PSY117 2017

Statistická analýza dat v psychologii

Přednáška 2

MÍRY CENTRÁLNÍ TENDENCE A VARIABILITY

He uses statistics as a drunken man uses lampposts – for support rather than illumination.

Andrew Lang

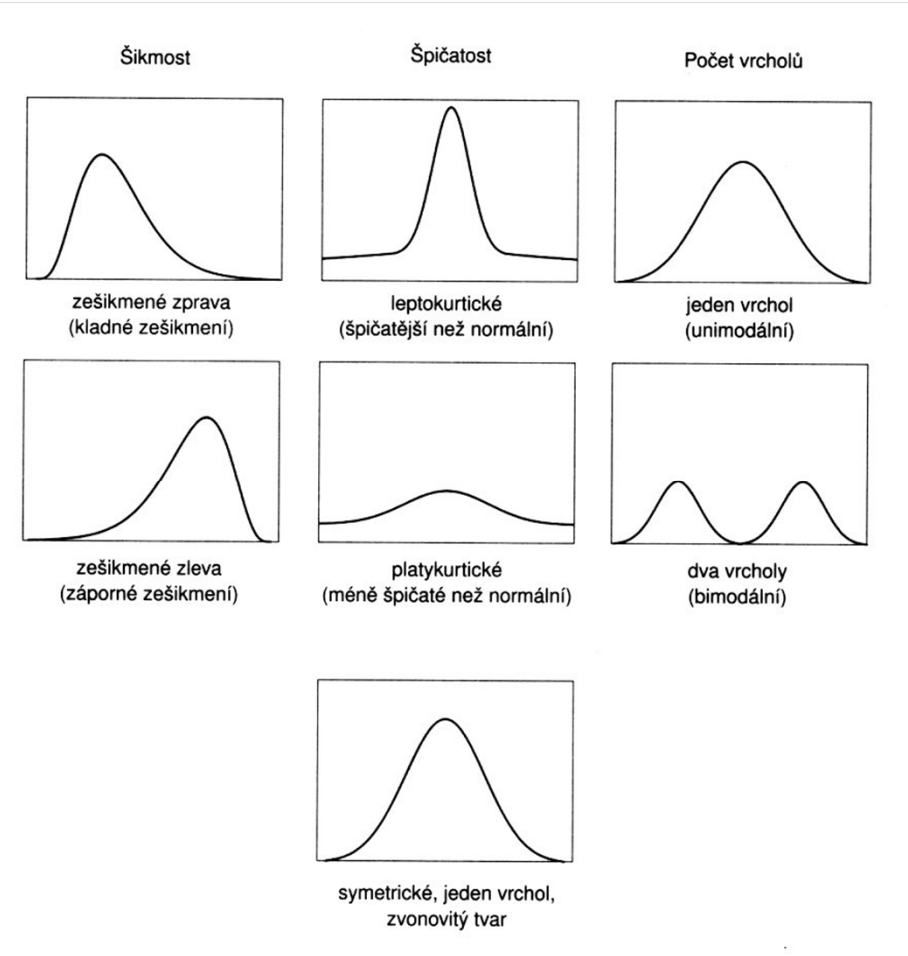
Rozložení *rozdělení, distribuce* četností

- 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.

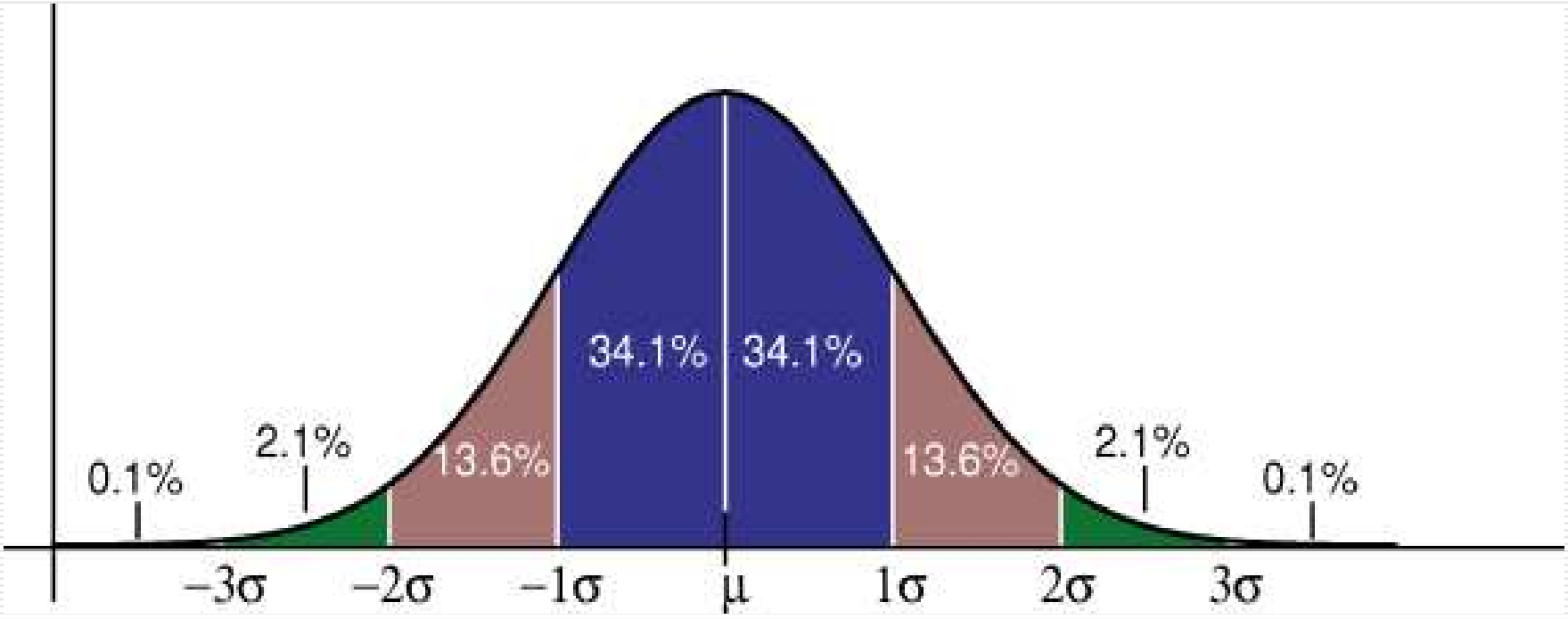
Tvar rozložení četností

- 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.

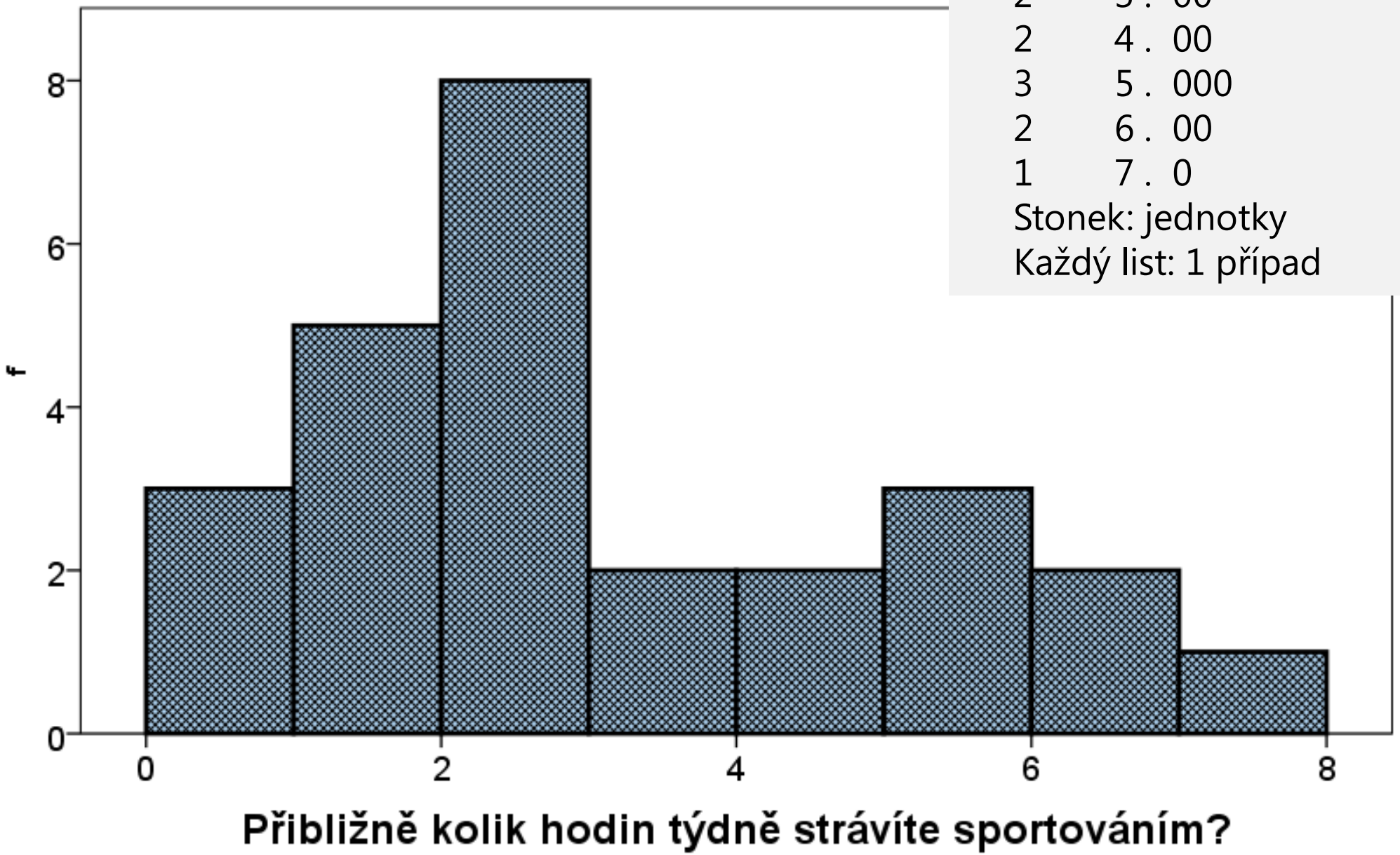
AJ: normal distribution, bell curve

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**.
- Nejčastěji diskutovaným rozložením je tzv. **normální rozložení**.

f	*Stonek a list*
3	0 . 002
5	1 . 00005
8	2 . 00000000
2	3 . 00
2	4 . 00
3	5 . 000
2	6 . 00
1	7 . 0

Stonek: jednotky
Každý list: 1 případ



Nedalo by se rozložení hodnot proměnné popsat úsporněji než pomocí tabulky četností, histogramu?

Kde na měřené škále se data nalézají?

UKAZATEL CENTRÁLNÍ TENDENCE

Jak moc jsou na ní rozptýlená?

UKAZATEL VARIABILITY

+ tvar rozložení (často implicitně)

Centrální tendence (=střední hodnoty, umístění)

- CT je jeden údaj, jímž se snažíme popsat rozložení četností jedné proměnné
  - Kouzlo i zrádnost je právě v tom, že je to 1 údaj.
- CT udává průměrnou, typickou, reprezentativní, *očekávanou* hodnotu
  - Co přesně tím míníme, záleží na tom, jakou míru CT se rozhodneme použít
- CT udává, kde na číselné ose si představujeme rozložení proměnné – odtud ukazatel *lokace*, umístění

Modus, medián a průměr

Modus - **kategoriální** typická hodnota

$\hat{X}$, Mo

- nejčastější hodnota, h. s nejvyšší četností
- jediná možnost u nominálních dat, u vyšších úrovní často užitečnou volbou

Medián – **pořadová** střední hodnota

$\tilde{X}$, Md

- hodnota prvku uprostřed uspořádaného souboru, 50. percentil (P_{50})
- při sudém počtu prvků je mediánem kterékoli číslo z intervalu mezi nejbližší vyšší a nejbližší nižší hodnotou (konsensuálně střed intervalu)
- pořadová data a výše

Aritmetický průměr – deviační, odchylková, **momentová** střední h.

$\overline{X}$, M, m

- jak ho znáte ze školy
- pouze intervalová a poměrová data
- velmi citlivý na extrémní hodnoty
- hodnota, od které je součet kvadratických odchylek nejmenší.

AJ: mode, median, mean

Jak spočítat *Mo*, *Md*, *M*

Mo

- vyčteme z tabulky četností
- Excel: =MODE(*rozsah_s_daty_proměnné*)

Md

- kategorické p.: vyčteme z tabulky četností, nejsnáze kum.
- spojité p. --> intervalové četnosti --> interpolujeme…
- formálně,
  - je-li N liché, je to X_k (k-tý prvek setříděné řady hodnot proměnné), kde $k=(N+1)/2$,
  - je-li N sudé, je to průměr X_k a X_{k+1}, kde $k=N/2$
- Excel: =MEDIAN(*rozsah_s_daty_proměnné*)
- =PERCENTIL(*rozsah_s_daty_proměnné*;0,5)

M

- Excel: =PRŮMĚR(*rozsah_s_daty_proměnné*)

Medián u intervalových četností

a spojitých škál s celými hodnotami

	f	%	cum %
0 - 0,99	3	11,5	11,5
1 - 1,99	4	19,2	30,8
2 - 2,99	9	30,8	61,5
3 - 3,99	2	7,7	69,2
…	…	…	…
7 - 7,99	1	3,8	100,0
Celkem	26	100,0	

1. Identifikujeme interval, v němž kumulativní četnost přesáhne 50% 2-2,99
2. Četnost tohoto intervalu = f_m = 9
3. Kumulativní četnost pro předchozí interval = f_p=7
4. Horní mez předchozího intervalu = L_p=1,99
5. Šířka intervalu = W = 1
6. Vypočítáme medián

$$Md = L_p + W(N/2 - f_p)/f_m =$$

$$= 1{,}99 + 1(26/2 - 7)/9 = 2{,}66$$

*Takto odhadnutý medián závisí na tom, jak jsou stanoveny hranice intervalů.

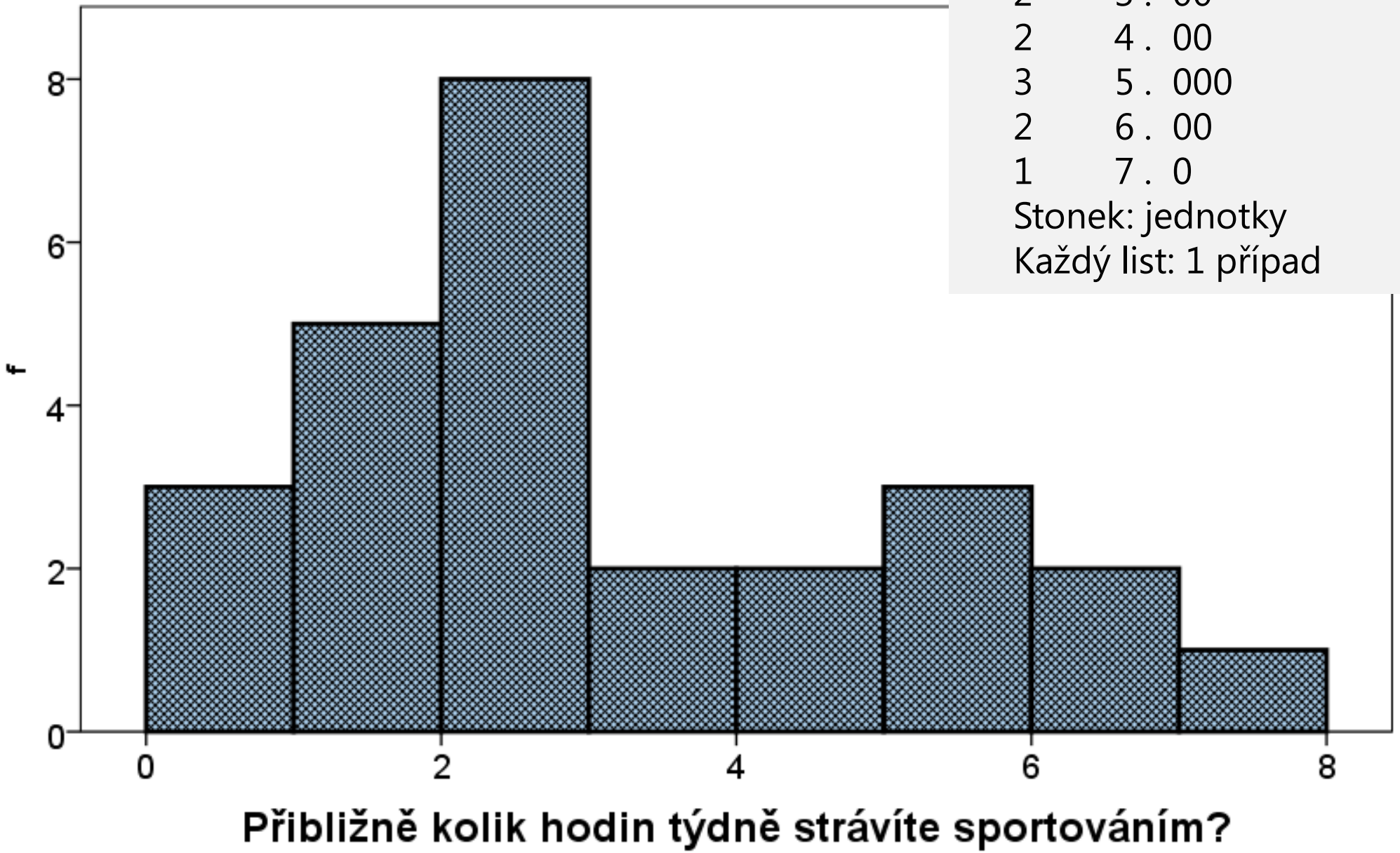
Mo=2 *Md=2* *M=2,68*

Md=2,66

f	*Stonek a list*
3	0 . 002
5	1 . 00005
8	2 . 00000000
2	3 . 00
2	4 . 00
3	5 . 000
2	6 . 00
1	7 . 0

Stonek: jednotky

Každý list: 1 případ



Míry variability (rozptýlenosti)

- Druhé číslo, jímž popisujeme rozložení hodnot proměnné
- Udává, jak moc či málo jsou data na škále rozptýlená.
  - Malá variabilita = většina hodnot v souboru je stejných nebo velmi blízkých
  - Vysoká variabilita = hodnoty jsou velmi rozmanité (n. rozložení je bimodální)

Rozpětí, rozptyl, směrodatná ochylka

Nominální statistika– entropie – nepoužívá se

Pořadové statistiky

- (variační) rozpětí = $X_{max} - X_{min}$ (extrémně roste s velikostí vzorku)
- (inter)**kvartilové rozpětí** = $Q_3 - Q_1$, IQR

Odchylkové (deviační, momentové) statistiky

- založené na odchylkách od průměru: $x = X - m$
- průměrná absolutní odchylka ($\Sigma|x| / n$) – nepoužívá se
- průměrná odchylka na druhou – **rozptyl** – s^2, $VAR(X)$
  - populační ($\Sigma x^2 / n$) vs. výběrový ($\Sigma x^2 / (n-1)$)
  - součet odchylek na druhou = **suma čtverců**
- **směrodatná odchylka** (standardní odchylka) – s, SD
  - odmocnina rozptylu - návrat k původní jednotce

AJ: measures of variability, entropy, rank-order, range, interquartile range, variance, standard deviation, sum of squares, square, square root

HC: Chyba v komputačním vzorečku na rozptyl.

Jak spočítat ukazatele variability

- IQR = Q_3-Q_1
  - Q_1=X_k[*)], kde k=(N+1)*0,25 zaokrouhleno dolů
  - Q_3=X_k, kde k=(N+1)*0,75 zaokrouhleno dolů
  - =PERCENTIL(*rozsah_s_daty_proměnné*; 0,25) resp. 0,75

  U spojitých proměnných lze využít intervalového výpočtu jako u mediánu.
- SD/VAR
  1. pro každý skór spočítáme deviační skór x_i=X_i-M
  2. deviační skóry umocníme na druhou
  3. druhé mocniny deviačních skórů sečteme a podělíme (N-1)
  4. pro SD výsledek ještě odmocníme
  - =VAR.VÝBĚR(*rozsah_s_daty_proměnné*)
  - =SMODCH.VÝBĚR(*rozsah_s_daty_proměnné*)

[*)] hodnota k-tého prvku seřazené řady hodnot proměnné X

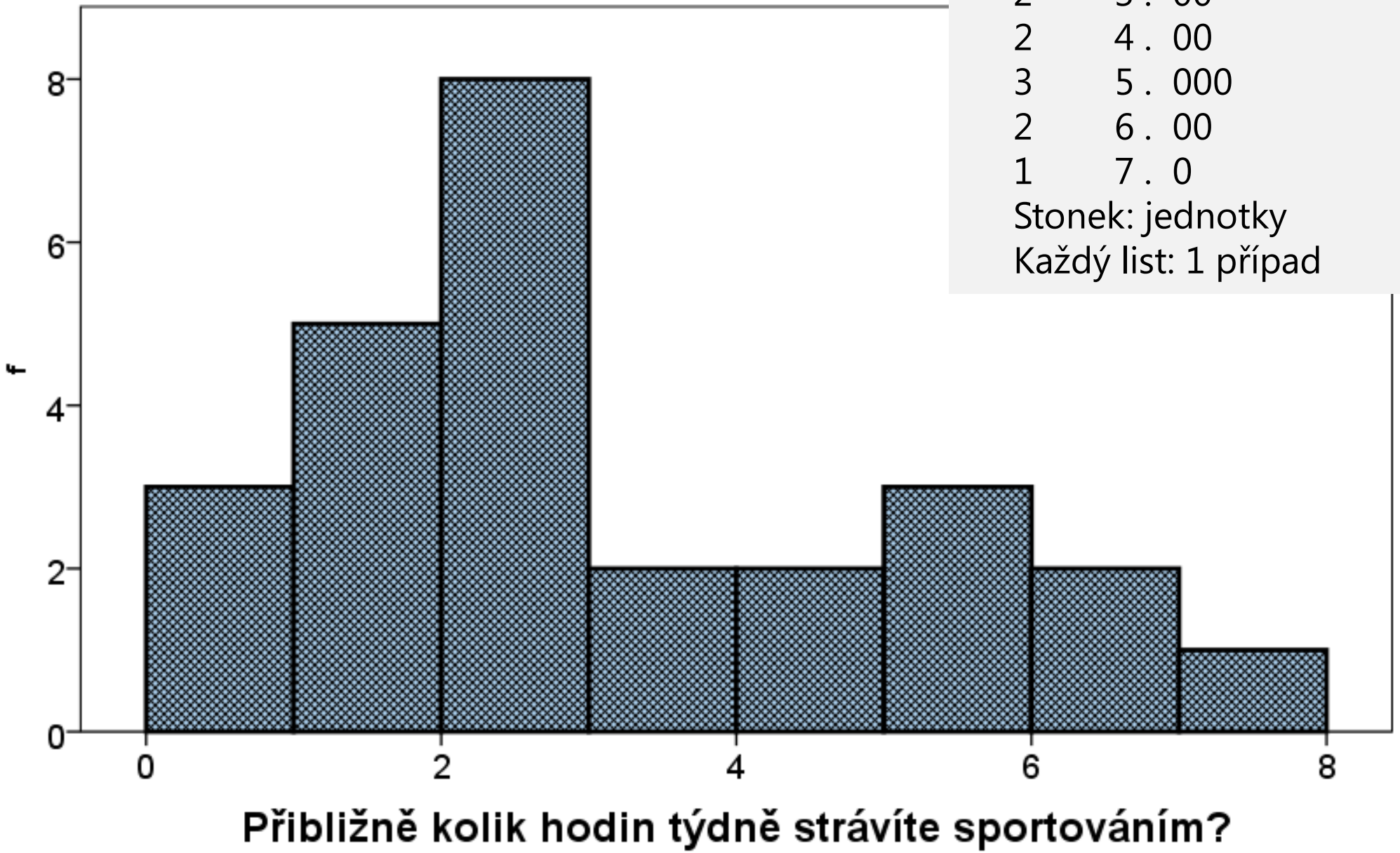
$Mo=2$ $Md=2$ $M=2,68$

$IQR=3$ $SD=1,97$

f	*Stonek a list*
3	0 . 002
5	1 . 00005
8	2 . 00000000
2	3 . 00
2	4 . 00
3	5 . 000
2	6 . 00
1	7 . 0

Stonek: jednotky

Každý list: 1 případ



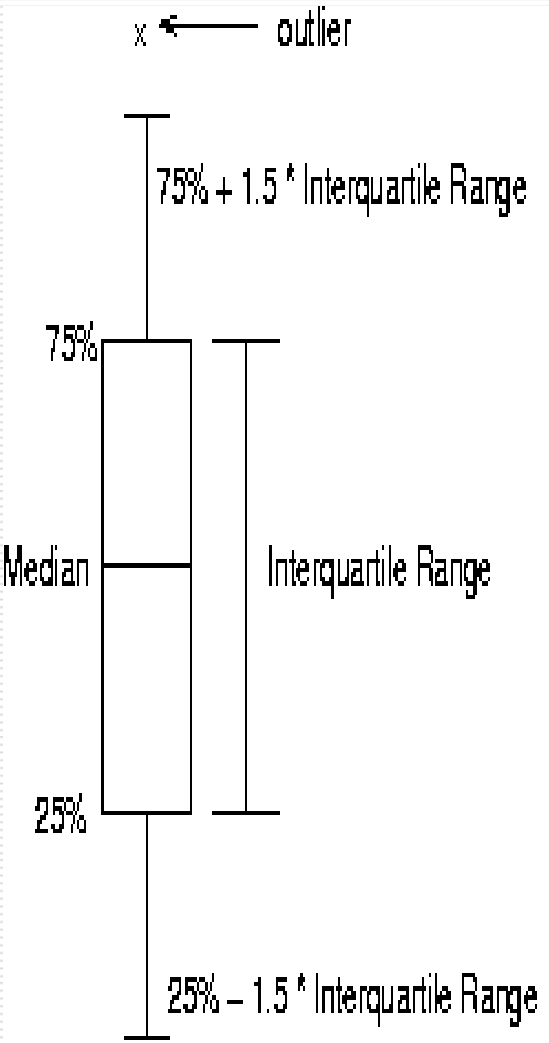
Ukazatele centrální tendence a variability - poznámky

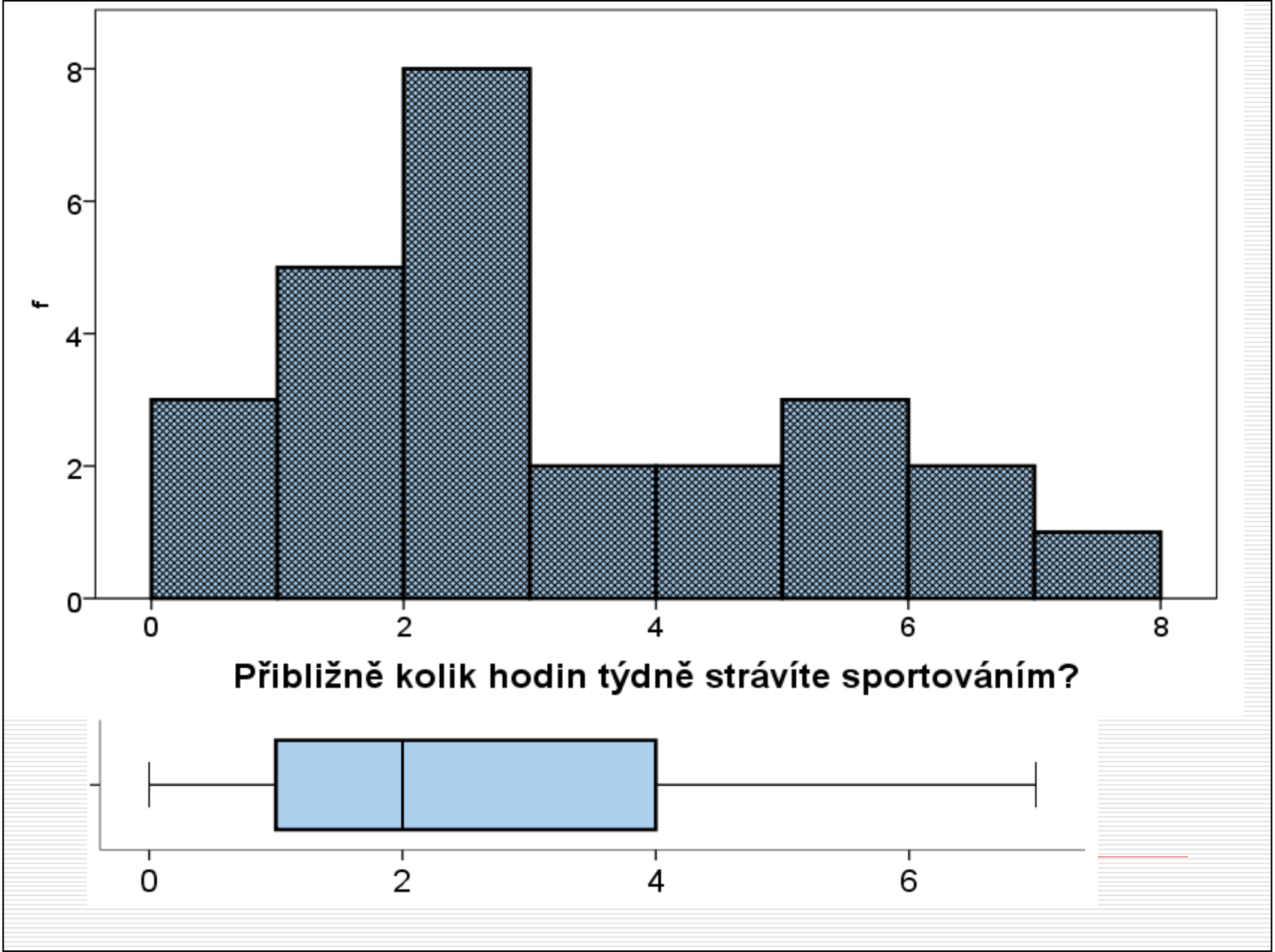
- je třeba je umět spočítat ručně (a zopakovat si práci se sumačním symbolem Σ)
- i vážený průměr
- jak je ovlivní datové transformace přičtení konstanty a násobení konstantou
- vhodnost použití ukazatelů centrální tendence (Hendl s.95)

AJ: weighted mean, add, multiply

Boxplot – krabicový graf s anténami

- krabice je od Q_1 do Q_3
- v krabici se značí medián
- antény jsou X_{min} do X_{max}, **maximálně!** však 1,5x délka krabice (kvartilového rozpětí)
- hodnoty vzdálenější se značí jako body – odlehlé hodnoty
- hodnoty ještě vzdálenější (více než 3x délka krabice od Q_1 nebo Q_3) jsou někdy označovány jako extrémně odlehlé hodnoty




f
8
6
4
2
0
0
2
4
6
8
Přibližně kolik hodin týdně strávíte sportováním?
0
2
4
6

Popis rozložení pomocí percentilů

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

Souhrn

- Kategoriální deskriptivy
  - modus, (entropie)
- Pořadové deskriptivy
  - medián, kvartily, percentily (a jiné *kvant**ily***)
  - kvartilové rozpětí
  - grafické znázornění rozložení pomocí pořadových deskriptiv - **BOXPLOT**
- Odchylkové (deviační), momentové deskriptivy
  - aritmetický průměr
  - rozptyl, směrodatná odchylka (k=2)
  - zešikmení (k=3)
  - špičatost (strmost) (k=4)

$= (\Sigma x^k)/ n$

Volba popisných statistik

- Zvažujeme
  - úroveň měření
  - tvar rozložení – symetrie, normalita
  - cíl studie – pouze popis X usuzování, porovnávání
- Podle komunikačních cílů…
  - Je-li cílem především deskripce dat(=rozložení), pak použijeme **POŘADOVÉ** ukazatele. Připojíme-li i odchylkové, nic nezkazíme.
    - ***N, min, Q_1, Md, Q_3, max***
    - **boxplot**
    - pro individuální skóry **percentily**
  - Je-li cílem další usuzování, porovnávání apod., používáme **ODCHYLKOVÉ** ukazatele … pokud to úroveň měření dovoluje
    - ***N, m, s*** (*N, M, SD*)
    - popis rozložení
    - pro individuální skóry **z-skóry**

Prezentace deskriptiv ve studiích

- **Vždy**! Bez ohledu na to, jak složité statistiky následují.
- Popis rozložení
  - Obvykle se neuvádějí tabulky četností a jejich grafické podoby, pokud ovšem není cílem studie právě statistická deskripce (např. manuál k testu inteligence).
  - Tvar rozložení obvykle podle potřeby zmiňujeme verbálně („přibližně normální, zleva zešikmené…"). Většinou se řeší pouze normalita a odchylky od ní.
- Obvykle pouze pro proměnné, s nimiž pracujeme (interpretujeme…)
  - Minimální triáda: N, m, s (či jejich pořadové ekvivalenty Q_1, Md, Q_3, IQR)
  - Vhodná pětice: N, X_{min}, X_{max}, m, s
  - V případě potřeby: N, X_{min}, X_{max}, m, s, zešikmení, špičatost, zajímavé kvantily
- Obvykle na 2-3 významné číslice (1-2 desetinná místa)
- V českém textu česky, v anglickém anglicky!
  - Pozor na konvence spojené s jazykem: značky, desetinné tečky, chybějící nuly
- Podoba tabulek je podchycena i normami, např. **publikační manuál APA**