PSY117 2016
Statistická analýza dat v psychologii
Přednáška 1

ÚVOD, ČETNOSTI A ROZLOŽENÍ ČETNOSTÍ

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

Andrejs Dunkels; wikiquote

Kostra PSY117 – Statistická analýza dat

- Pochopení základních statistických pojmů
- Použití základních statistických postupů
- Aktivní i pasivní komunikace statistických zjištění

- 2 seminární práce (20b)
- 3 průběžné písemky (3x10b)
- Závěrečný test (50b)

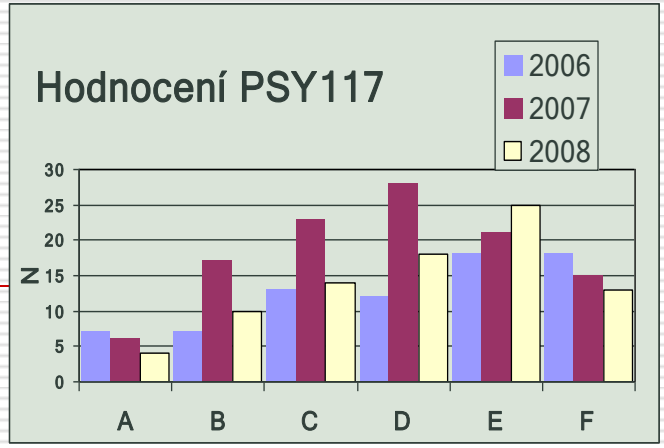
Obtížnost statistiky

Kód	Zapsáno	A	B	C	D	E	F	-
2015	78	7	10	19	18	7	6	9
2014	73	4	6	20	13	11	10	8
2013	98	6	18	16	15	9	16	14
2012	84	8	25	8	12	4	16	9
2011	76	9	11	12	11	4	12	15
2010	81	8	17	12	13	8	11	9

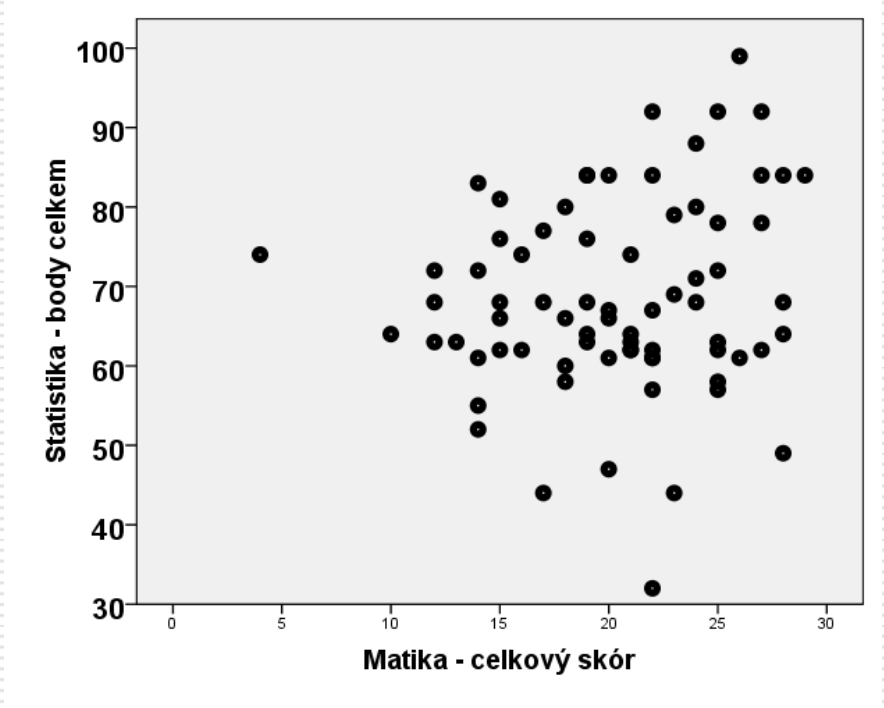
Obtížnost statistiky

Kód	Zapsáno	A	B	C	D	E	F	-
2015	78	9%	13%	24%	23%	9%	8%	12%
2014	73	5%	8%	27%	18%	15%	14%	11%
2013	98	6%	18%	16%	15%	9%	16%	14%
2012	84	10%	30%	10%	14%	5%	19%	11%
2011	76	12%	14%	16%	14%	5%	16%	20%
2010	81	10%	21%	15%	16%	10%	14%	11%

Obtížnost statistiky



- Statistika je obtížná … i pro přírodovědně orientované
- Matematické dovednosti kamenem úrazu nejsou, většinou je máte (r_s=0,13)
- Statistika koreluje s ostatními Áčky – společným jmenovatelem je snaha a obecné předpoklady.



	101	102	103	104	105	106	107	108	112	113	118
r_s	0,36	0,53	0,52	0,59	0,51	0,53	0,56	0,49	0,42	0,33	0,36

Jak se učit statistiku

- S. = lehká matematika, těžké myšlení
- …jako cizí jazyk
  - po malých kouscích, pravidelně
  - pozor na slovíčka
  - prakticky: tužka-papír-kalkulačka + počítač (Excel, SPSS, Statistica…)
- Neexistuje dobrá učebnice v češtině
  - Hendl – i ve čtvrtém vydání žádná cvičení, obtížně stravitelný text
  - zbývá angličtina: např. **Howell**; Howitt&Cramer; Glass&Hopkins, Field
  - web: wiki, statsoft.com
- …sám i společně
  - diskuzní fórum FB: http://goo.gl/Mt95eT
  - poskytovna: sdílení materiálů

KEEP
CALM
AND
STUDY
STATISTICS

Co je to vlastně statistika?

- **Popis** získaných **dat** o **jevech**, které se vyskytují ve větších množstvích
  - Popis **proměnných**: jaké podoby jevu, jak časté?
  - Popis **vztahů** mezi proměnnými/jevy

- Statistické **usuzování** ze vzorku na populaci
  - Pravděpodobnostní usuzování
  - Konfrontace očekávání (modelů) se získanými daty
  - Testování hypotéz

AJ: description, data, variable, statistical inference, population

K čemu je statistika jako taková?

- Formalizované **zpracování zkušenosti**, když
  - počet zkušeností, výskytů jevu přesáhne 7±2 (automat)
  - hledané je malé (mikroskop)
  - záludnosti naší kognice představují problém (zvl. paměť)
- Motivuje vytváření záznamů o zkušenosti (a.k.a. dat)
- „Objektivní“ (=v komunitě srozumitelný) popis výskytu jevů
- Hledání společného, typického, normálního i jedinečného
- Hledání vztahů, souvislostí mezi jevy
- Trénuje myšlení
  - kritické myšlení, modely vzniku jevů
  - myšlení o variabilitě jevů (≈rozdílech mezi lidmi)
  - uvědomění si všudypřítomnosti chyby měření (vnímání)
  - **pravděpodobnostní myšlení**

K čemu je statistika psychologům?

1. V běžném životě – statistická gramotnost (literacy)

2. Ve výzkumu
   - hledání pravidelností + identifikace jedinců, kteří se těmto pravidelnostem vzdalují

3. V aplikovaných disciplínách a praxi
   - formalizovaná reflexe praxe - zjišťování efektů, výsledků – co se mi osvědčuje a co ne?

4. Při diagnostice, poznávání lidí
   - diagnostické metody mají statistické základy
   - statistické pojetí normality a odchylky od ní
   - pravděpodobnost správného určení diagnózy

Malá mapa semestru

- Jaké hodnoty(podoby jevu) se vyskytují a jak často?
  - Je v tom nějaká pravidelnost?
- Existuje souvislost mezi výskytem jednoho jevu a výskytem nějakého jiného?
  - Dokážeme z existence jednoho jevu usuzovat na ten druhý?
- Jak velké zkreslení asi vzniklo tím, že máme data jen o zlomku všech výskytů zkoumaného jevu?

12,08	1	2	2	15	11	5	1
12,58	1	1	1	24	13	4	1
11,92	1	2	2	7	13	6	2
12,33	1	2	2	10	17	4	2
12,08	1	1	1	7	13	6	1
11,92	1	2	2	10	11	4	1
12,67	1	2	1	16	11	3	1
12,08	1	2	2	7	1	6	1
12,25	1	1	1	24	11	4	1
12,67	1	1	2	6	1	6	1
12,08	1	2	2	7	10	4	2
12,67	1	1	2	10	17	6	1

Data, proměnné

- Data vznikají měřením(záznamem) jevů
- Data mají obvykle podobu proměnných
  - Proměnné vznikají(jsou) **kódováním dat**
  - Z jedněch dat můžeme udělat více proměnných
- Proměnné reprezentují *znaky, charakteristiky, atributy, vlastnosti* zkoumaných jevů či objektů, popř. jejich kombinace
- Proměnné nabývají různých hodnot, pokud ne, jsou to **konstanty**

AJ: data, measurement, variable, coding, value, constant

Data, proměnné

- Data vznikají měřením(záznamem) jevů

Měření: Standardizovaný postup, procedura

Procedura, kt. dává číslům smysl

Tato procedura je **vždy** zatížena chybou

Někdy je měření prostý **záznam**

Data, proměnné

- Data vznikají měřením(záznamem) jevů
- Data mají obvykle podobu proměnných
  - Proměnné vznikají(jsou) **kódováním dat**
  - Z jedněch dat můžeme udělat více proměnných
- Proměnné reprezentují *znaky, charakteristiky, atributy, vlastnosti* zkoumaných jevů či objektů, popř. jejich kombinace
- Proměnné nabývají různých hodnot, pokud ne, jsou to **konstanty**

AJ: data, measurement, variable, coding, value, constant

věk	národnost	mat	cj	pr_oblib	pr_neobl	ocek_vzd	stav_r
12,08	1	2	2	15	11	5	1
12,58	1	1	1	24	13	4	1
11,92	1	2	2	7	13	6	2
12,33	1	2	2	10	17	4	2
12,08	1	1	1	7	13	6	1
11,92	1	2	2	10	11	4	1
12,67	1	2	1	16	11	3	1
12,08	1	2	2	7	1	6	1
12,25	1	1	1	24	11	4	1
12,67	1	1	2	6	1	6	1
12,08	1	2	2	7	10	4	2
12,67	1	1	2	10	17	6	1

Co ta čísla-kódy znamenají?
Úrovně měření (typy měřítka, škály)

Úroveň	***Operace***	***Příklady***
1 Nominální	= ≠	pohlaví, tramvaj, preference
2 Ordinální (pořadová)	= ≠ > <	známky, souhlasení
3 Intervalová	= ≠ > < + −	°C, IQ, „dobré“ psychotesty
4 Poměrová	= ≠ > < + − × ÷	K, váha, počty, frekvence

1+2: kategorické, 3+4: metrické, kardinální;

Howitt&Cramer: nominal category data (1) vs score data (2-4)

Více viz extrakt z Urbánek, Denglerová, Širůček v ISu

AJ: level of measurement, nominal, ordinal, interval, ratio

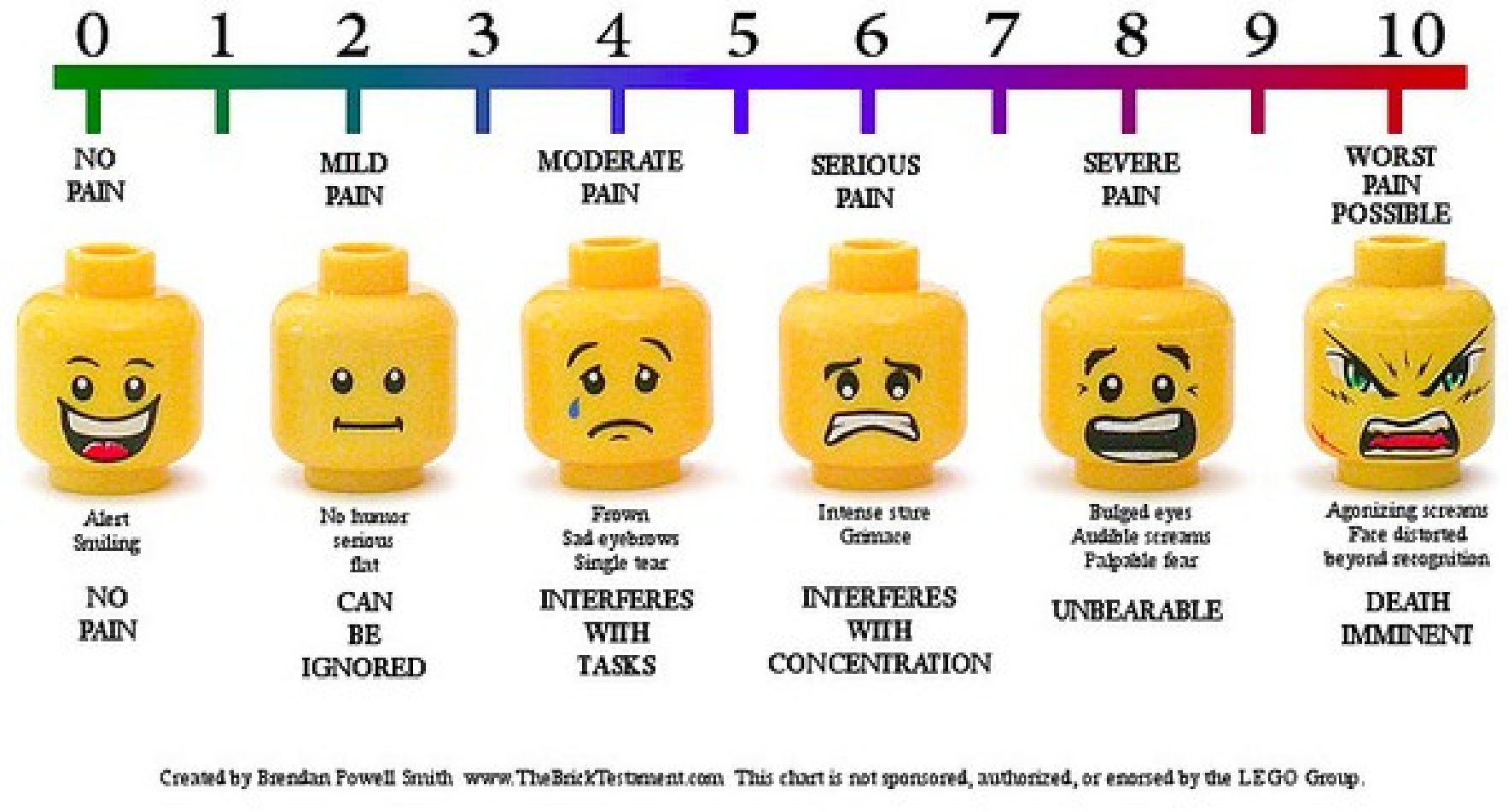
Typy proměnných podle počtu možných hodnot

- **Spojité** proměnné
  - Nekonečně mnoho hodnot – reálná čísla
- **Diskrétní** proměnné
  - [Nekonečně] mnoho hodnot, jen některá (typicky celá) čísla – často se k nim chováme jako ke spojitým
  - Nemnoho hodnot
    - jen 2 možné hodnoty: **dichotomické** (alternativní)
    - „pár“ možných hodnot: **polytomické**

AJ: discrete, continuous, dichotomous, alternative, polytomous

Zacházení s proměnnými podle jejich typu

- Reálné proměnné na ideální typy často přesně nepasují
- Rozlišujeme měřenou charakteristiku a škálu, pomocí které byla změřena
  - Často je v psychologii charakteristika uvažována jako intervalová spojitá proměnná, kterou měříme diskrétní polytomickou škálou
  - Př. Postoj
- Hledáme argumenty pro to, abychom mohli škálu považovat za intervalovou – jednodušší statistiky, více informace, riziko zkreslení.
  - Flexibilní, argumentující, opatrný přístup – žádné dogma.



Shrnutí

- Při hledání odpovědí na otázky a řešení problémů je užitečné využít data – psychologie jako empirická věda
- I při reflexi vlastních zkušeností je užitečné nespoléhat jen na paměť
- Každá statistika má smysl jen jako podklad pro odpověď na určitou otázku – ne sama o sobě – a v kontextu této otázky má smysl ji i komunikovat
- Tyto principy jsou užitečné stejně občanovi jako psychologovi i jako výzkumníkovi v psychologii
- Data tvoříme (my nebo někdo jiný) a tomu, co potřebujeme vědět, odpovídají vždy nedokonale
- Tvoříme různé typy dat, pro které máme různé statistiky – kategorie vs. škály

Máme data

„účetnictví“ může začít

Jaké hodnoty máme v datech?

- Jaké hodnoty proměnné/ých se v datech vyskytují? – *třídění, kódování*
  - 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?

Tabulka četností (frekvencí)

hodnota/ interval	**(absolutní) četnost**		**relativní četn. (%)**	**kumulativní rel. č.**
Minimum / interval1				
Hodnota2 / interval2				
…				
Maximum / posl. interv.				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
- Tabulka četností zobrazuje téměř všechna data
  - Použitím intervalů již data mírně redukujeme
- Minimální podoba tabulky četností: **absolutní a relativní četnosti, součtový poslední řádek**

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

Věk	Interval			kumulativní	relativní	kumulativní relativní
2	Dolní mez	Horní mez	Četnost	četnost	četnost %	četnost *cum %*
3			n_i	N_i	f_i	F_i
5	0	20	8	8	40	40
5	21	40	1	9	5	45
7	41	60	4	13	20	65
16	61	80	5	18	25	90
17	81	100	2	20	10	100
19	Celkem		20		100	

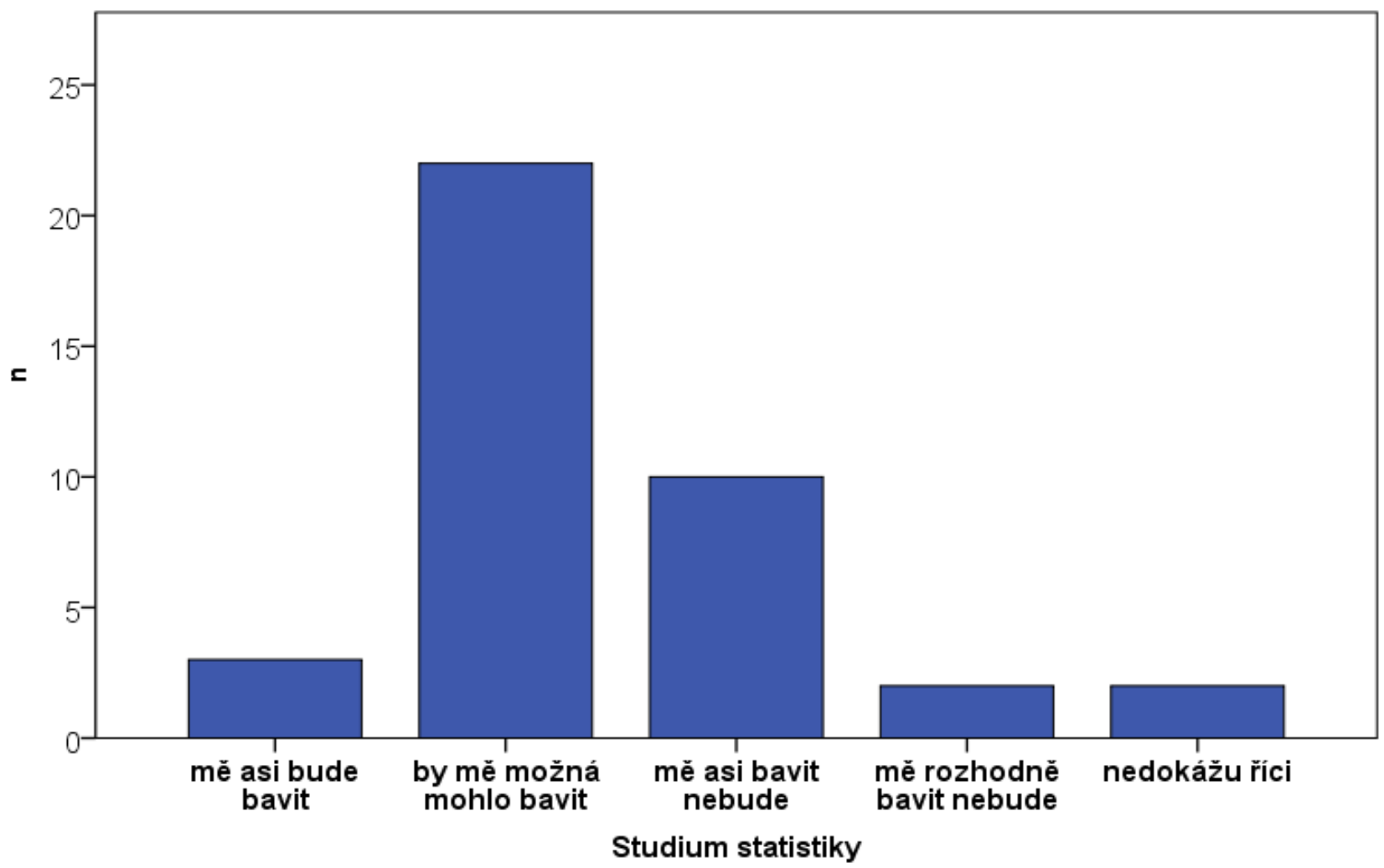
Též Datíčka.xls, list „četnosti“.

Grafické podoby tabulky četností

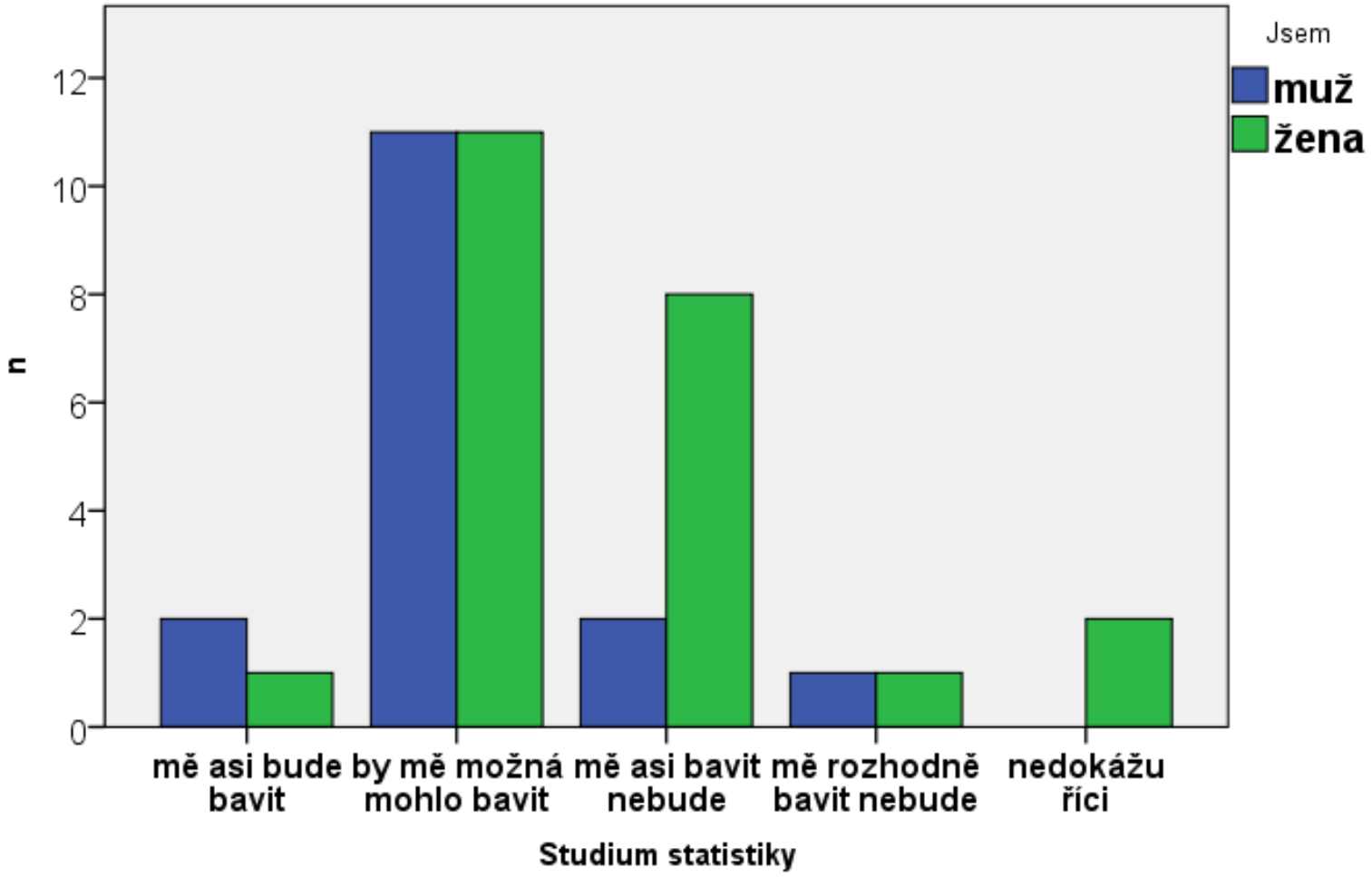
- Kategorické proměnné
  - sloupcový graf (diagram)
  - koláčový diagram – zřídkakdy, neukazuje rozložení
- Metrické proměnné
  - Histogram – jako sloupcový, ale šíře sloupců reprezentuje šíři intervalů
  - stem-and-leaf – rozdělení hodnot do intervalů

AJ: bar chart, histogram, pie chart, frequency distribution, stem-and-leaf plot

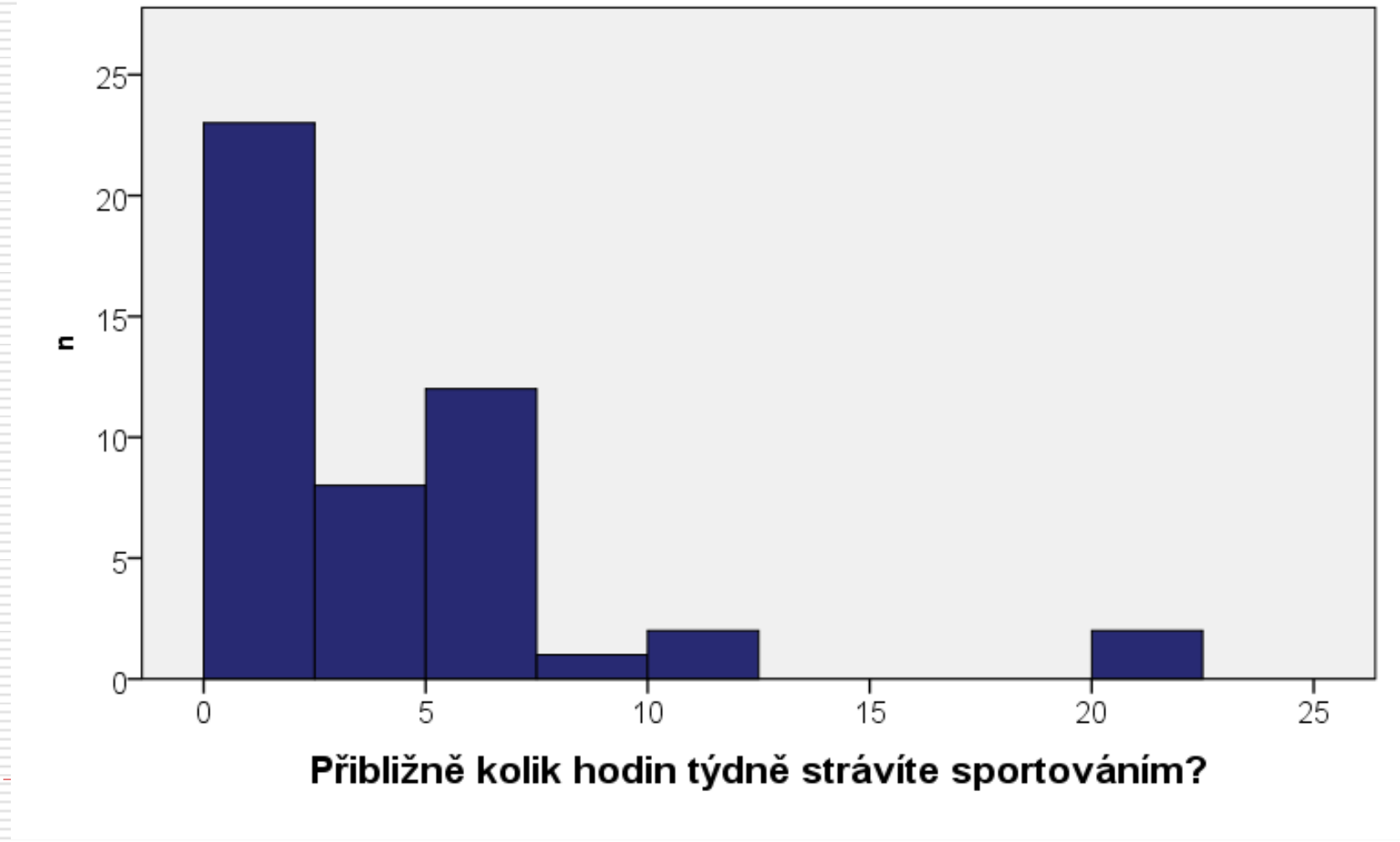
Sloupcový diagram

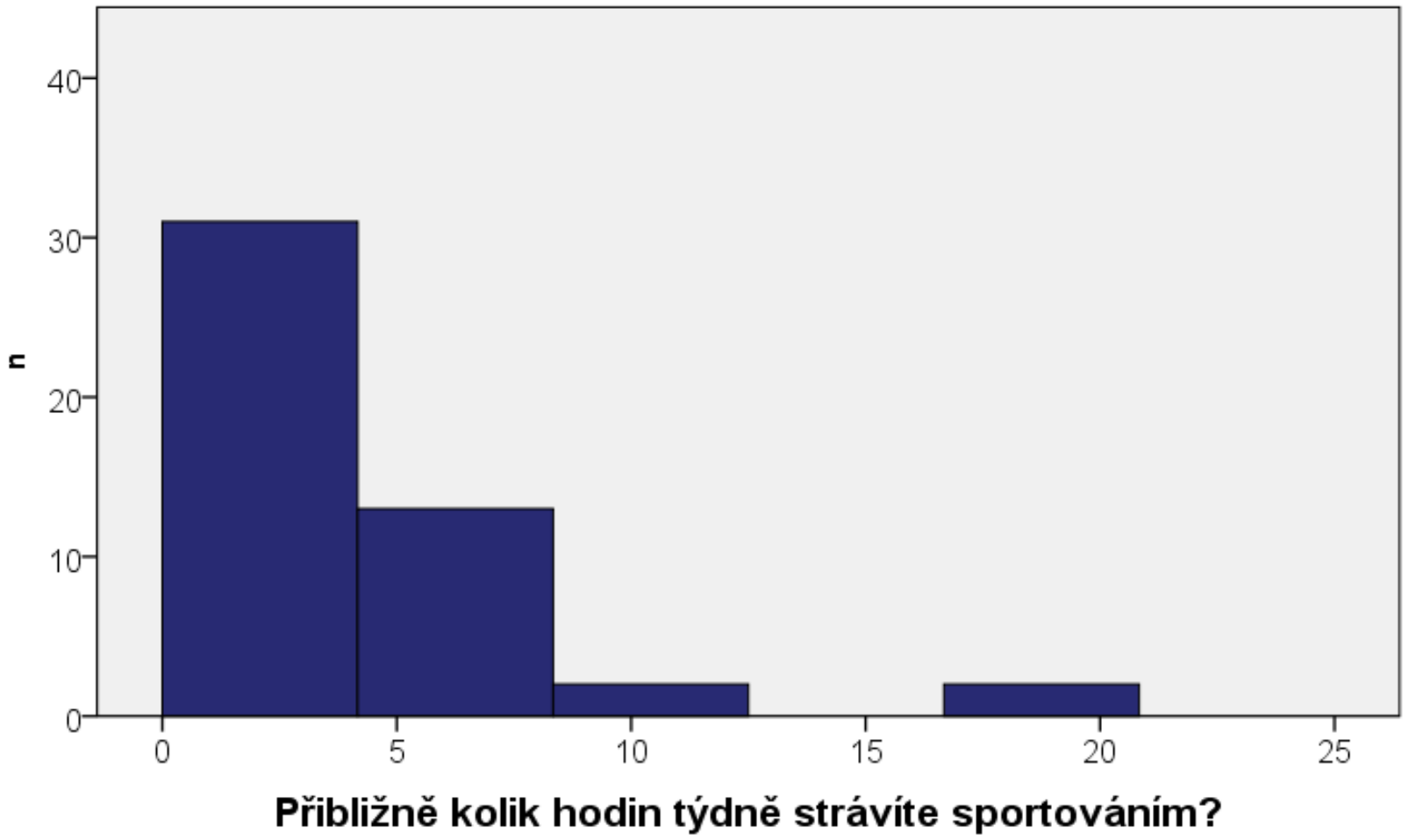


Sloupcový diagram s tříděním

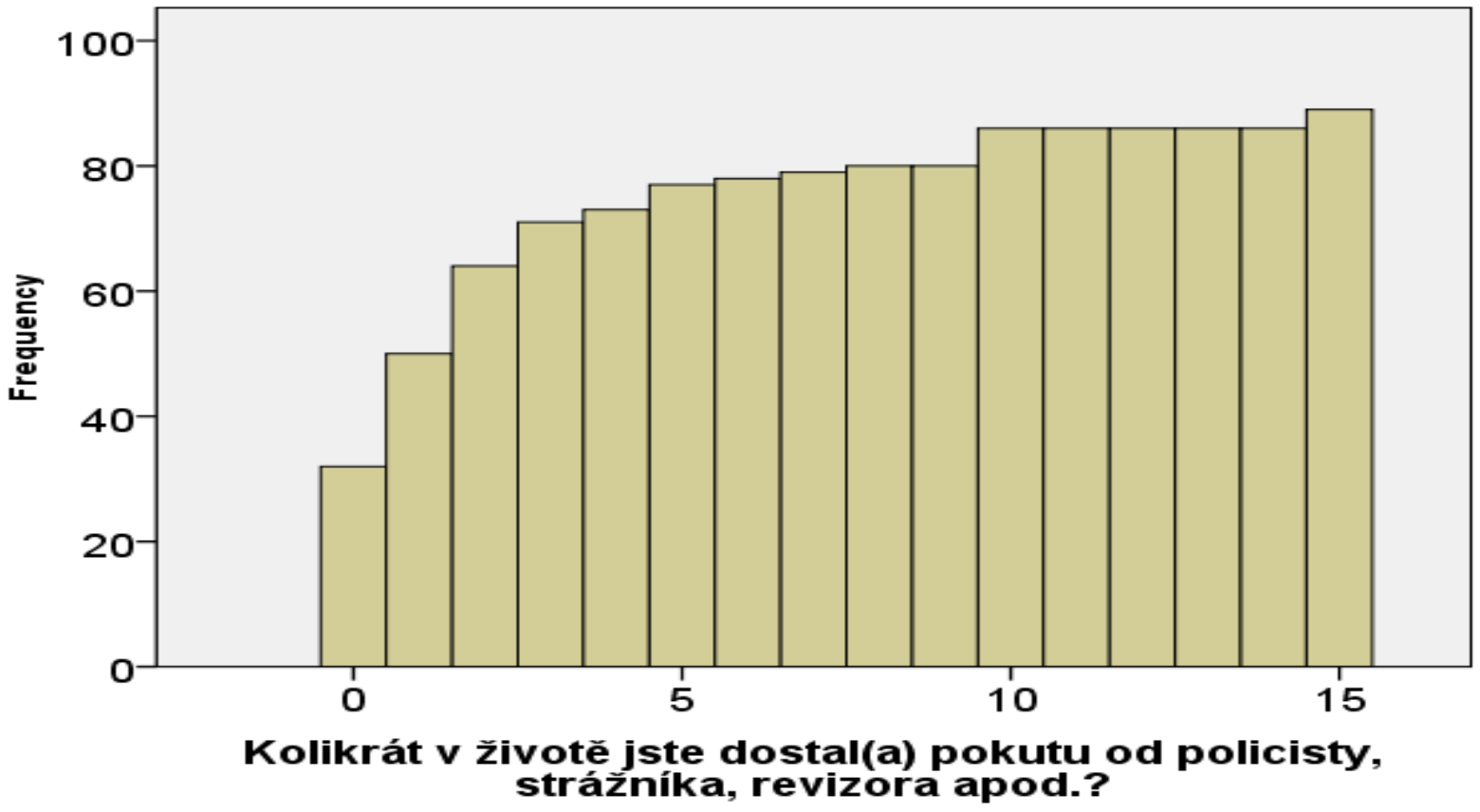


?





Kumulativní histogram



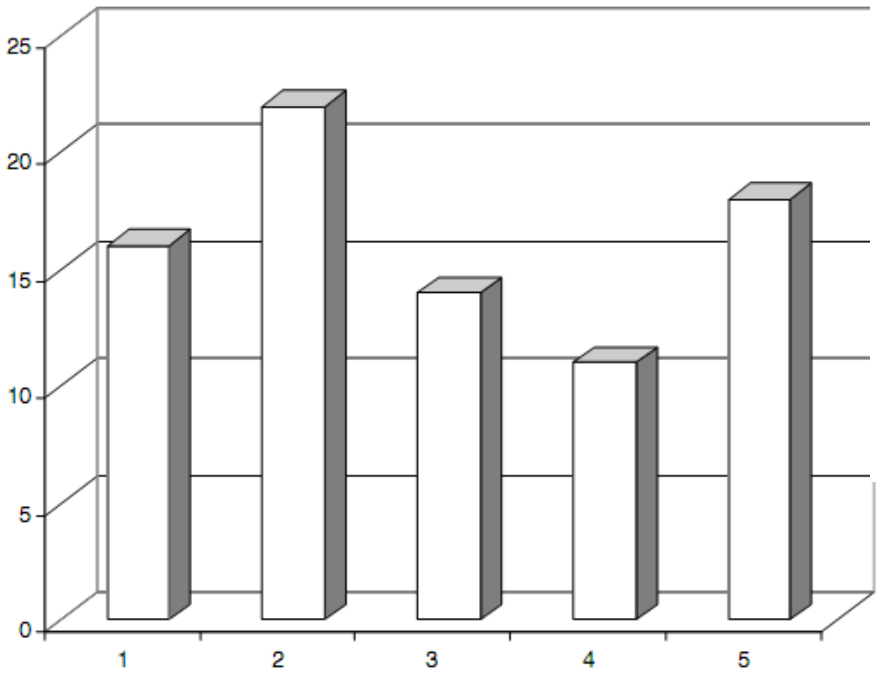
Číslicový histogram „stonek a list“

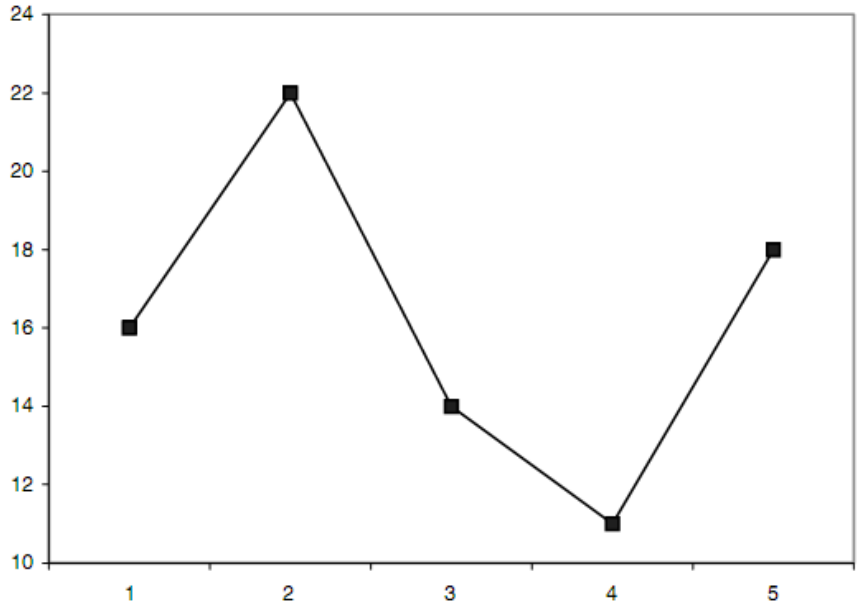
```
Frequency    Stem &  Leaf
   32,00        0 .  00000000000000000000000000000000
   18,00        1 .  0000000000000000
   14,00        2 .  00000000000000
    7,00        3 .  0000000
    2,00        4 .  00
    4,00        5 .  0000
    1,00        6 .  0
    1,00        7 .  0
   10,00 Extremes    (>=8,0)

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

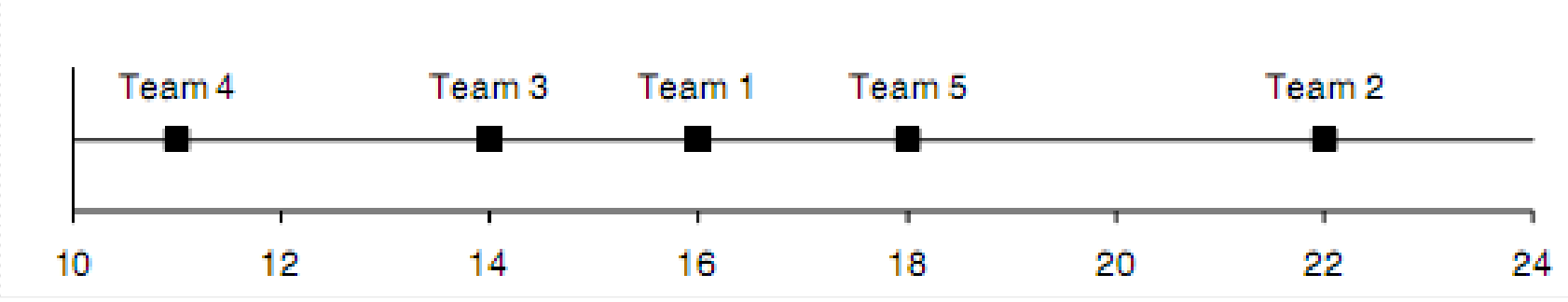
```
Větev°list  (jed. list1,000000, např. 6°5 = 6,50000
   0° 00000000000000000000000000000000        ·
   1° 0000000000000000  ·          ·          ·
   2° 00000000000000    ·          ·          ·
   3° 0000000   ·       ·          ·          ·
   4° 00        ·       ·          ·          ·
   5° 0000      ·       ·          ·          ·
   6° 0         ·       ·          ·          ·
   7° 0         ·       ·          ·          ·
   8° 0         ·       ·          ·          ·
   9°           ·       ·          ·          ·
  10° 000000    ·       ·          ·          ·
  11°           ·       ·          ·          ·
  12°           ·       ·          ·          ·
  13°           ·       ·          ·          ·
  14°           ·       ·          ·          ·
  15° 000       ·       ·          ·          ·
min = 0,000000    max = 15,00000             Celk.
```

„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
  - 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
- Další
  - Hans Rosling na TEDu: http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html
  - Nathan Yau: Visualise this… http://www.amazon.com/o/ASIN/0470944889?tag=adapas02-20
  - **Howitt & Cramer s. 21**

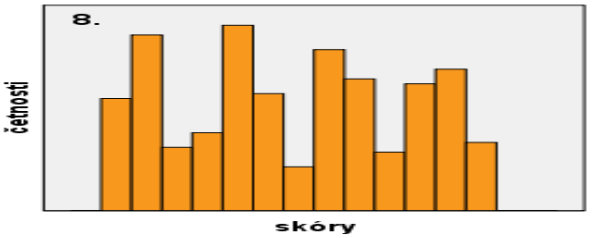


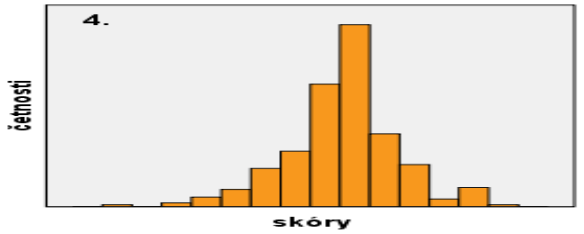




Rozložení *rozdělení, distribuce* četností

- 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.

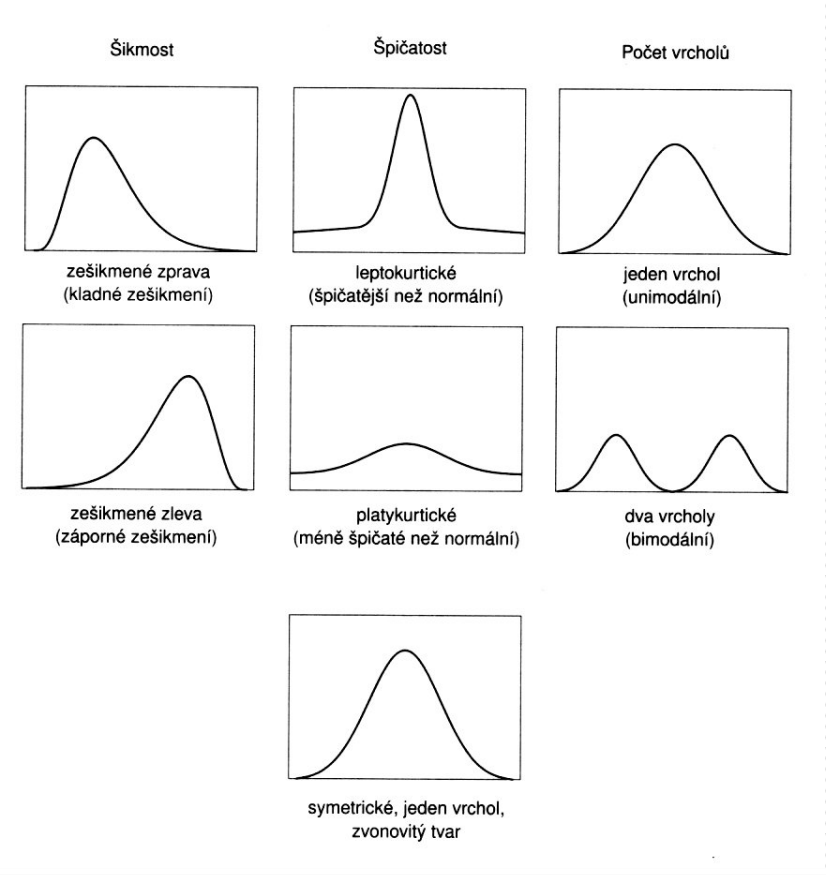




- **Rozdělení** četností je tedy to, kolik relativně (či absolutně) máme kterých hodnot měřené proměnné.
  - Typicky lze přibližně popsat slovy, např.: vyskytlo se hodně středních hodnot a relativně málo extrémních hodnot.
  - Toto **rozložení** jevů na měřené škále je nejlépe vidět na grafech.
  - Obvykle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

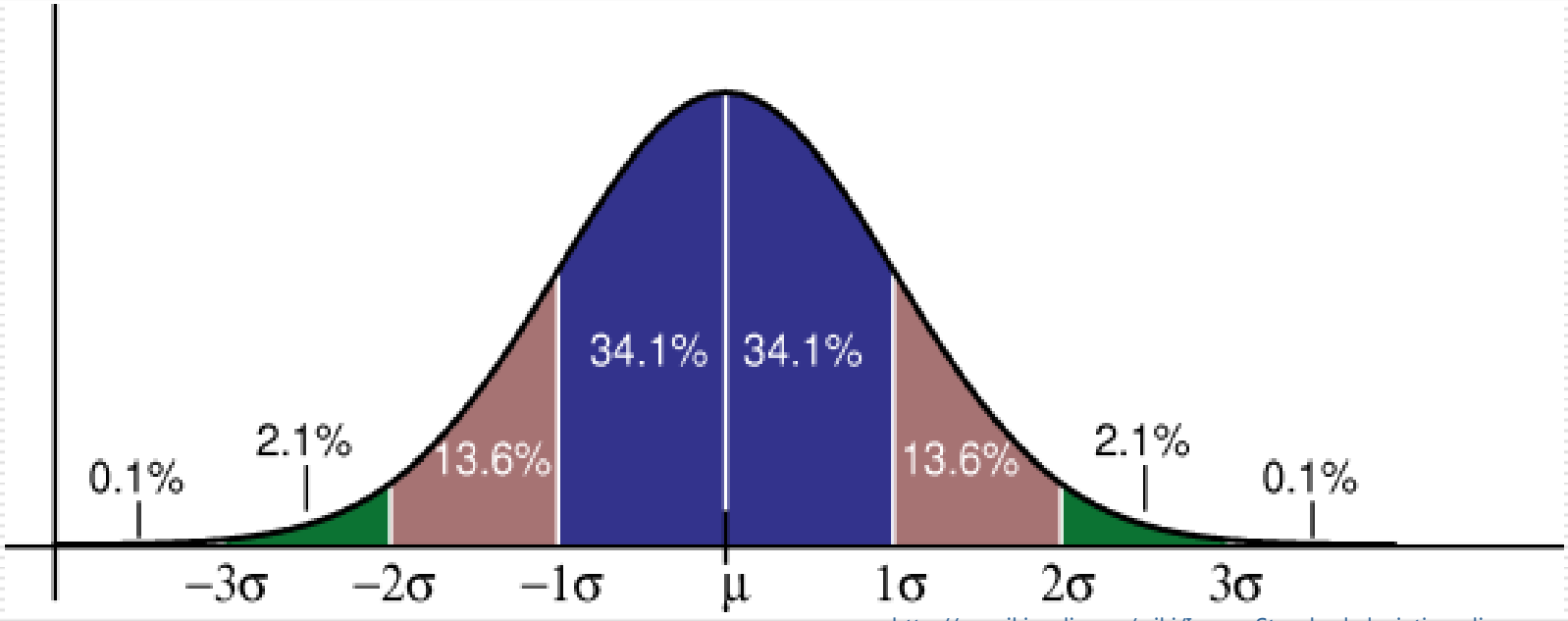
Tvar rozložení četností

- 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

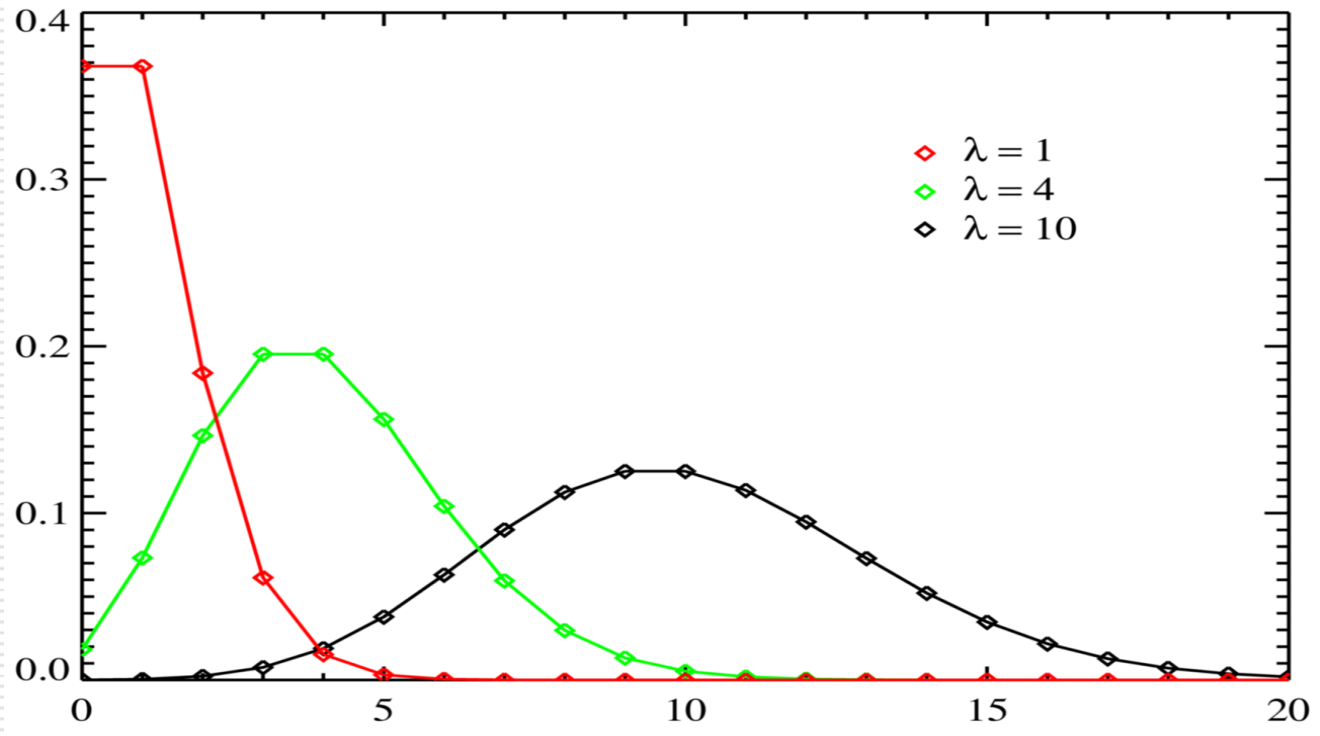


http://en.wikipedia.org/wiki/Image:Standard_deviation_diagram.png

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

AJ: Poisson distribution

Rozložení

- Známky ze statistiky
- Výška studentů psychologie
- Depresivita
- Postoje k interrupcím
- Spokojenost se studiem
- Pohlaví na psychologii
- Počet návštěv u lékaře

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**.