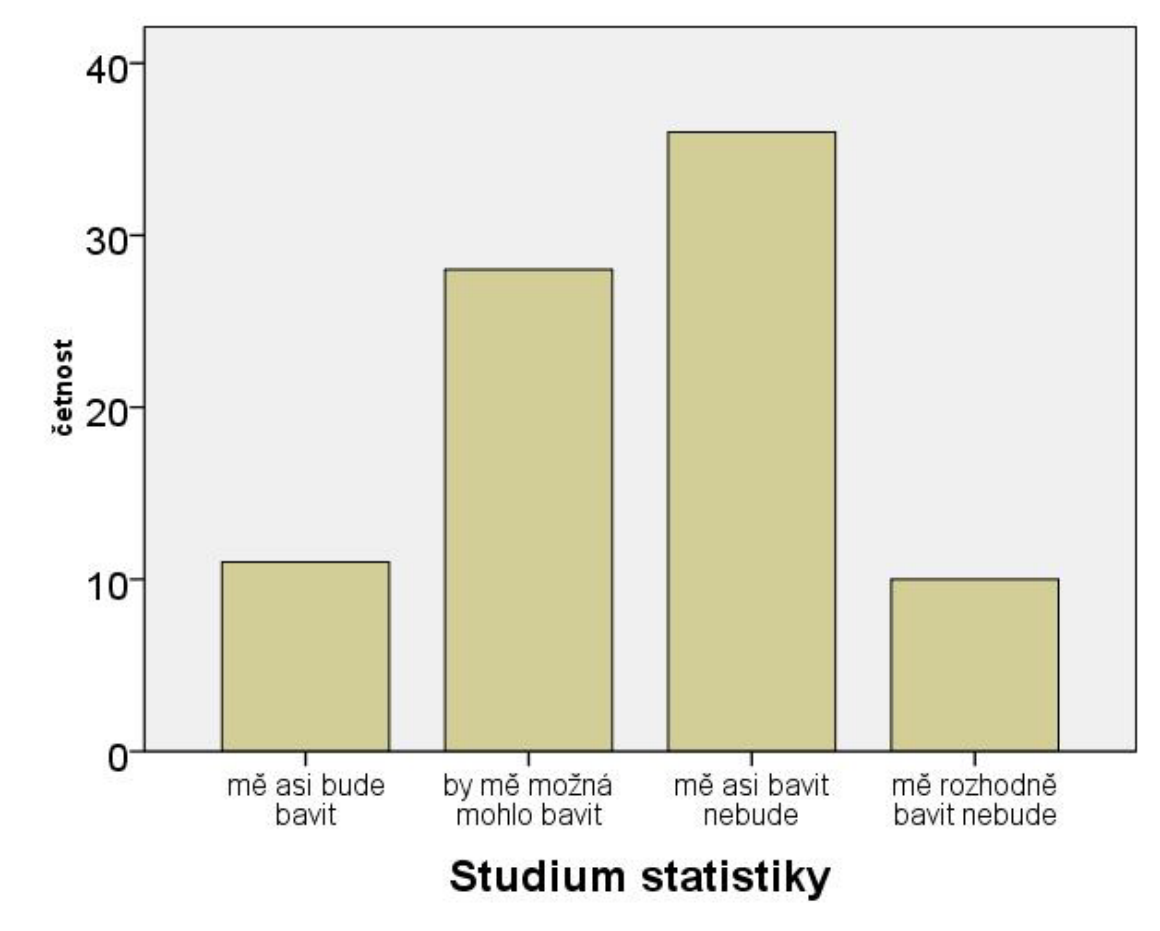
AJ: minimum, maximum, valid values, valid percent, loutlier, interval (bin, class), interval size

Grafické podoby tabulky četností

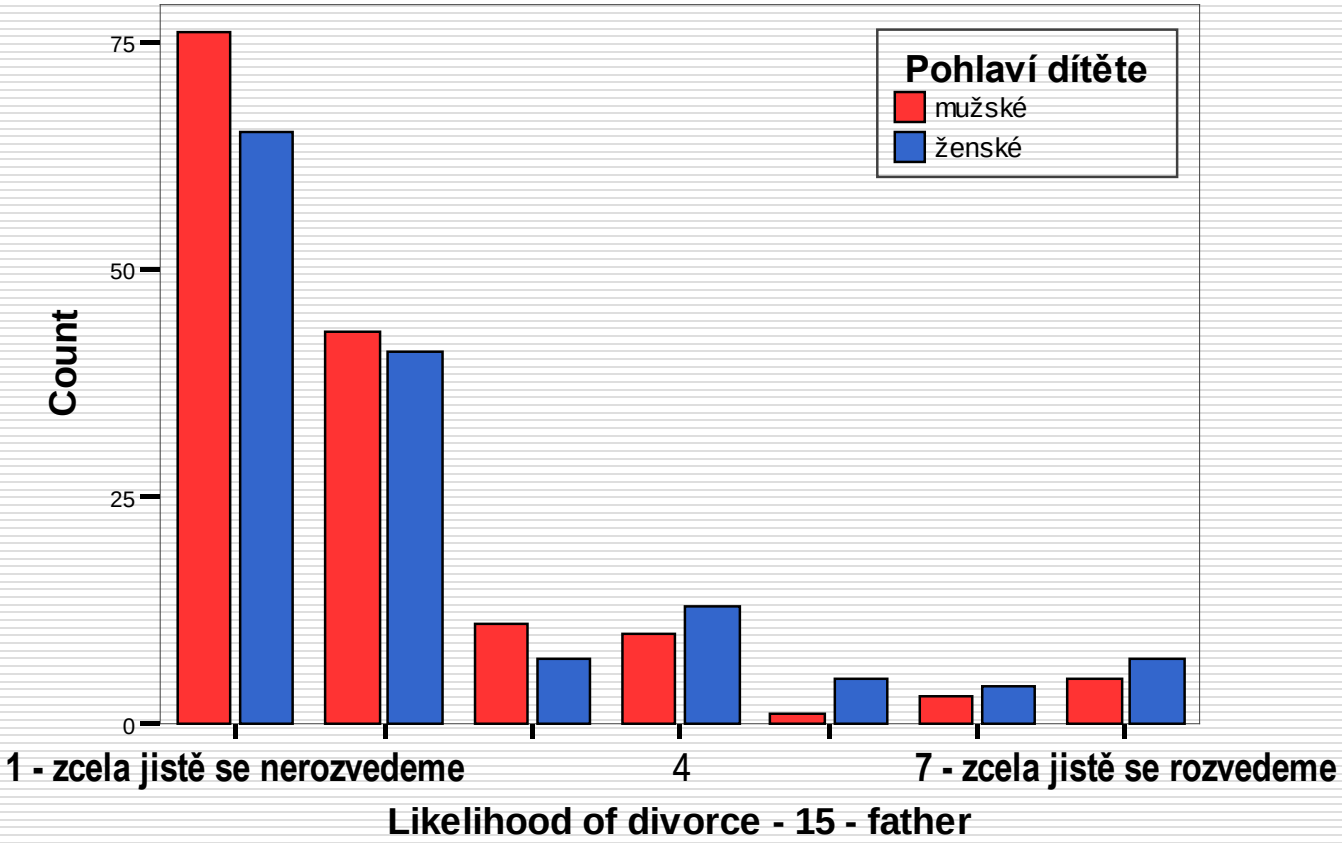
- Kategorické proměnné
  - sloupcový graf (diagram)
  - koláčový diagram – zřídkakdy, neukazuje rozložení
- Metrické proměnné
  - histogram / stem-and-leaf – rozdělení hodnot do intervalů

AJ: bar chart, histogram, pie chart, frequency distribution

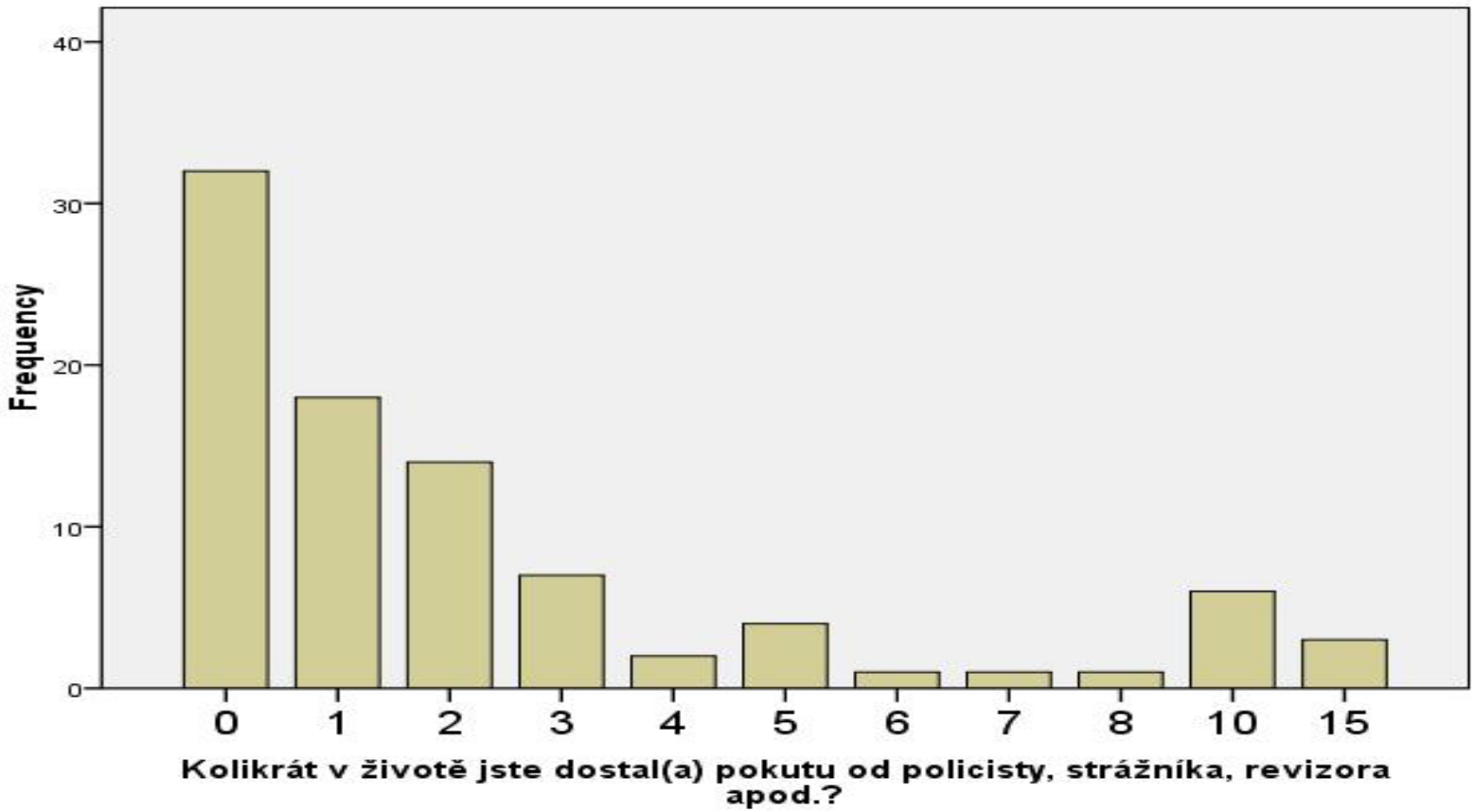
Sloupcový diagram



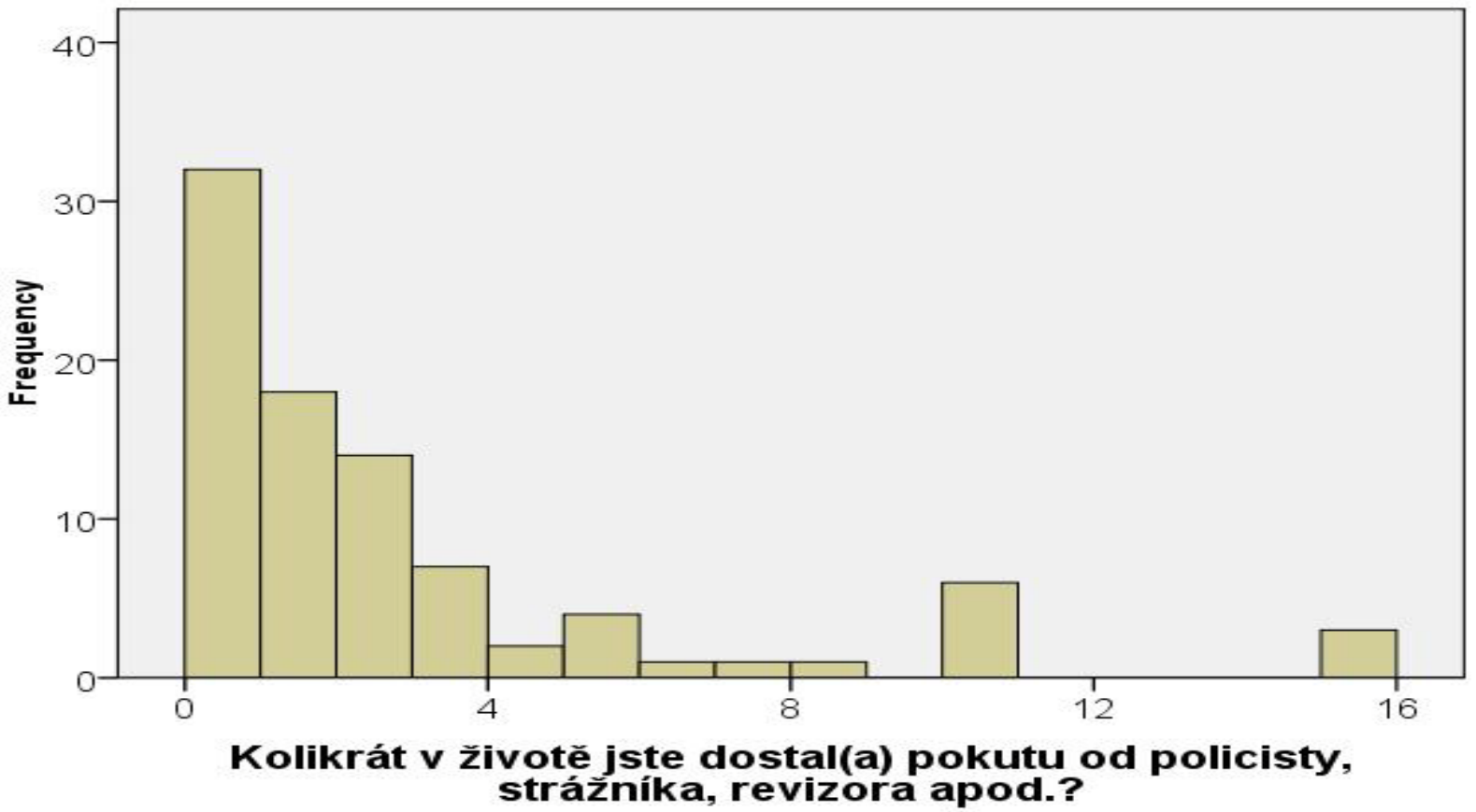
Sloupcový diagram s tříděním



?



Histogram



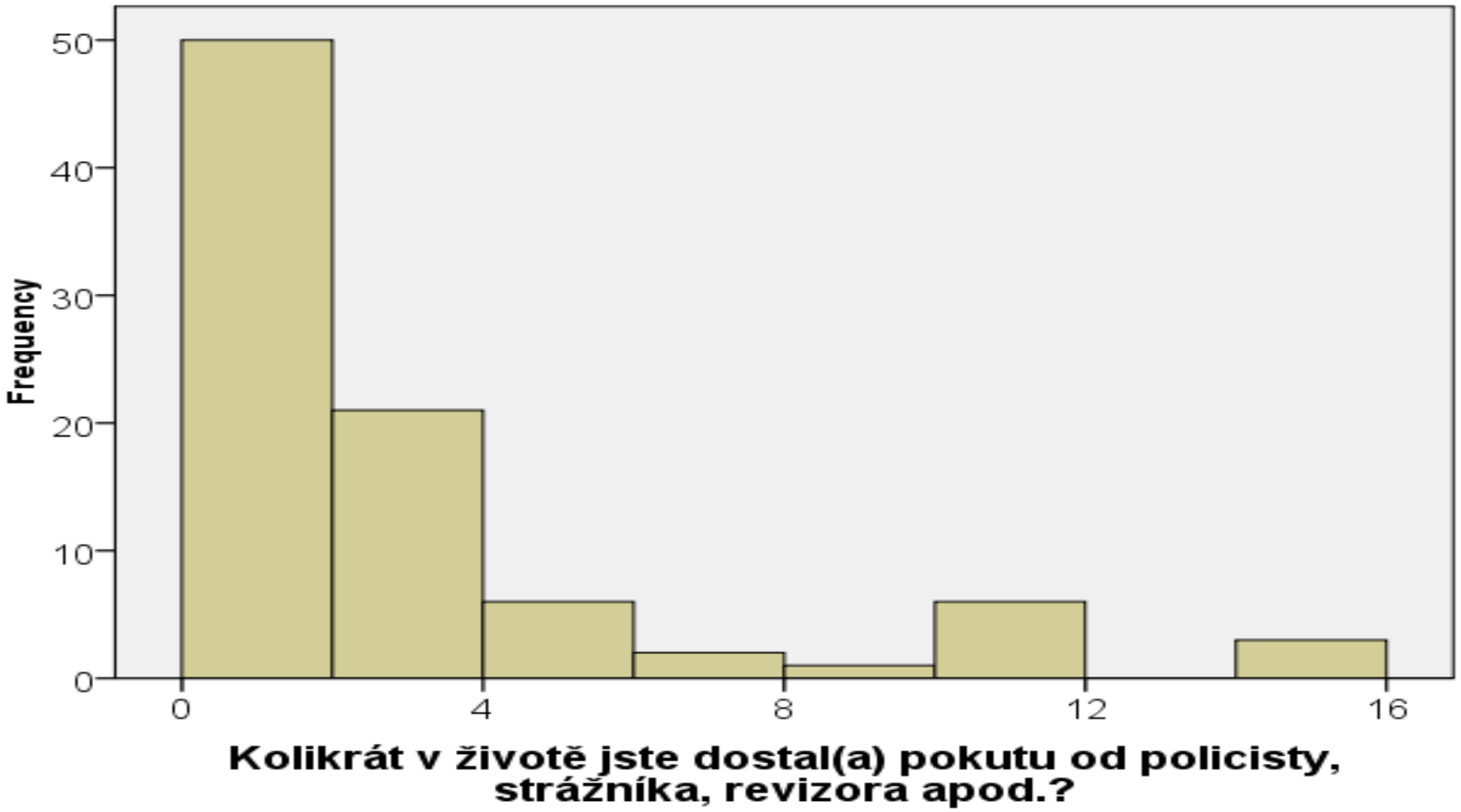
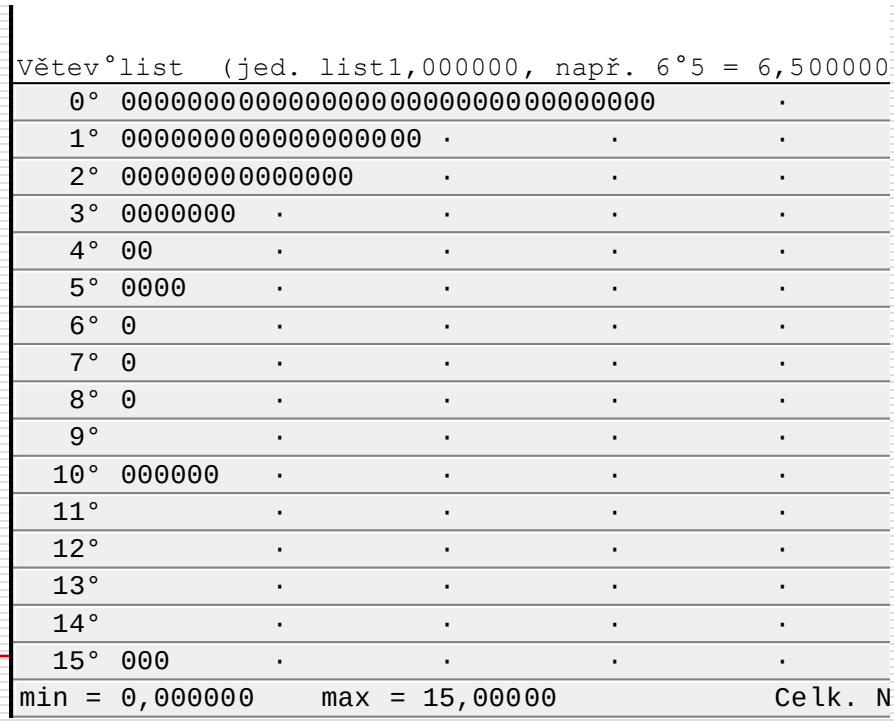


Diagram „stonek a list“

```
Frequency    Stem &  Leaf
   32,00        0 .  00000000000000000000000000000000
   18,00        1 .  000000000000000000
   14,00        2 .  00000000000000
    7,00        3 .  0000000
    2,00        4 .  00
    4,00        5 .  0000
    1,00        6 .  0
    1,00        7 .  0
   10,00 Extremes    (>=8,0)

Stem width:         1
Each leaf:       1 case(s)
```

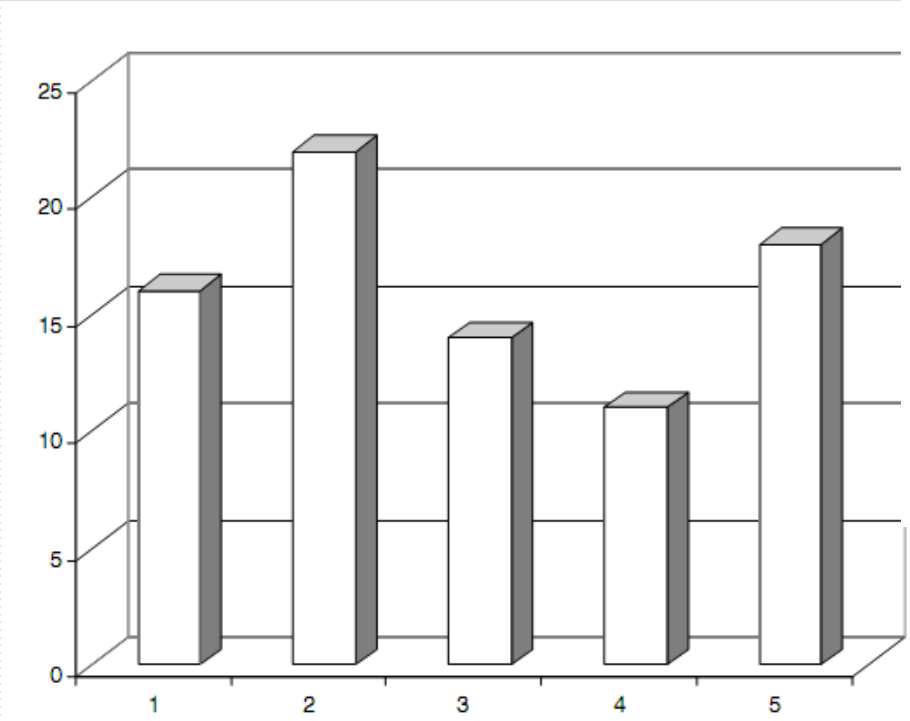


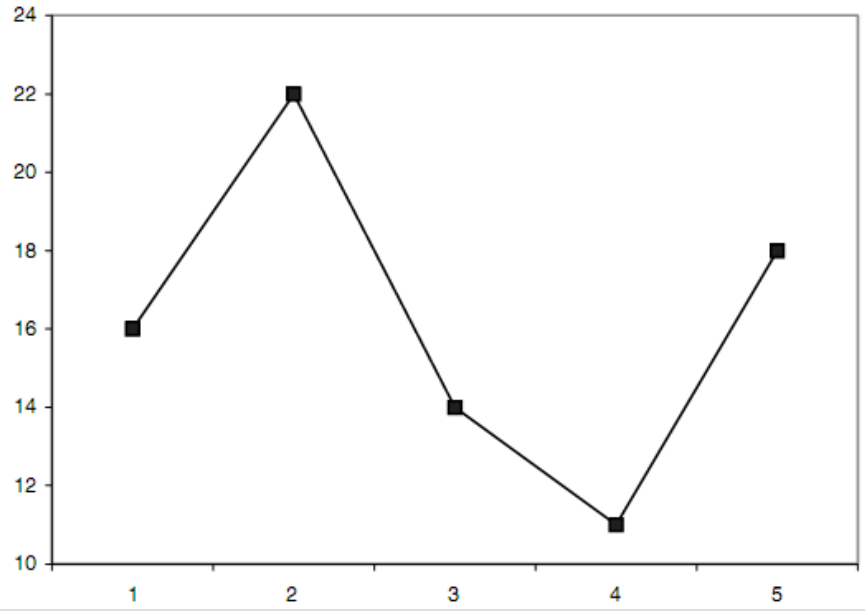
Větev°list (jed. list1,000000, např. 6°5 = 6,500000				
0°	00000000000000000000000000000000			·
1°	000000000000000000	·	·	·
2°	00000000000000	·	·	·
3°	0000000	·	·	·
4°	00	·	·	·
5°	0000	·	·	·
6°	0	·	·	·
7°	0	·	·	·
8°	0	·	·	·
9°		·	·	·
10°	000000	·	·	·
11°		·	·	·
12°		·	·	·
13°		·	·	·
14°		·	·	·
15°	000	·	·	·
min = 0,000000	max = 15,00000			Celk. N

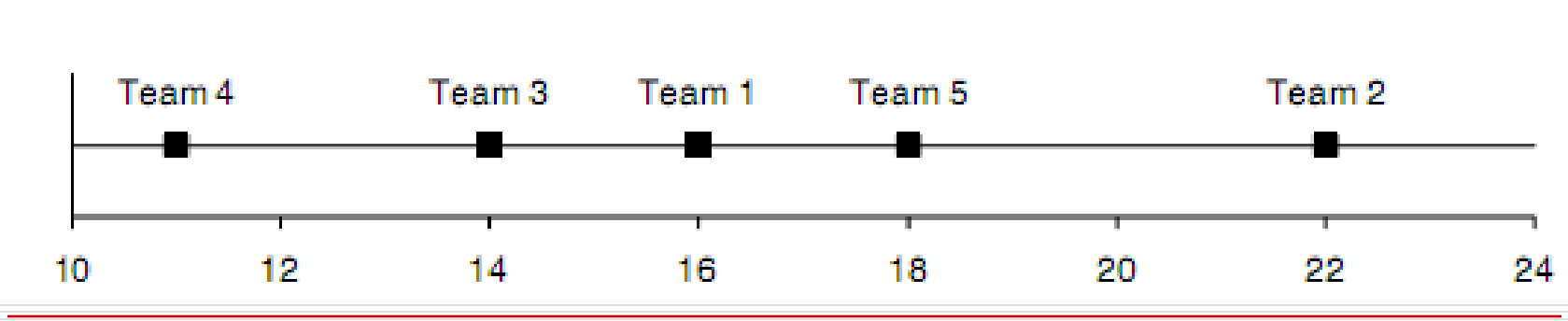
„Férové“ zobrazení dat

- Každý graf (i tabulka) musí být natolik přehledně popsán (nadpis + popisky uvnitř), aby byl srozumitelný i bez čtení textu

- Rozličné rady, např. Good, Hardin
  - V zobrazení by nemělo být více dimenzí než v zobrazovaných datech (často zbytečný 3. rozměr)
  - Popisky dat by neměly stínit datové body
  - Rozsah škál by měl být volen smysluplně, aby byla plocha užitečně využita („nulové“ body na škálách).
  - Numerické osy naznačují spojité proměnné, u kategorií volme raději textové popisky.
  - Nepropojujme datové body, jde-li o diskrétní škály, pokud nemá interpolace smysl, nebo pokud nemáme v úmyslu srovnání profilů

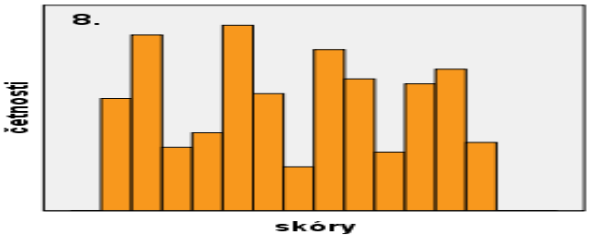


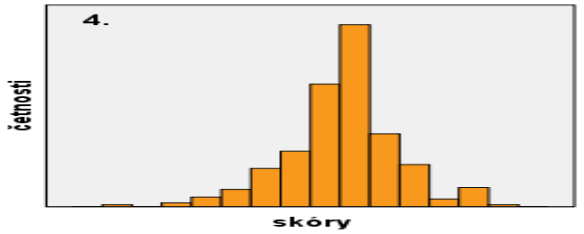




Rozložení *rozdělení, distribuce* četností

- Měřené jevy jsou nějak rozděleny do kategorií (intervalů) a tyto kategorie jsou různě „populární“ – četné.
- Četnosti u reálných ordinálních a vyšších proměnných obvykle nebývají **distribuovány** nahodile – jejich rozdělení zobrazené histogramem má popsatelný tvar.

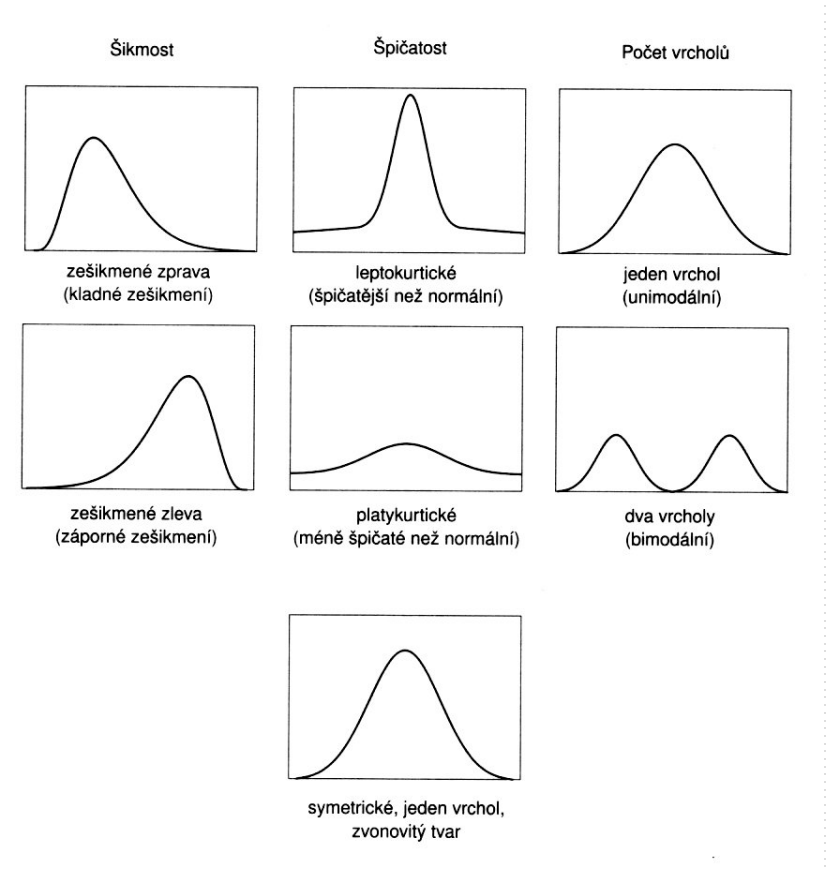




- **Rozdělení** četností je tedy to, kolik relativně (či absolutně) máme kterých hodnot měřené proměnné.
  - Typicky lze přibližně popsat slovy, např.: vyskytlo se hodně středních hodnot a relativně málo extrémních hodnot.
  - Toto **rozložení** jevů na měřené škále je nejlépe vidět na grafech.
  - Obvykle nějaké konkrétní rozložení očekáváme.

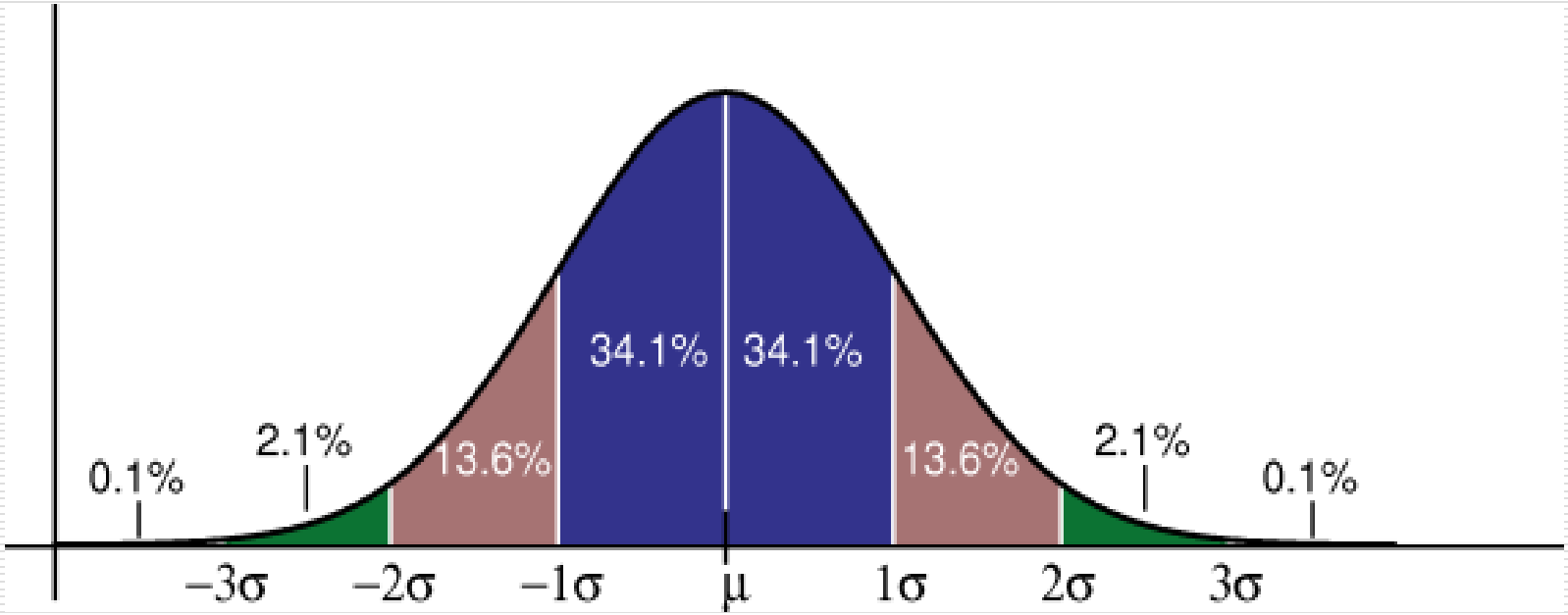
Tvar rozložení četností

- Normální
- Uniformní
- Počet vrcholů
  - Unimodální, bimodální, multimodální
- Zešikmení
  - Zešikmené zprava (pozitivně), efekt podlahy
  - Zešikmené zleva (negativně), efekt stropu
- Strmost
  - Leptokurtické, platykurtické



AJ: frequency distribution, normal, rectangular, unimodal, bimodal, positively/negatively skewed, lepto(platy)kurtic, floor/ceiling effect

Normální (Gaussovo) rozložení

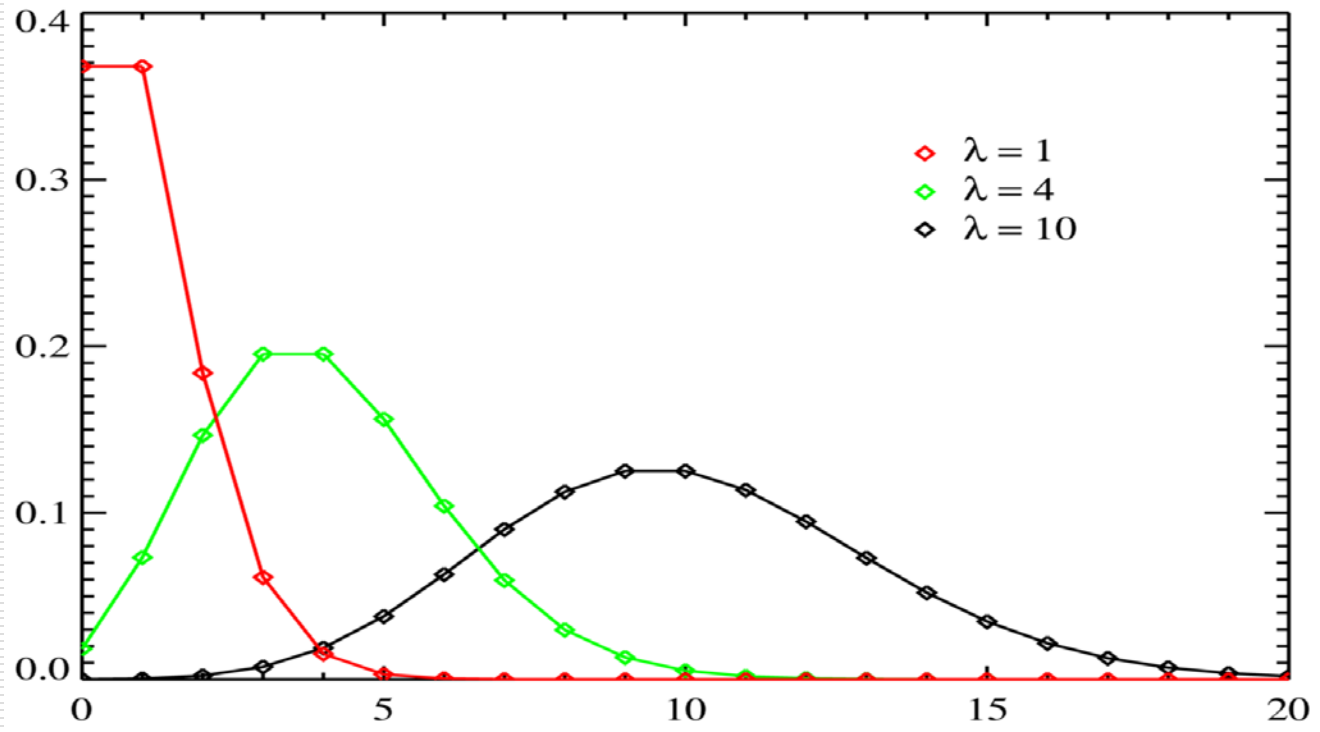


http://en.wikipedia.org/wiki/Image:Standard_deviation_diagram.png

- „Normální“ ve smyslu „velmi běžné“
- Tam, kde se setkává mnoho nezávislých vlivů.
- Ne vždy, nesouvisí s „kvalitou“ dat.

AJ: normal distribution, bell curve

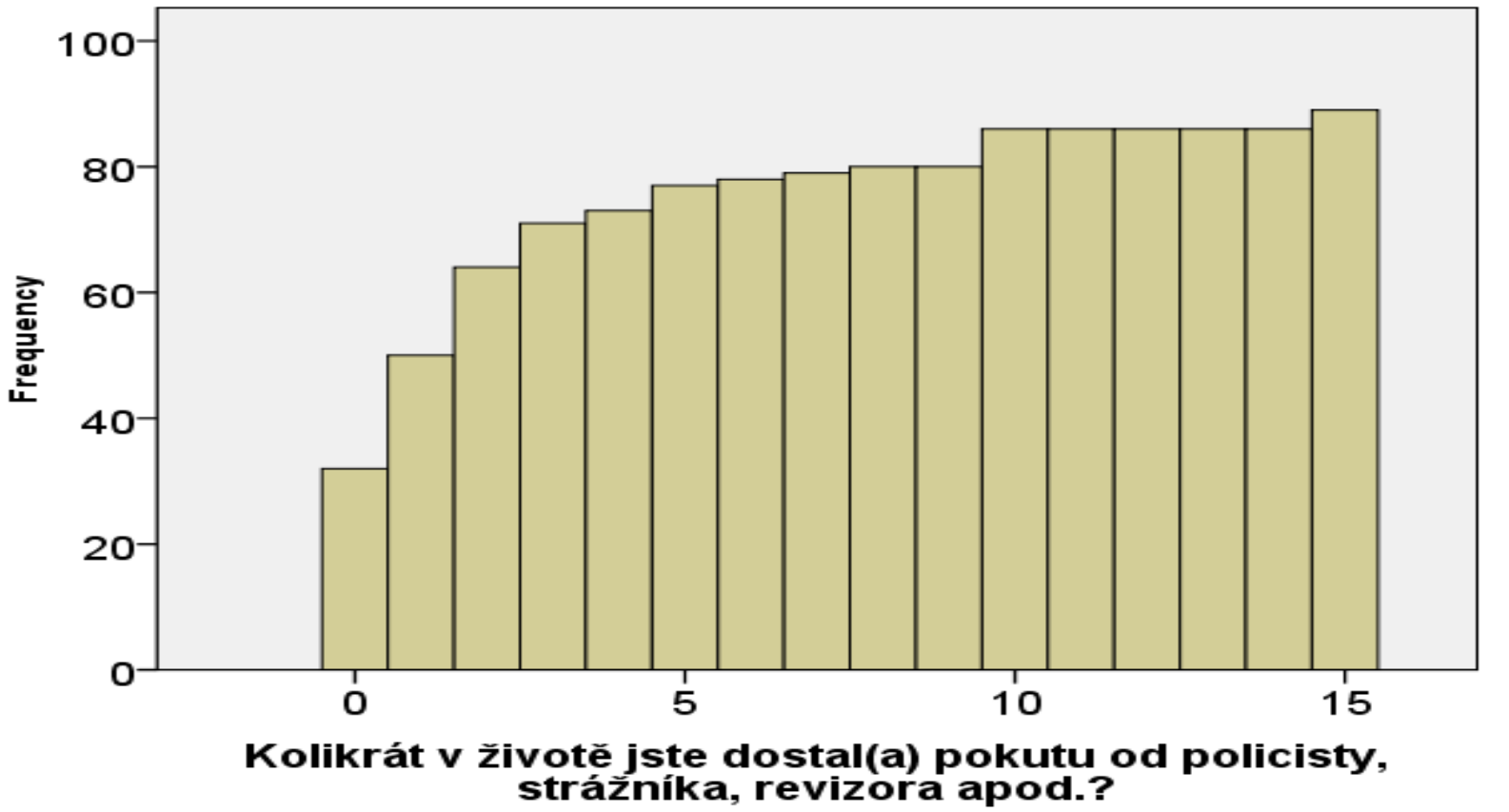
Poissonovo rozložení



- Rozložení řídkých událostí (ta lambda v grafu = frekvence za jednotku času)
- Děje-li se událost častěji, než 10x za časovou jednotku, která nás zajímá, je jeho dobrou aproximací normální rozložení.

AJ: Poisson distribution

Kumulativní histogram



Popis rozložení pomocí percentilů

- *X*-tý percentil
  - hodnota, pro kterou platí, že *X* % lidí (jevů) ve vzorku má/získalo tuto nebo menší hodnotu
  - lze snadno odečíst z kumulativního histogramu či patřičného sloupce tabulky četností
- Typicky rozložení popisujeme
  - 10., 20., …, 80.,90. percentilem – obecně
  - min, 25., 50., 75., max – nejčastěji
  - min., 1., 5., 10., 25., 50., 75., 90., 95., 99. – v normách

Shrnutí

- První informací (*statistikou*), která nás zajímá je **četnost** výskytu jednotlivých hodnot (resp. hodnot uvnitř jednotlivých intervalů)
- Konfiguraci **četností** nazýváme **rozložení (rozdělení)**.
- Rozložení popisujeme (=komunikujeme je)
  - tabulkou četností
  - graficky – histogram, sloupcový diagram
  - pomocí percentilů
- O typu, tvaru **rozložení** hodnot proměnné uvažujeme většinou graficky – **histogram, sloupcový diagram**.
- Nej… rozložením je tzv. **normální rozložení**.
- Byť tohle je 5. třída ZŠ – už tady se **podvádí**.