PSY117 2019
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

Vyučující



Kostra PSY117 – Statistická analýza dat

- Pochopení základních statistických pojmů a myšlenek – statistická gramotnost
- Použití základních statistických postupů
- Aktivní i pasivní komunikace statistických zjištění

1 seminární práce (10b)

3 průběžné písemky (3x10b)

Závěrečný test (50b)

Obtížnost statistiky

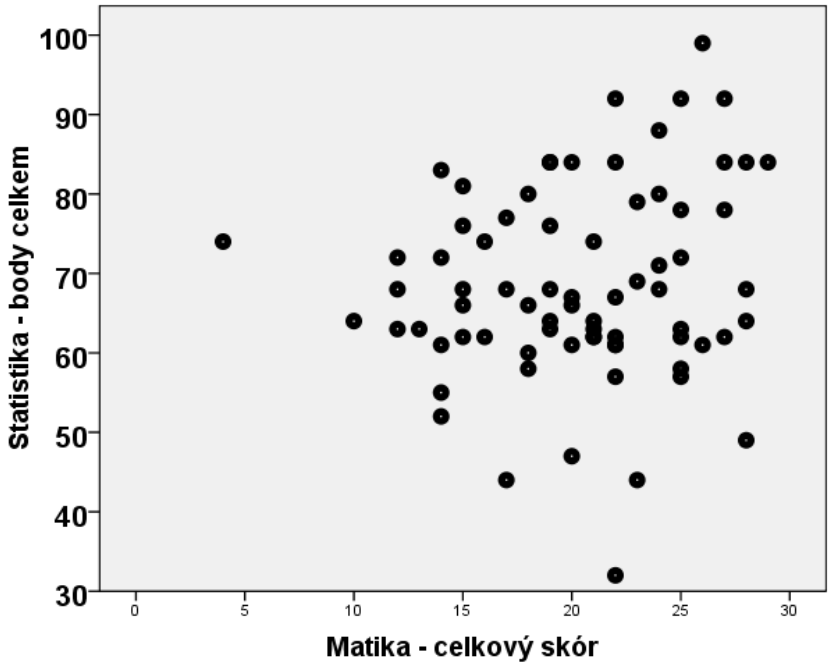
Rok	Zapsáno	A	B	C	D	E	F	-
2018	102	7	19	18	13	10	15	17
2017	99	6	18	24	17	11	13	7
2016	86	2	15	17	15	12	13	7
2015	78	7	10	19	18	7	6	9
2014	73	4	6	20	13	11	10	8
2013	98	6	18	16	15	9	16	14
2012	84	8	25	8	12	4	16	9
2011	76	9	11	12	11	4	12	15
2010	81	8	17	12	13	8	11	9

Obtížnost statistiky

Rok	Zapsáno	A	B	C	D	E	F	-
2018	102	7%	19%	18%	13%	10%	15%	17%
2017	99	6%	18%	24%	17%	11%	13%	7%
2016	86	2%	15%	17%	15%	12%	13%	7%
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

- 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, jamovi, R…)
- Nemáme dobrou učebnici v češtině
  - Hendl – i ve čtvrtém vydání žádná cvičení, obtížně stravitelný text
  - zbývá angličtina: **Howell**; Howitt&Cramer; Glass&Hopkins, Field
  - web: wiki, statsoft.com, Coursera, Khan Academy….
- …sám i společně
  - diskuzní fórum FB: http://goo.gl/Mt95eT
  - poskytovna: sdílení materiálů
  - MU Math and Stats Support Centre http://mathstat.econ.muni.cz/

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 **souvislostí** mezi proměnnými/jevy – *závislé* a *nezávislé* proměnné

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

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

Data? Jaká data?

O charakteristikách účastníků výzkumu.

O výkonech žáků ve škole

O vybraném vzorku lidí, s nimiž budeme srovnávat své klienty

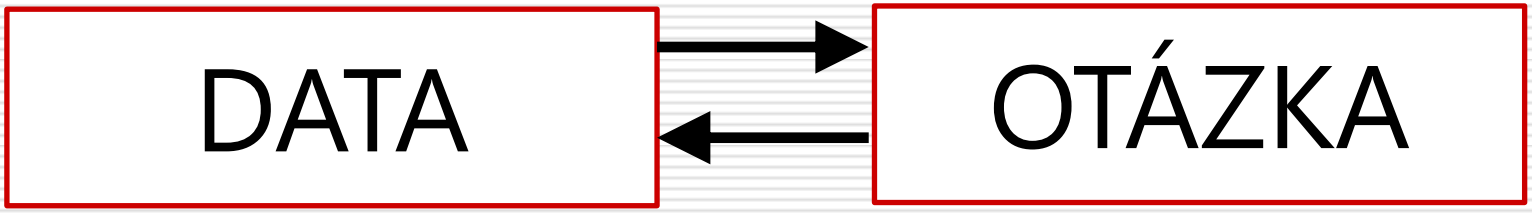
O jednom klientovi a jeho změnách v čase

O sobě, svém životě, nebo praxi

Záznamy, pozorování, testování, vzpomínky…

Záměrně vytvářená data pro výzkum

K čemu data?



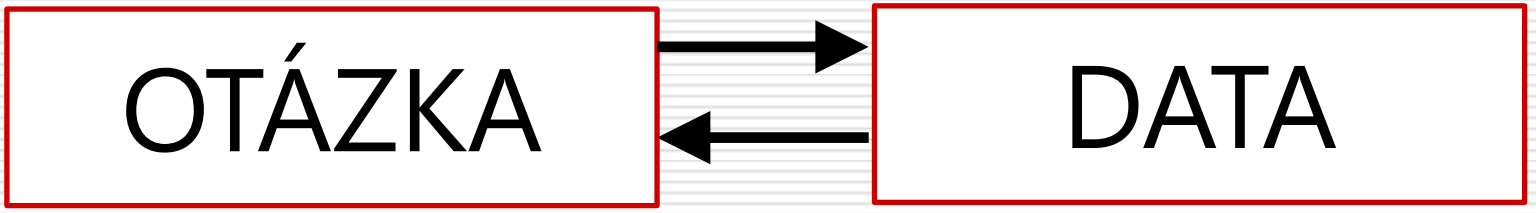
Známky.
Počet spotřebovaných termínů.
Čas strávený učením.
Hrdost při zvládnutí…

Jak náročná je PSY117?

Výkony (známky, body…)
+
IQ, známky ze SŠ, seberegulace
Využití zdrojů
Vynaložené úsilí

Jak zvládnout PSY117?

K čemu data?



Kladení (si) otázek je základní prvkem smysluplnosti dat.

Data nesou (omylnou) informaci o jevech, o kterých si klademe otázky.

Analýzou dat se snažíme zpracovat informaci obsaženou v datech tak, abychom získali podklad pro odpověď na svou otázku.

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, analýz)
- „Objektivní“ (=v komunitě srozumitelný) popis výskytu jevů
- Hledání společného, typického, normálního i jedinečného, odlišného
- 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 – chyba měření
   - 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?

0	2	2	2	4	11	3	10
10	5	3	1	3	4	3	2
4	3	0	1	4	1	4	2
5	24	3	0	0	0	2	6
5	5	1	4	7	4	4	2
2	2	3	3	4	7	4	2
5	14	0	10	1	10	15	

Data, proměnné

- **Data** vznikají (standardizovaným) záznamem jevů
- **Data** se člení do **proměnných**
- **Proměnné** reprezentují jednotlivé *znaky, charakteristiky, atributy, vlastnosti* zkoumaných jevů či objektů, popř. jejich kombinace
  - Proměnné vznikají **kódováním hrubých dat**
  - Z jedněch dat můžeme udělat více proměnných
- **Proměnné** nabývají různých hodnot, pokud ne, jsou to **konstanty**

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

Odpovědi 54 lidí na otázku:
Přibližně kolik hodin týdně strávíte sportováním?

0	2	2	2	4	11	3	10
10	5	3	1	3	4	3	2
4	3	0	1	4	1	4	2
5	24	3	0	0	0	2	6
5	5	1	4	7	4	4	2
2	2	3	3	4	7	4	2
5	14	0	10	1	10	15	

Co ta čísla-kódy-hodnoty 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;

Howell: categorical(qualitative) data vs. measurement (quantitative) data

Více viz extrakt z Urbánek, Denglerová, Širůček v ISu

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

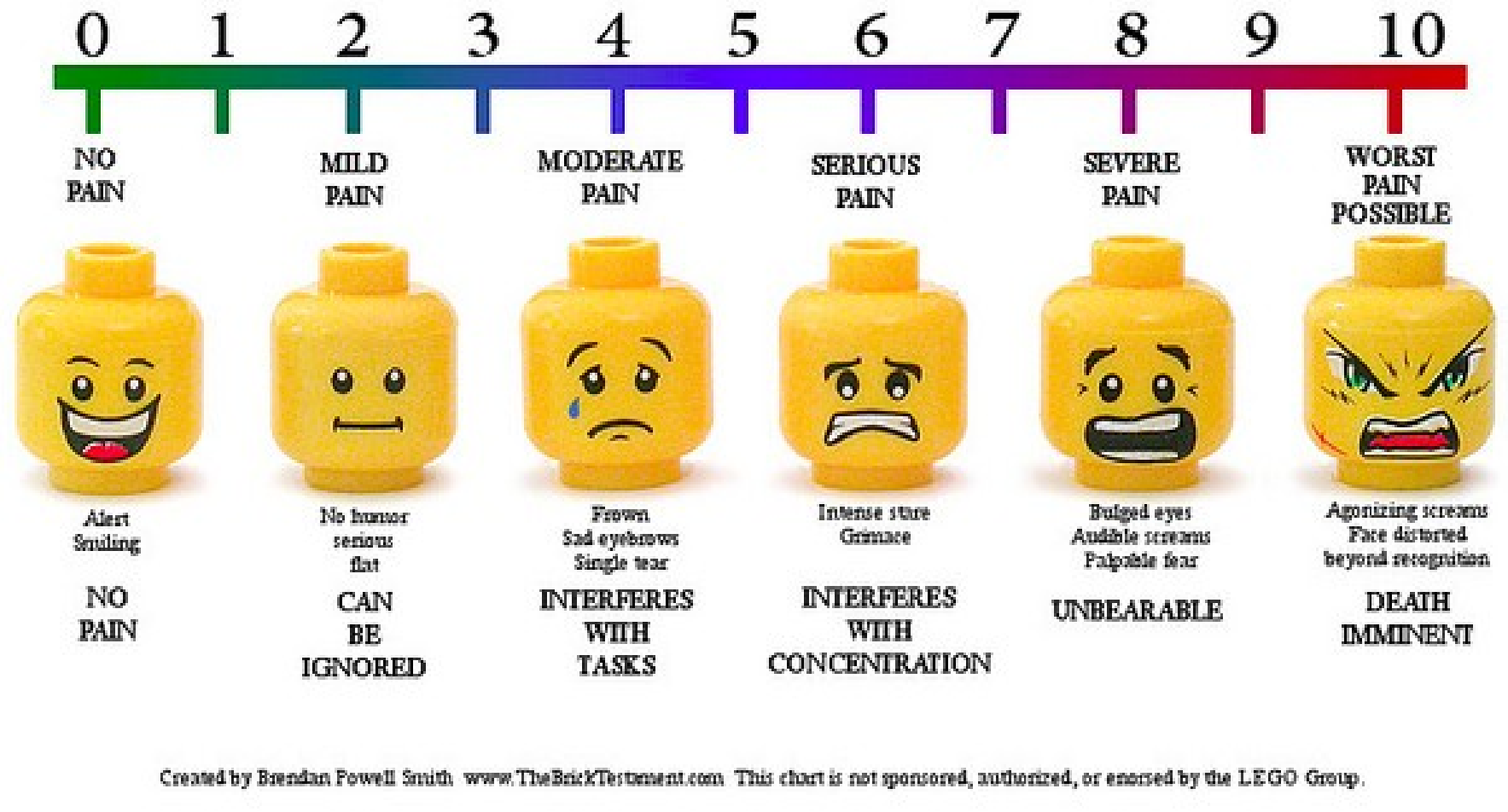
Typy proměnných podle počtu možných hodnot

- **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

Usuzování na úroveň měření v praxi

- *Úroveň měření* je ideál.
- U skutečné proměnné *argumentovaně* **předpokládáme** úroveň měření, **považujeme ji za** N/O/I/P
- Rozlišujeme měřenou charakteristiku a škálu, pomocí které byla změřena
  - Často je v psychologii charakteristika uvažována jako intervalová spojitá proměnná, kterou měříme diskrétní nebo dokonce polytomickou škálou
  - př. postoj
- Často hledáme argumenty, abychom mohli škálu považovat za I/P – jednodušší statistiky, více informace, (-) riziko zkreslení.
  - Flexibilní, argumentující, opatrný přístup – žádné dogma.
  - Více detailů v psychometrice



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
  - Intervaly jsou obvykle zprava uzavřené. např. (0,1)
- Tabulka četností zobrazuje téměř všechna data
  - Použitím intervalů již data mírně redukujeme
- Minimální podoba tabulky četností: **abs. a rel. četnosti, součtový poslední řádek**

„Ruční“ tvorba tabulky četností

1. Seřazení hodnot (od nejmenší do největší)
2. Rozhodnutí o rozdělení na intervaly (I, P)
3. Spočítání abs. četností hodnot/intervalů
4. Spočítání relativních četností
5. Spočítání kumulativních četností
6. Spočítání kumulativních relativních četností

Hodin					
30	Popisky řádků	f	%	cum %	
7	0	2	2,82%	2,82%	
1	1	7	9,86%	12,68%	
5	2	9	12,68%	25,35%	
8	3	7	9,86%	35,21%	
16	4	6	8,45%	43,66%	
6	5	16	22,54%	66,20%	
8	6	1	1,41%	67,61%	
1	7	3	4,23%	71,83%	
1	8	7	9,86%	81,69%	
4	9	1	1,41%	83,10%	
2	10	7	9,86%	92,96%	
36	16	1	1,41%	94,37%	
2	21	1	1,41%	95,77%	

Hodin		Interval			kumulativní		kumulativ
30		Dolní mez	Horní mez	Četnost	četnost	relativní četnost	relativní četn
7				*f*	*cum f*	*%*	*cum %*
1		0	5	47	47	66,2	
5		5	10	19	66	26,8	
8		10	15	0	66	0,0	
16		15	20	1	67	1,4	
6		20	100	4	71	5,6	1
8		Celkem		71		100	
1							
1							
4							
2							
36							
2							
8							
5							
5							
2							

Poklikáním se tabulka otevře jako editovatelný objekt v Excelu.

Též Datíčka.xls, list „četnosti“.

Grafické podoby tabulky četností

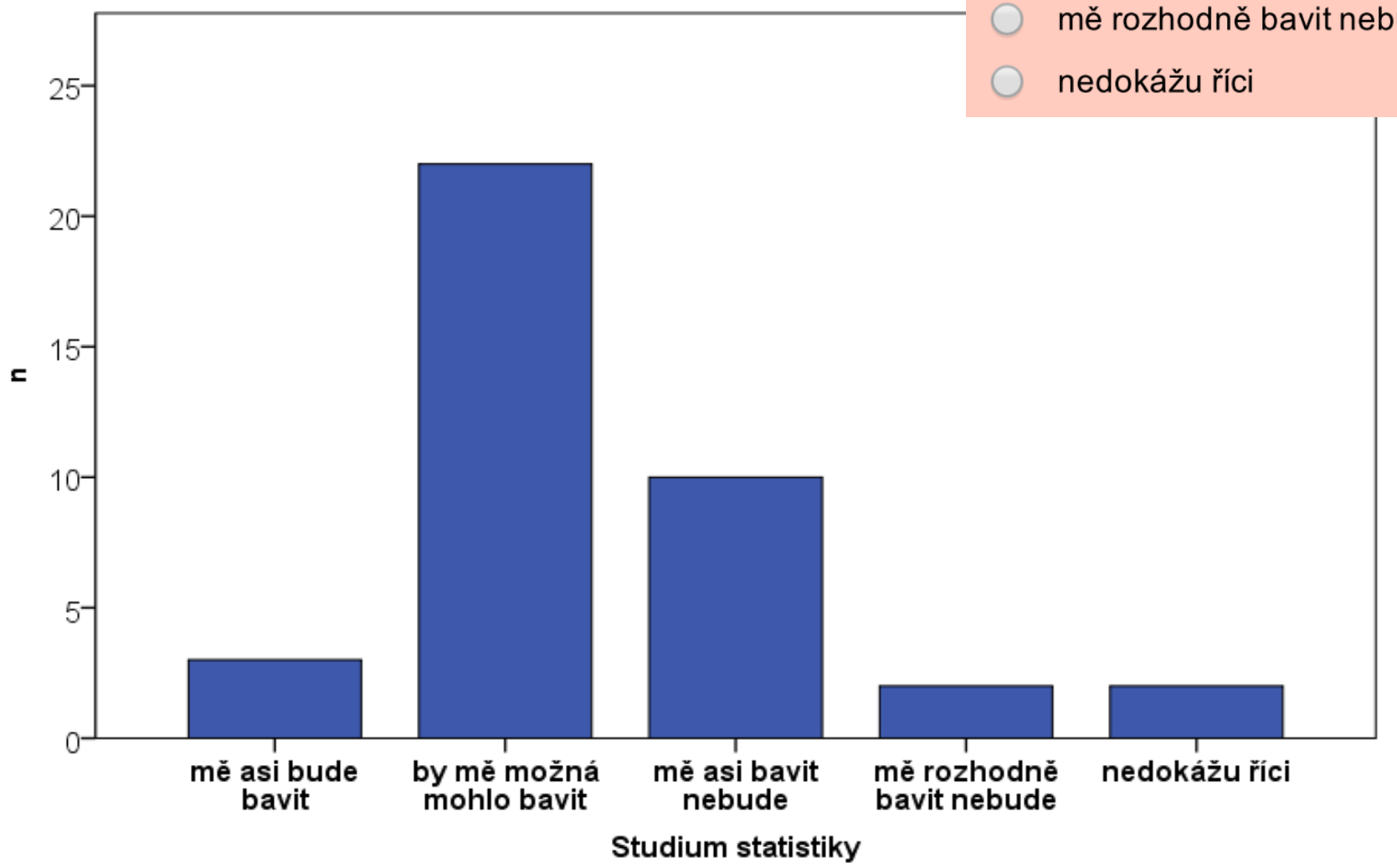
- Kategorické proměnné
  - sloupcový graf (diagram)
  - koláčový diagram – zřídkakdy, neukazuje rozložení
- Metrické proměnné
  - Histogram – jako sloupcový, ale šíře sloupců reprezentuje šíři intervalů
  - stem-and-leaf – rozdělení hodnot do intervalů

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

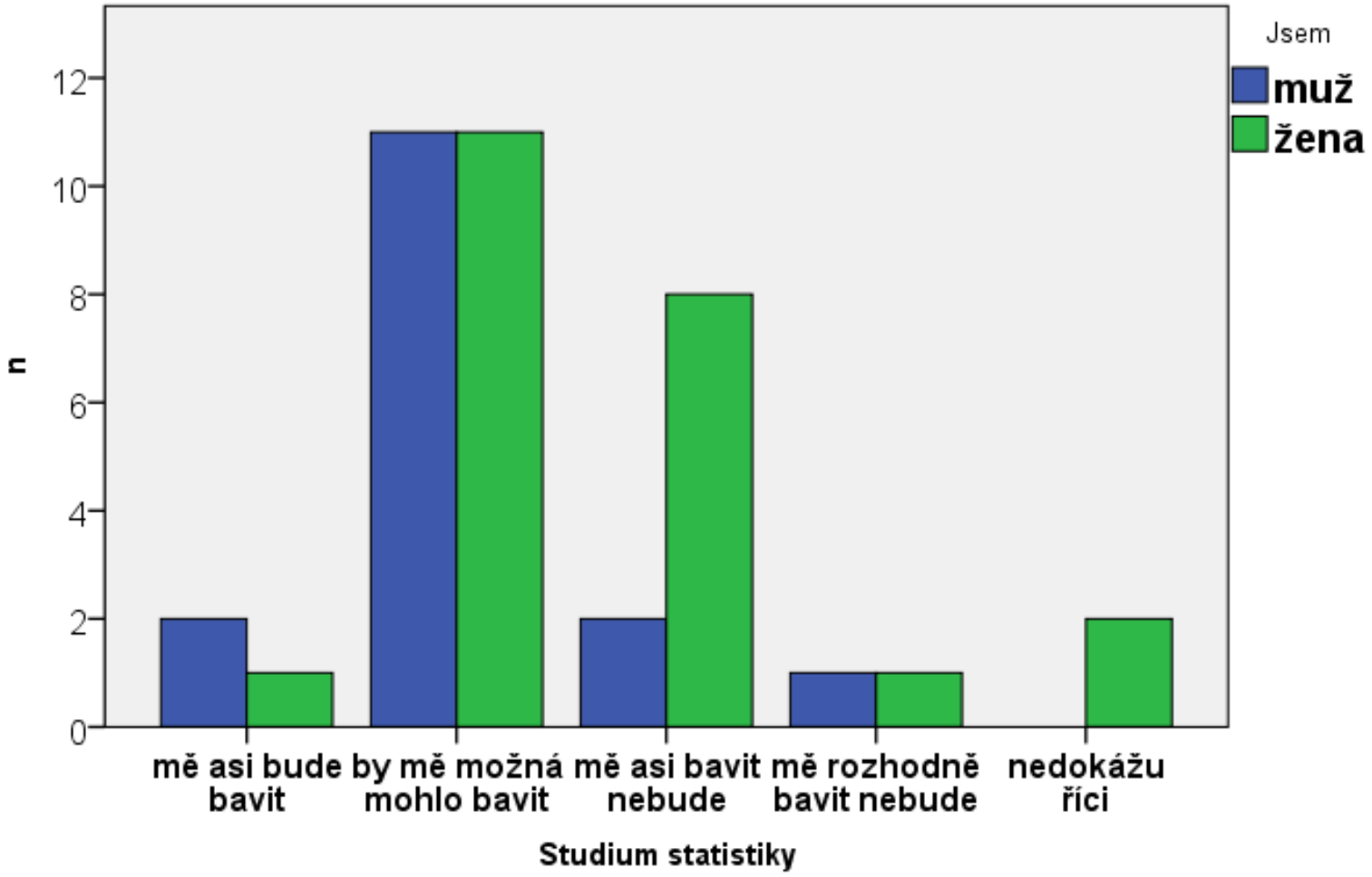
Sloupcový diagram

26. Studium statistiky

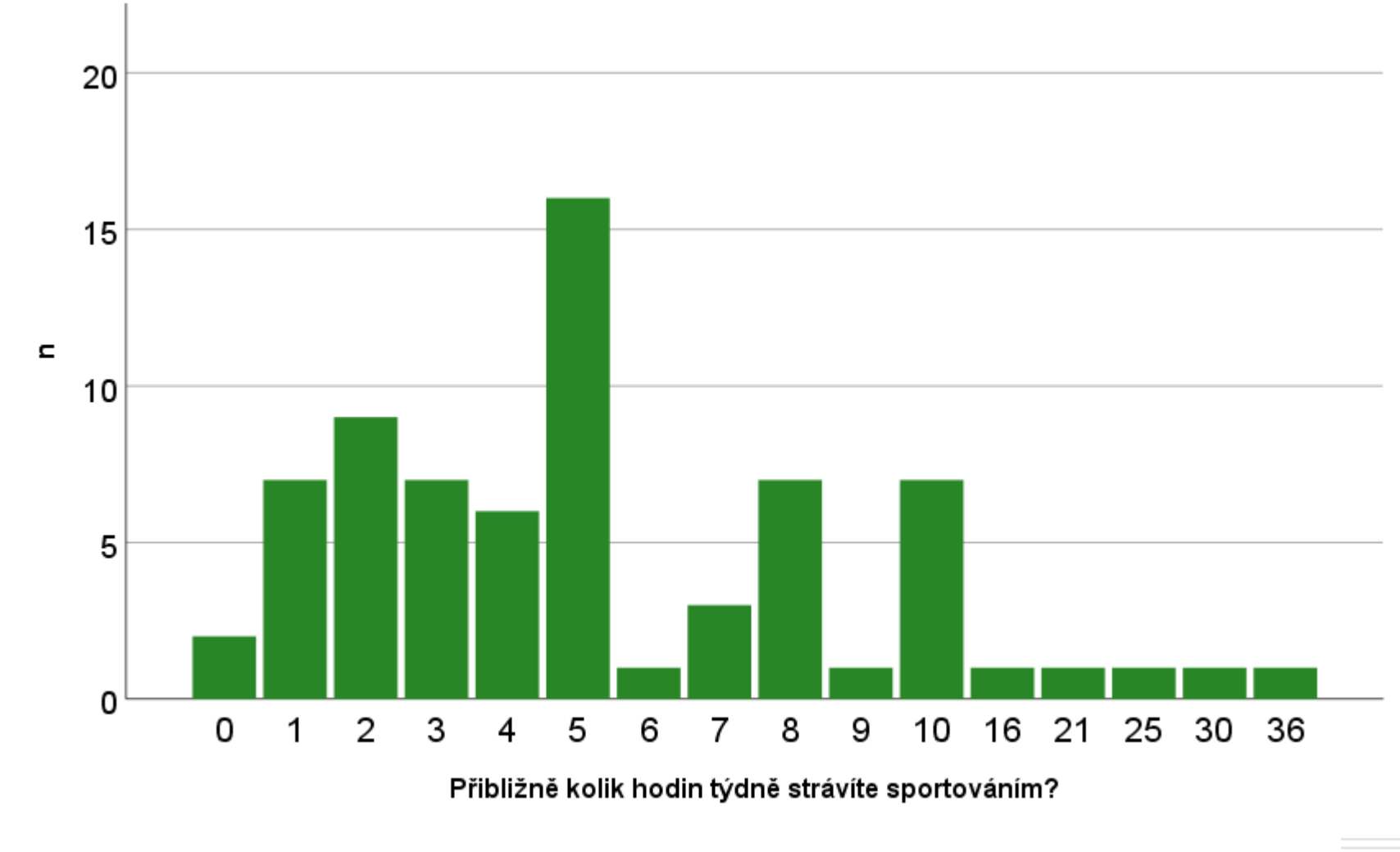
- mě asi bude bavit
- by mě možná mohlo bavit
- mě asi bavit nebude
- mě rozhodně bavit nebude
- nedokážu říci



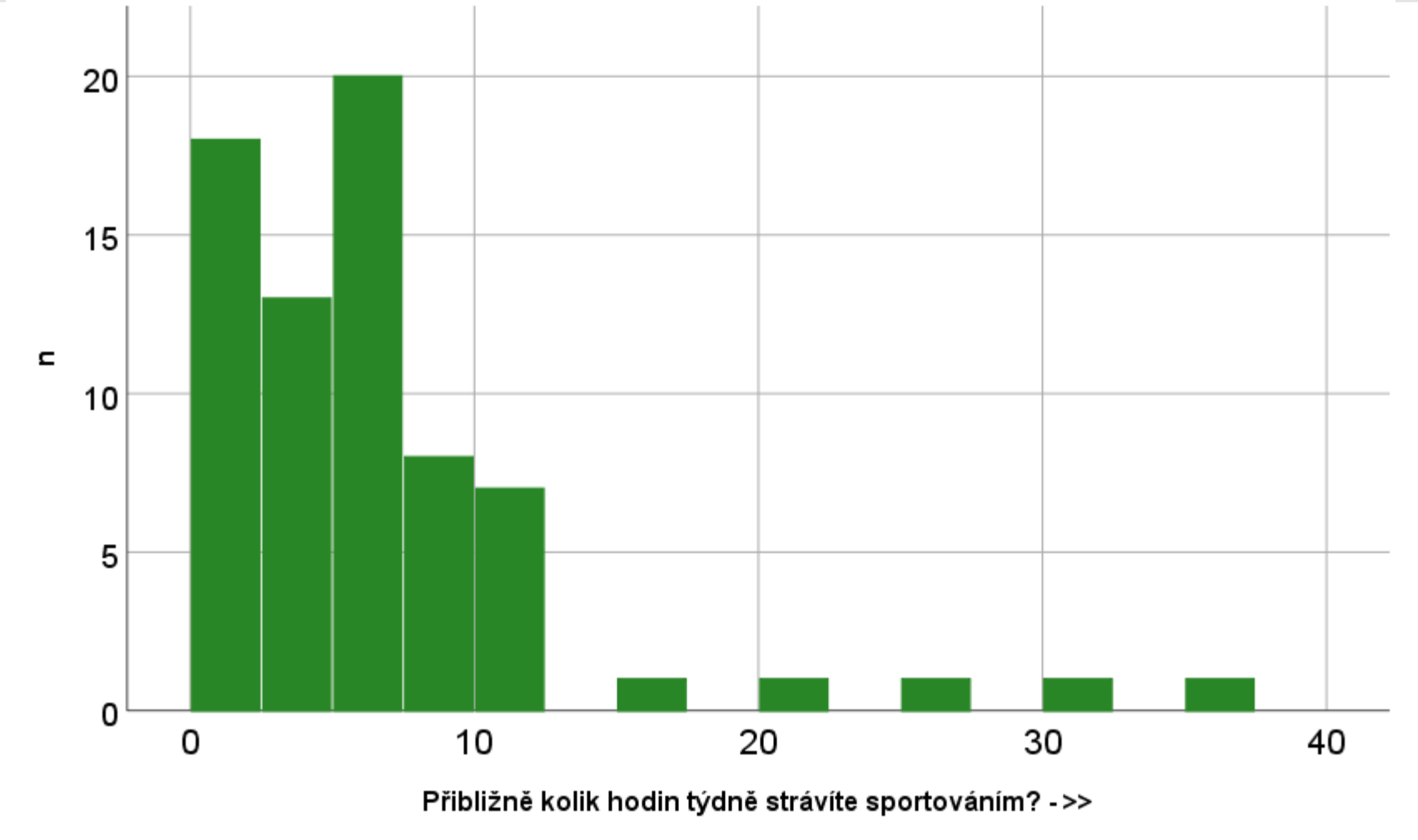
Sloupcový diagram s tříděním



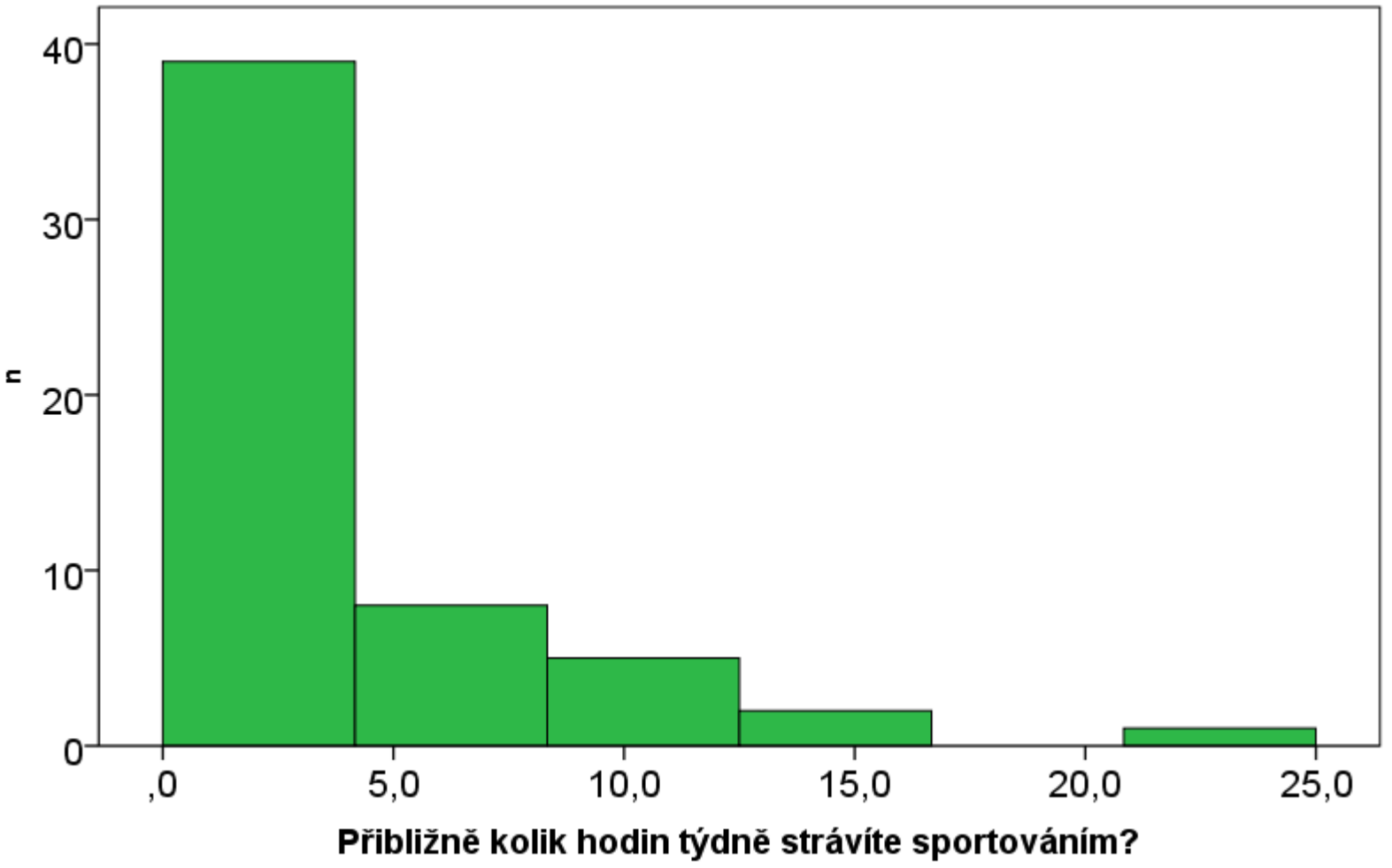
?



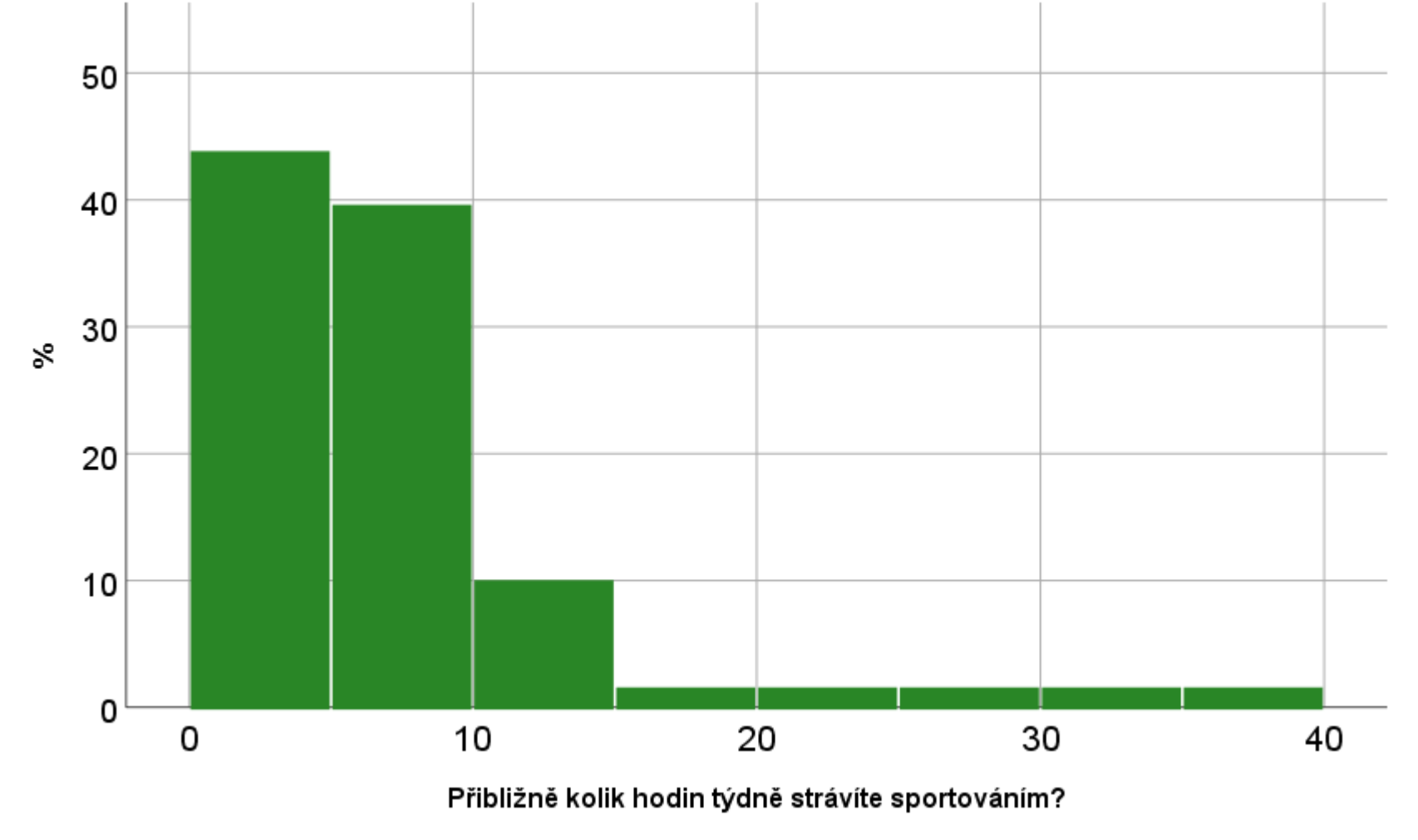
Histogram



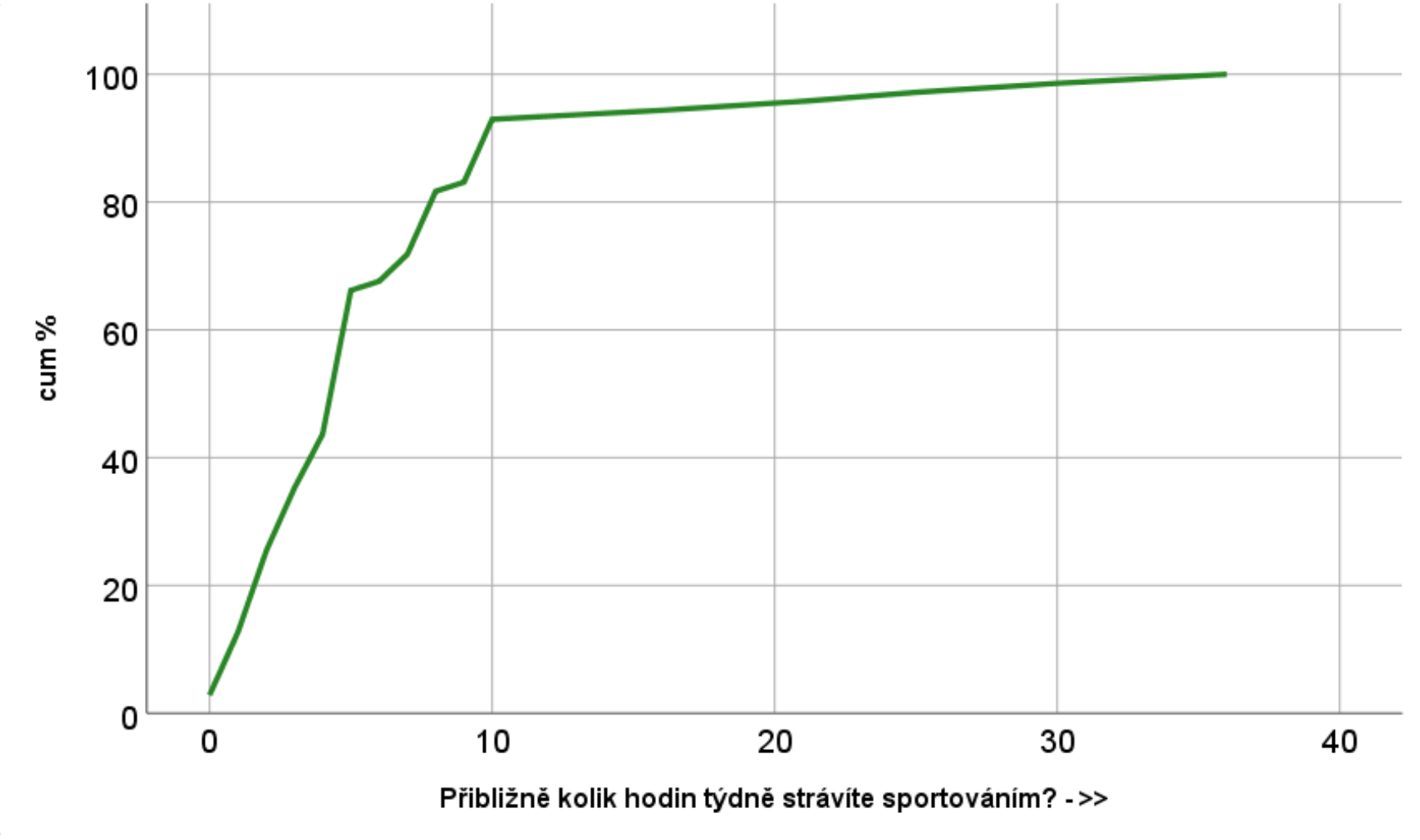
Histogram s širšími intervaly



Histogram s relativními četnostmi (%)



Kumulativní frekvenční polygon
(empirická kumulativní distribuční funkce)

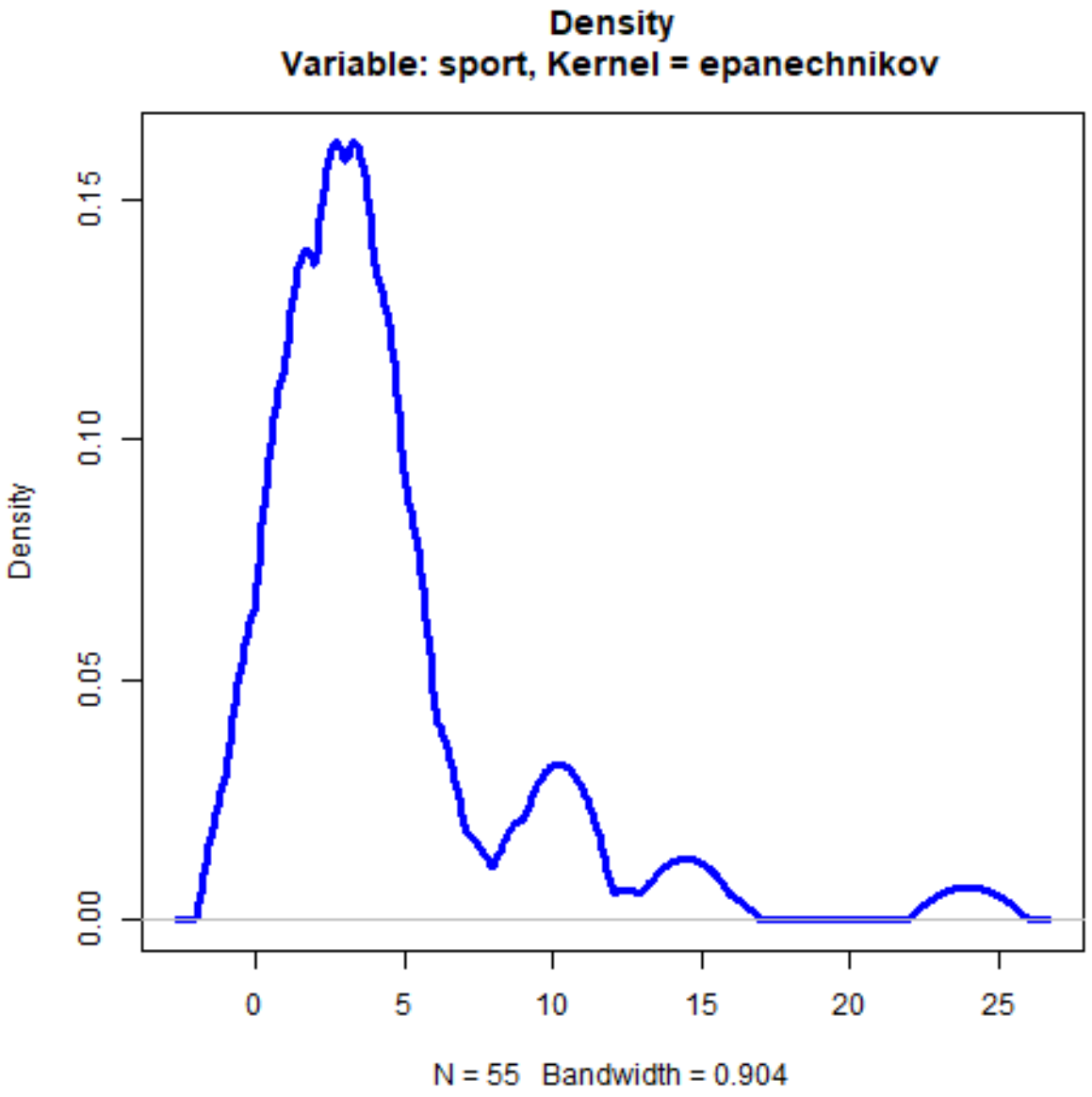


Číslicový histogram „stonek a list“

Přibližně kolik hodin týdně strávíte sportováním? - >> Stem-and-Leaf Plot

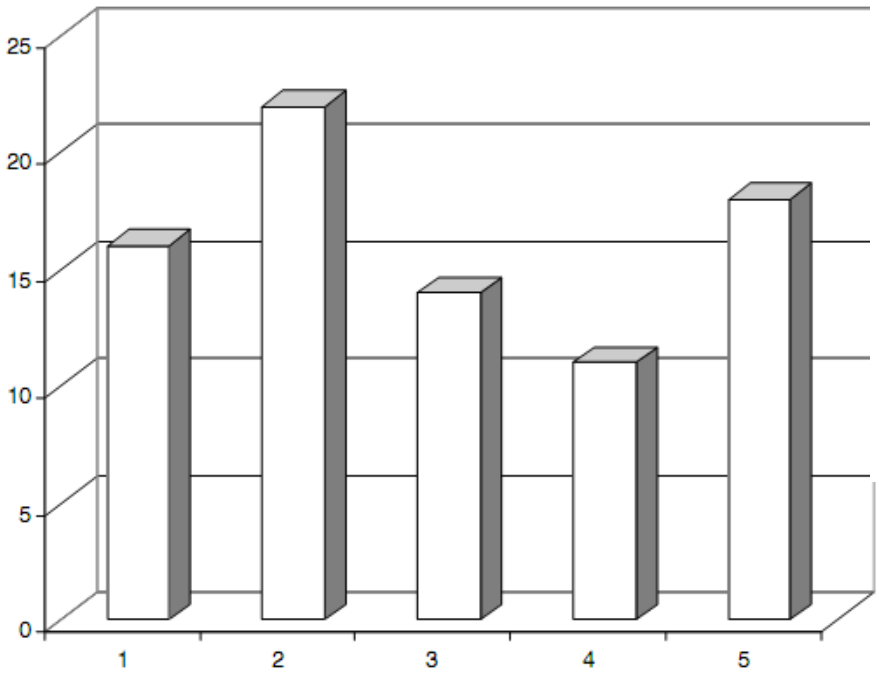
```
 Frequency    Stem &  Leaf

     9,00        0 .  001111111
    16,00        0 .  2222222223333333
    22,00        0 .  44444455555555555555555
     4,00        0 .  6777
     8,00        0 .  88888889
     7,00        1 .  0000000
      ,00        1 .
      ,00        1 .
     1,00        1 .  6
     4,00 Extremes    (>=21)

 Stem width:        10
 Each leaf:       1 case(s)
```

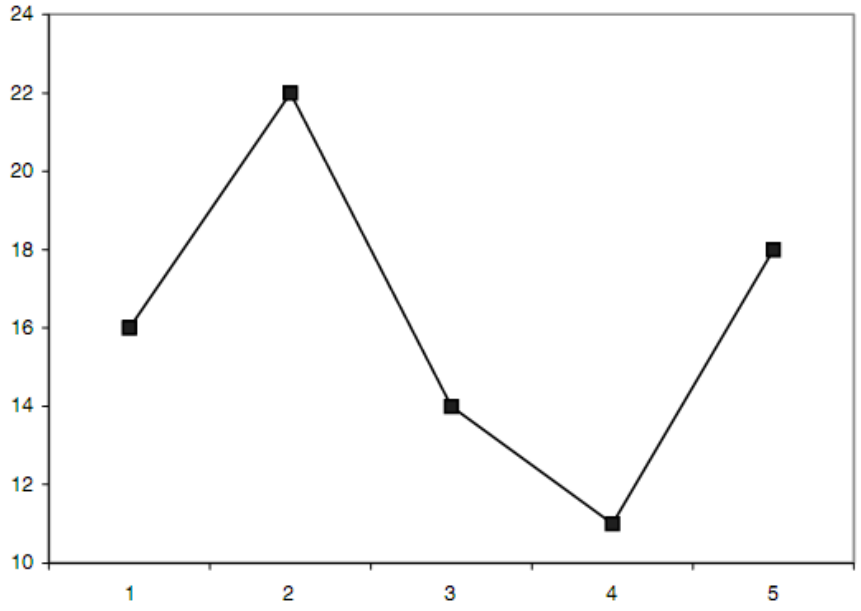
Kernel density plot

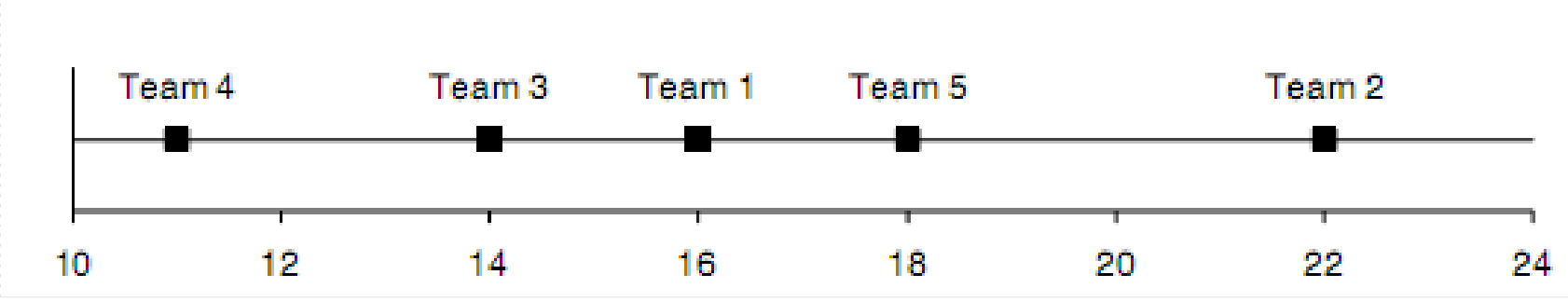


„Férové“ zobrazení dat

- 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





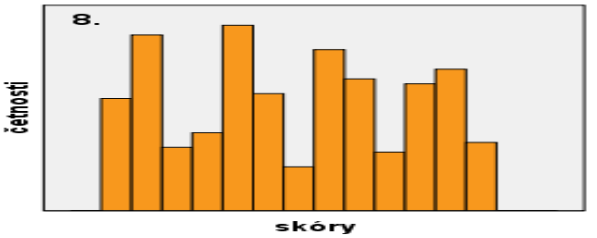


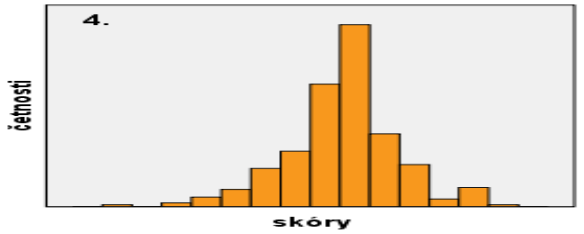
Shrnutí

- **Data** mají typicky podobu **proměnných**, které nesou informaci o různých aspektech jevu, který nás zajímá.
- První informací (*statistikou*), která nás zajímá, je **četnost** výskytu jednotlivých hodnot (resp. hodnot uvnitř jednotlivých intervalů)
- Četnosti popisujeme (=komunikujeme je)
  - tabulkou četností
  - graficky – histogram, sloupcový diagram
  - (pomocí percentilů)

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.

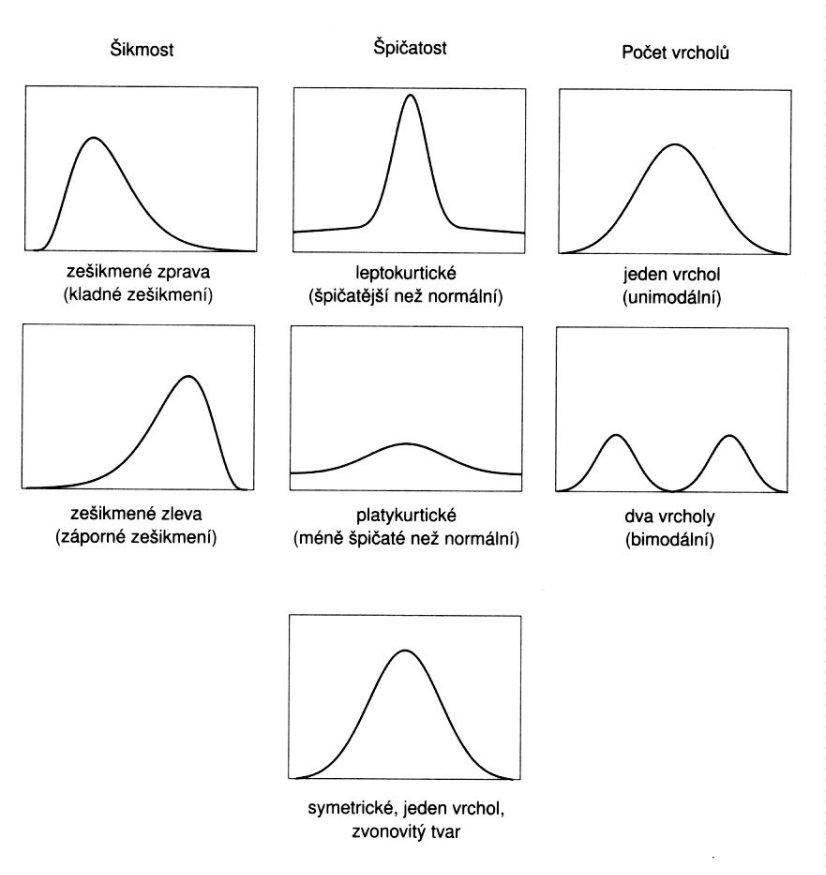




- **Rozdělení** četností je tedy to, kolik relativně (či absolutně) máme kterých hodnot měřené proměnné.
  - Typicky lze přibližně popsat slovy, např.: vyskytlo se hodně středních hodnot a relativně málo extrémních hodnot.
  - Toto **rozložení** jevů na měřené škále je nejlépe vidět na grafech.
  - Obvykle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

Tvar rozložení četností

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



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

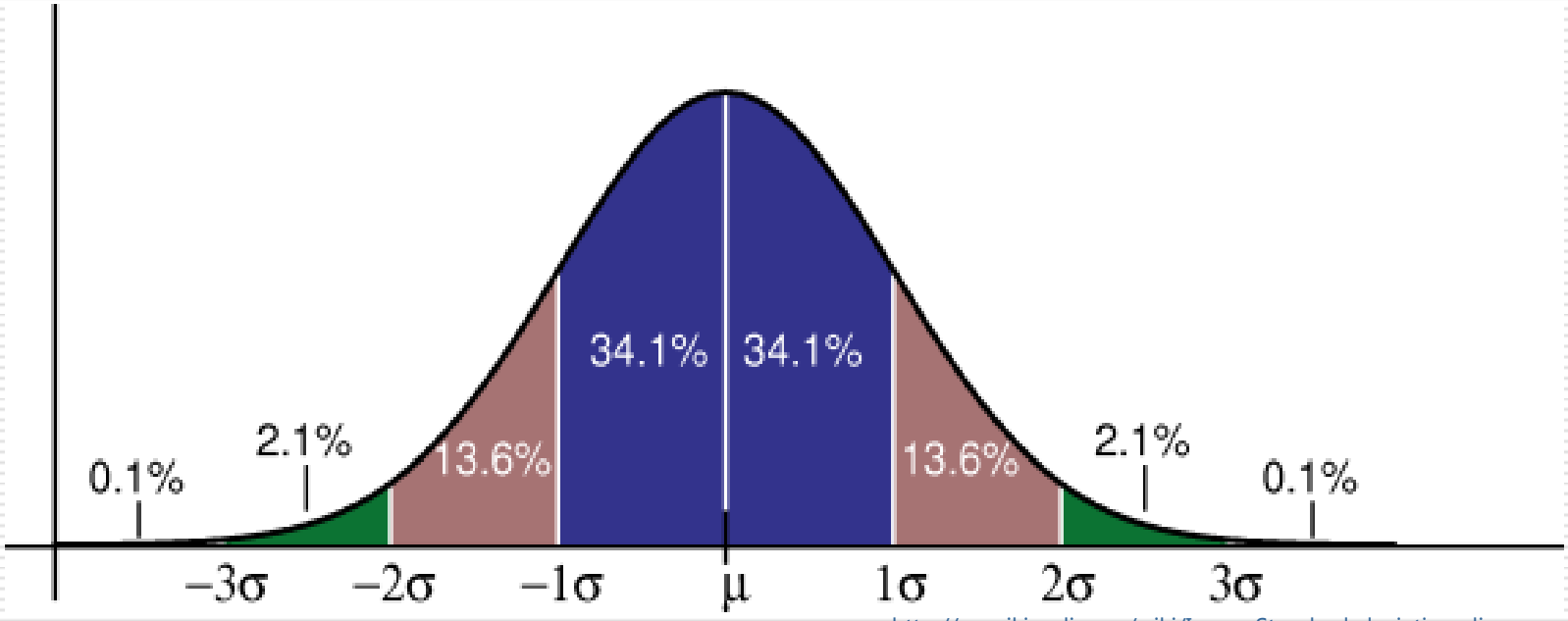
Parametrický popis rozložení

- **Rozložení** je úplně popsáno (určeno) četnostmi jednotlivých hodnot, popř. intervalů.
- Je tedy popsáno množstvím statistik (četností), přesněji k-1 četnostmi, pokud proměnná nabývá k hodnot (či k intervalů).

Lze rozložení popsat efektivněji, méně statistikami (**parametry**)?

- Všechny hodnoty jsou stejně četné (1 číslo)
  - $f_k = k/N$ kde k je konstanta ..... **UNIFORMNÍ** rozložení
- Četnosti jsou výsledkem procesu, který se dá připodobnit k opakovanému házení korunou, kdy nás zajímá počet „hlav“
  - $p_k = p^k(1-p)^{n-k}(n!/(n!-k!))$ kde n = počet hodů, k= počet hlav p=pravděpodobnost „hlavy“
  - **BINOMICKÉ** rozložení pro diskrétní proměnné
- Normální rozložení

Normální (Gaussovo) rozložení

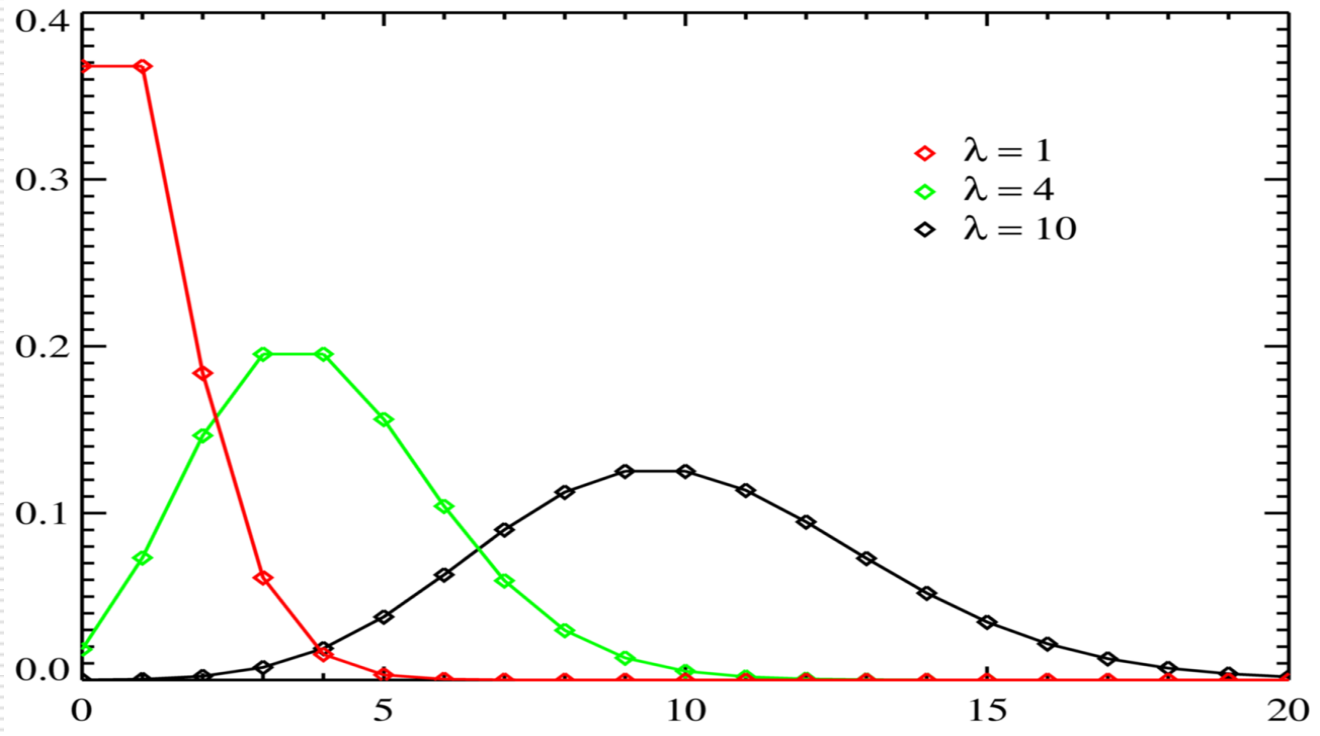


http://en.wikipedia.org/wiki/Image:Standard_deviation_diagram.png

- „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**.