

PSY117 2018

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



is it normal

is it normal to talk to yourself

is it normal for your period to be brown

is it normal to miss a period

is it normal to be sexually attracted to numbers

is it normal to bleed during intercourse

is it normal to get your period late

is it normal to poop green

is it normal to have headaches everyday

is it normal to have hair on your bum

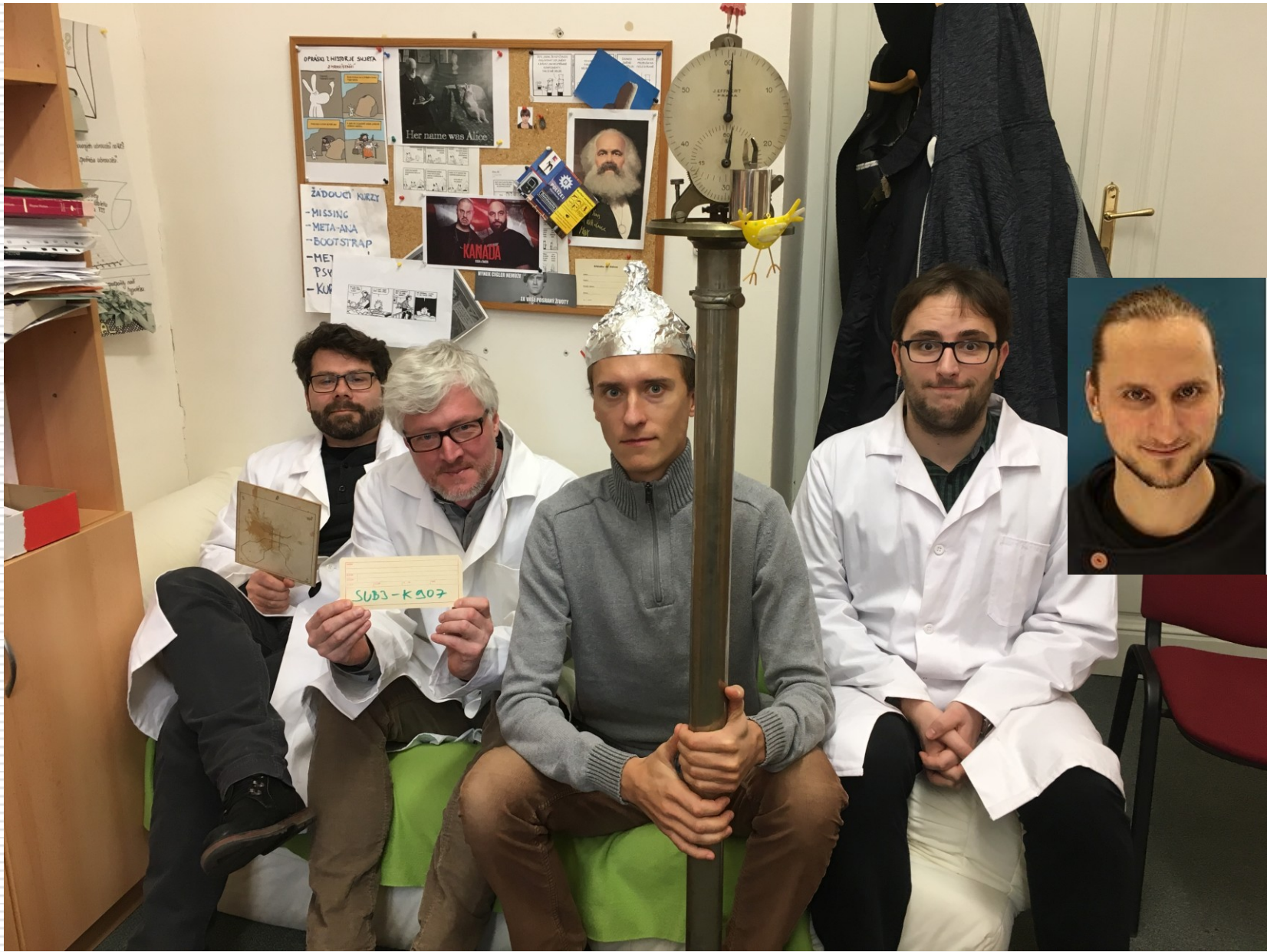
is it normal to spot during pregnancy

FAIL

Google Search

I'm Feeling Lucky

Vyučující



Kostrá PSY117 – Statistická analýza dat

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

2 seminární práce (20b)

3 průběžné písemky (3x10b)

Závěrečný test (50b)

Obtížnost statistiky

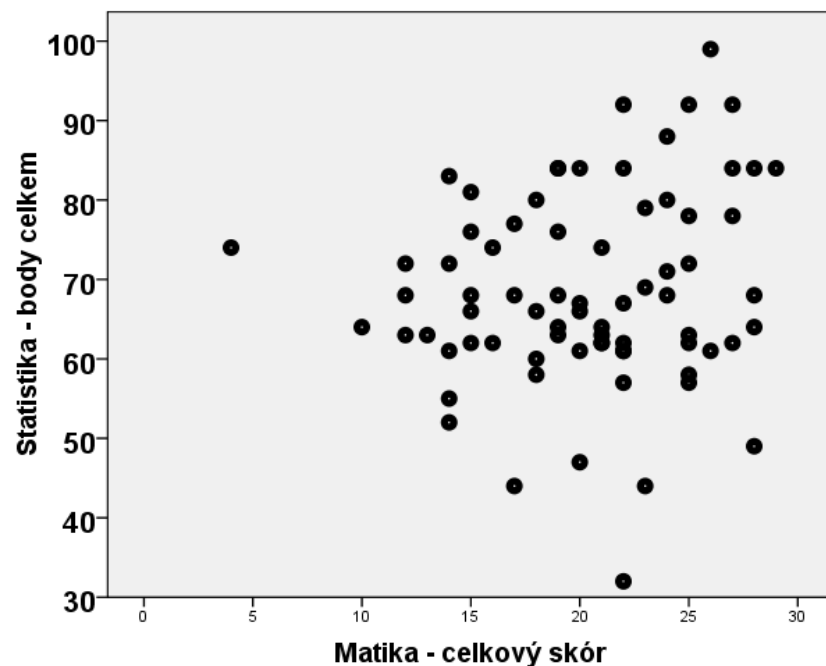
Kód	Zapsáno	A	B	C	D	E	F	-
2017	99	6	18	24	17	11	13	7
2016	86	2	15	17	15	12	13	7
2015	78	7	10	19	18	7	6	9
2014	73	4	6	20	13	11	10	8
2013	98	6	18	16	15	9	16	14
2012	84	8	25	8	12	4	16	9
2011	76	9	11	12	11	4	12	15
2010	81	8	17	12	13	8	11	9

Obtížnost statistiky

Kód	Zapsá no	A	B	C	D	E	F	-
2017	99	6%	18%	24%	17%	11%	13%	7%
2016	86	2%	17%	20%	17%	14%	15%	8%
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, jamovi, R...)
- Nemáme dobrou učebnici 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, 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 **vztahů** mezi proměnnými/jevy – *závislé a nezávislé* proměnné

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

Data? Jaká data?

O charakteristikách účastníků výzkumu.

O výkonech žáků ve škole

O vybraném vzorku lidí, s nimiž budeme srovnávat své klienty

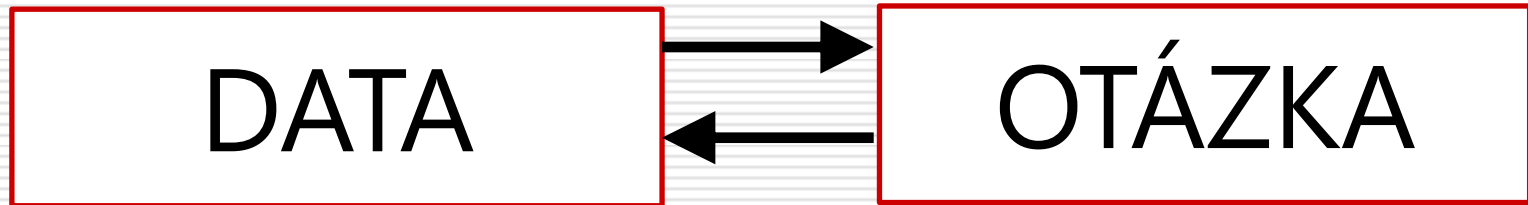
O jednom klientovi a jeho změnách v čase

O sobě, svém životě, nebo praxi

Záznamy, pozorování, testování, vzpomínky...

Záměrně vytvářená data pro výzkum

K čemu data?



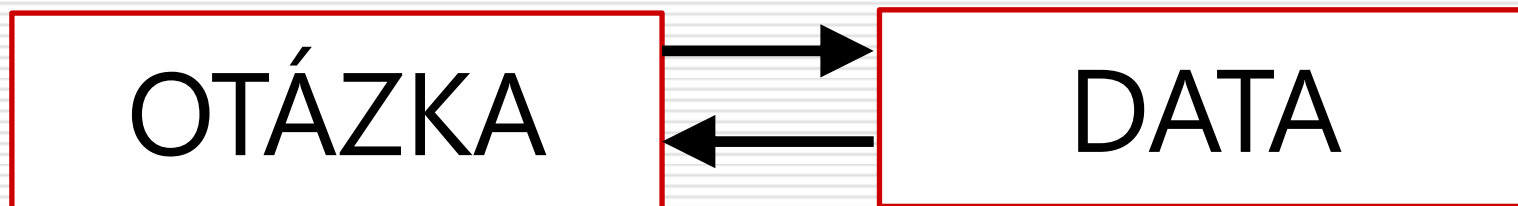
Známky.
Počet spotřebovaných termínů.
Čas strávený učením.
Hrdost při zvládnutí...

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 (\approx 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 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**

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**

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;

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

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

Více viz extrakt z Urbánek, Denglerová, Širůček v ISu

Typy proměnných podle počtu možných hodnot

□ **Spojité** proměnné

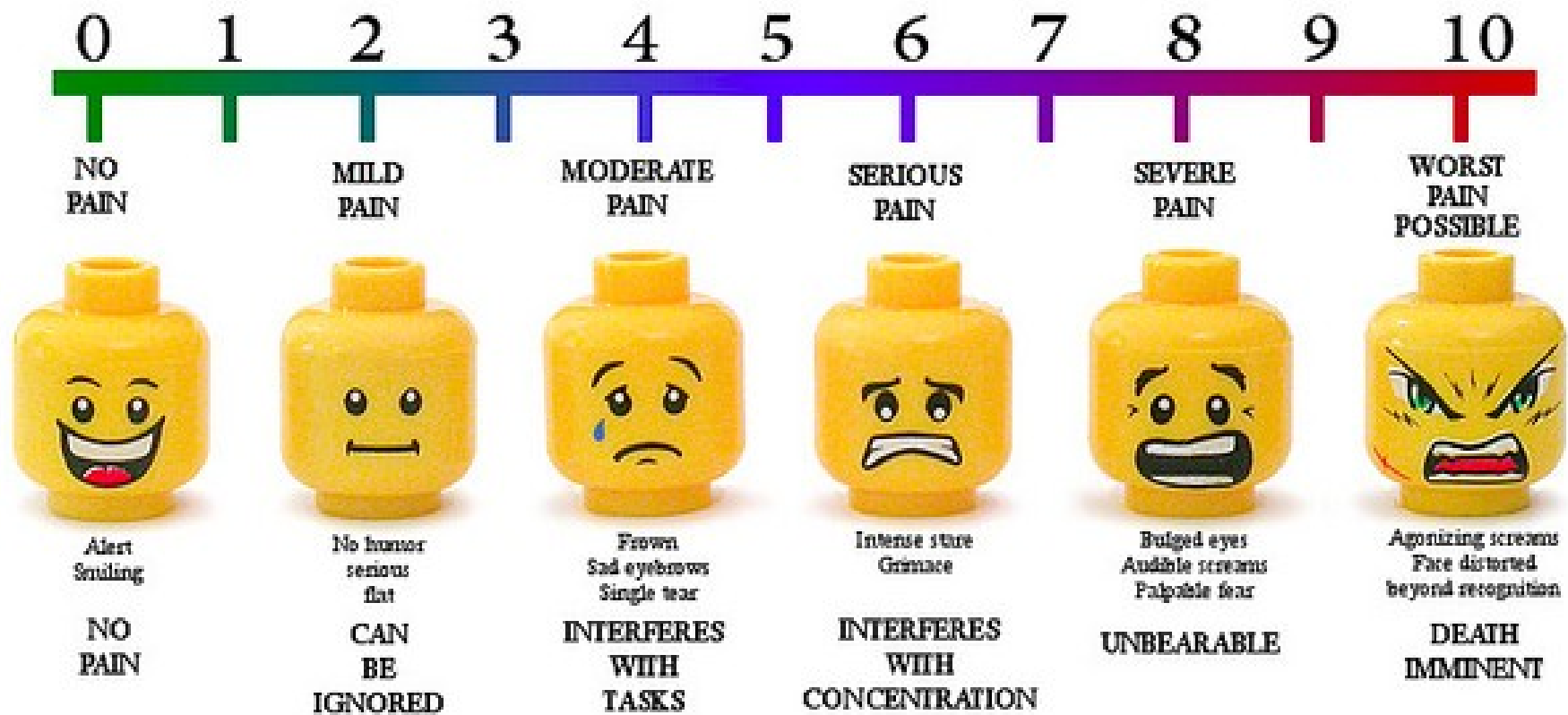
- Nekonečně mnoho hodnot – reálná čísla

□ **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**

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



Created by Brendan Powell Smith www.TheBrickTestament.com This chart is not sponsored, authorized, or endorsed by the LEGO Group.

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 méně?
 - 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	<i>N</i>		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.
- Tabulka četností zobrazuje téměř všechna data
 - Použitím intervalů již data mírně redukuje
- Minimální podoba tabulky četností: **abs. a rel. četnosti, součtový poslední řádek**

„Ruční“ tvorba tabulky četností

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

Hodin					
0		Popisky řádků	f	%	cum %
10		0	6	10,91%	10,91%
4		1	5	9,09%	20,00%
5		2	10	18,18%	38,18%
5		3	8	14,55%	52,73%
2		4	10	18,18%	70,91%
5		5	5	9,09%	80,00%
2		6	1	1,82%	81,82%
5		7	2	3,64%	85,45%
3		10	4	7,27%	92,73%
24		11	1	1,82%	94,55%
5		14	1	1,82%	96,36%
2		15	1	1,82%	98,18%
14		24	1	1,82%	100,00%
2		Celkový součet	55	100,00%	
3					
0					
3					

Hodin	Interval		Četnost <i>f</i>	kumulativní		relativní četnost %	kumulativní	
	Dolní mez	Horní mez		četnost <i>cum f</i>	relativní četnost <i>cum %</i>			
0								
10								
4	0	5	44	44	80,0	80,0		
5	5	10	7	51	12,7	92,7		
5	10	15	3	54	5,5	98,2		
2	15	20	0	54	0,0	98,2		
5	20	25	1	55	1,8	100,0		
2	Celkem		55		100			
5								
3								
24								
5								
2								
14								
2								
3								
0								
3								
1								
3								
0								
2								
1								
1								

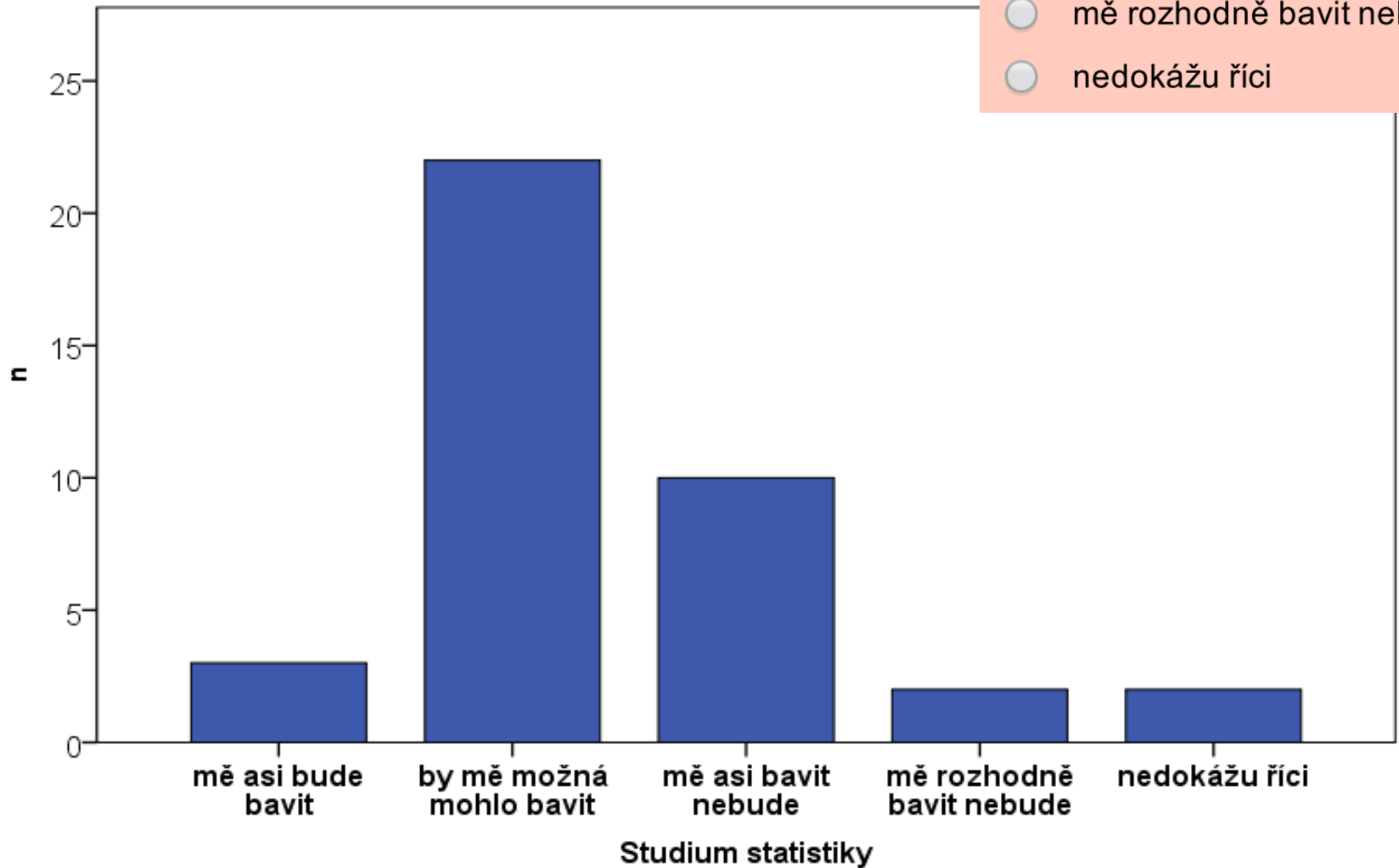
Poklikáním se tabulka otevře jako editovatelný objekt v Excelu.

Těž Datíčka.xls, list „četnosti“.

Grafické podoby tabulky četností

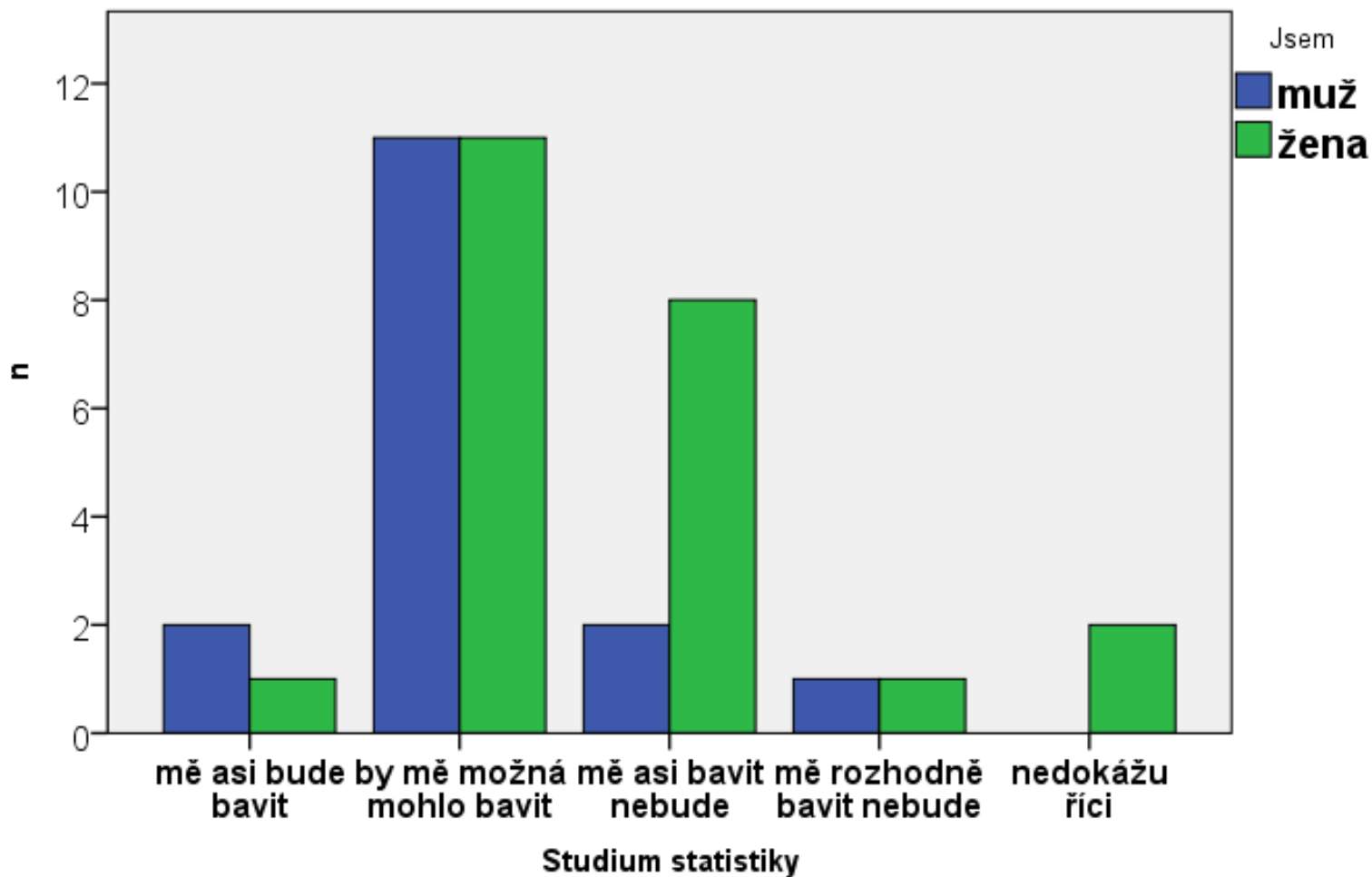
- Kategorické proměnné
 - sloupcový graf (diagram)
 - koláčový diagram – zřídka, neukazuje rozložení
- Metrické proměnné
 - Histogram – jako sloupcový, ale šíře sloupců reprezentuje šíři intervalů
 - stem-and-leaf – rozdělení hodnot do intervalů

Sloupcový diagram

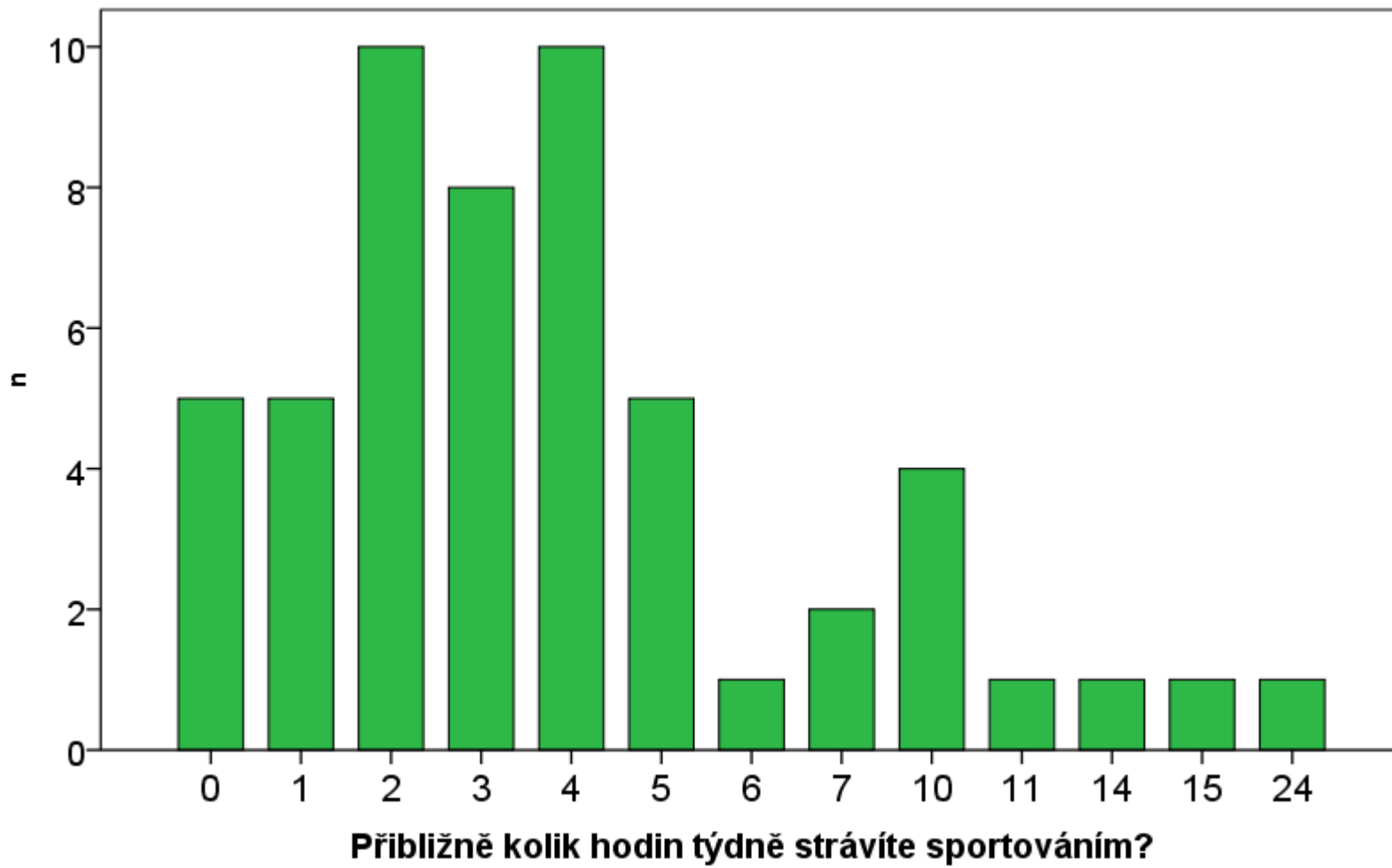


- mě asi bude bavit
- by mě možná mohlo bavit
- mě asi bavit nebude
- mě rozhodně bavit nebude
- nedokážu říci

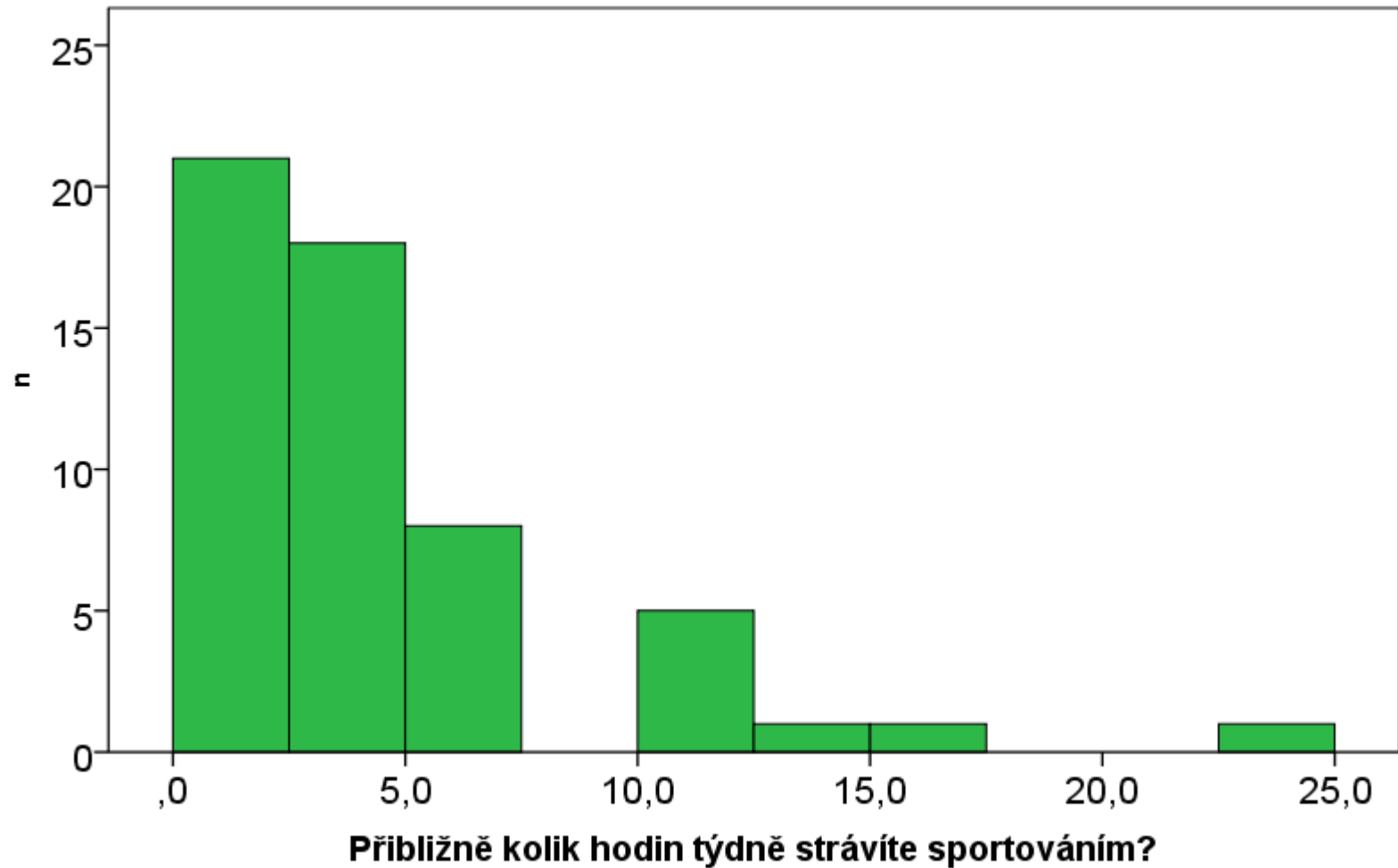
Sloupcový diagram s tříděním



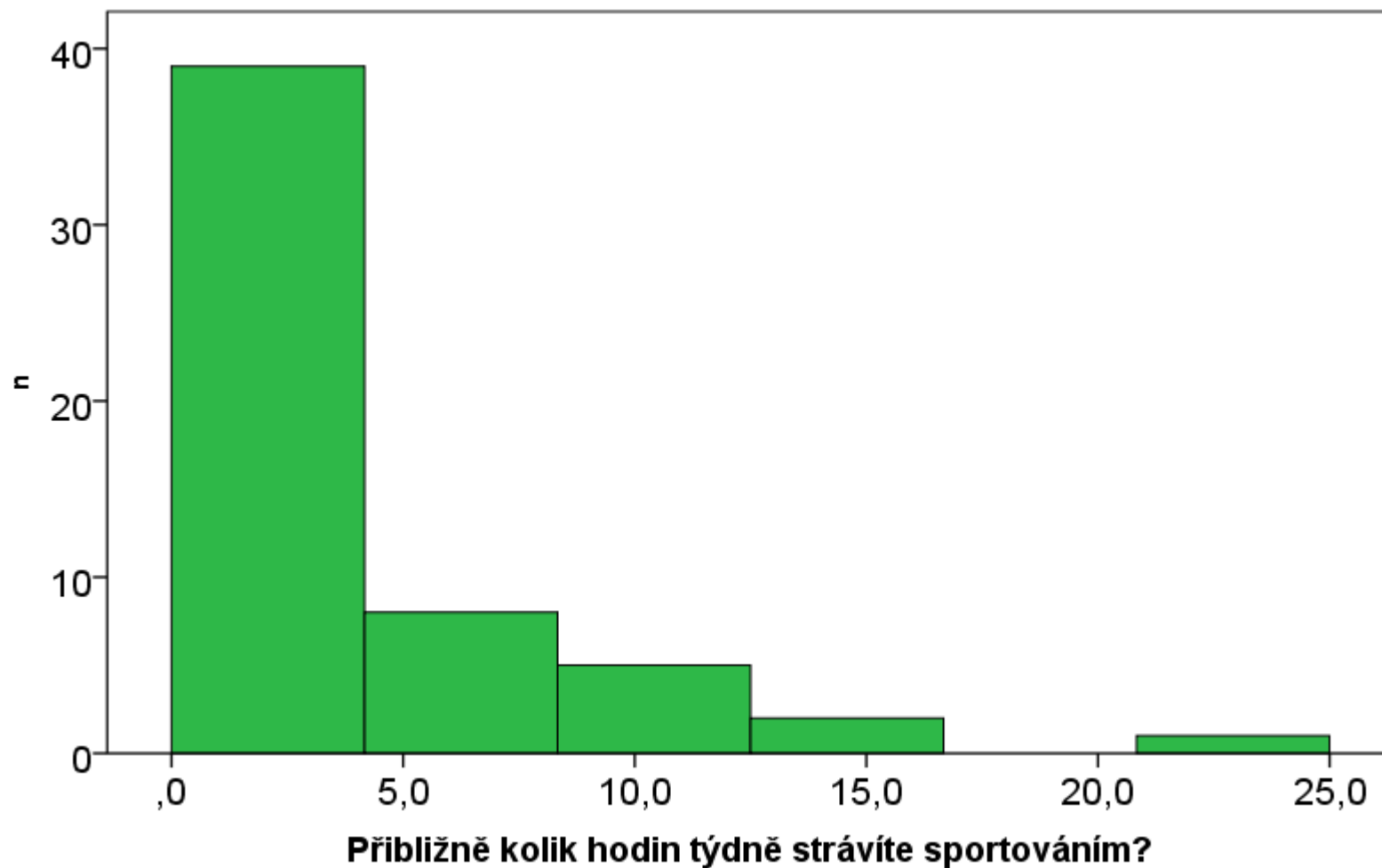
?



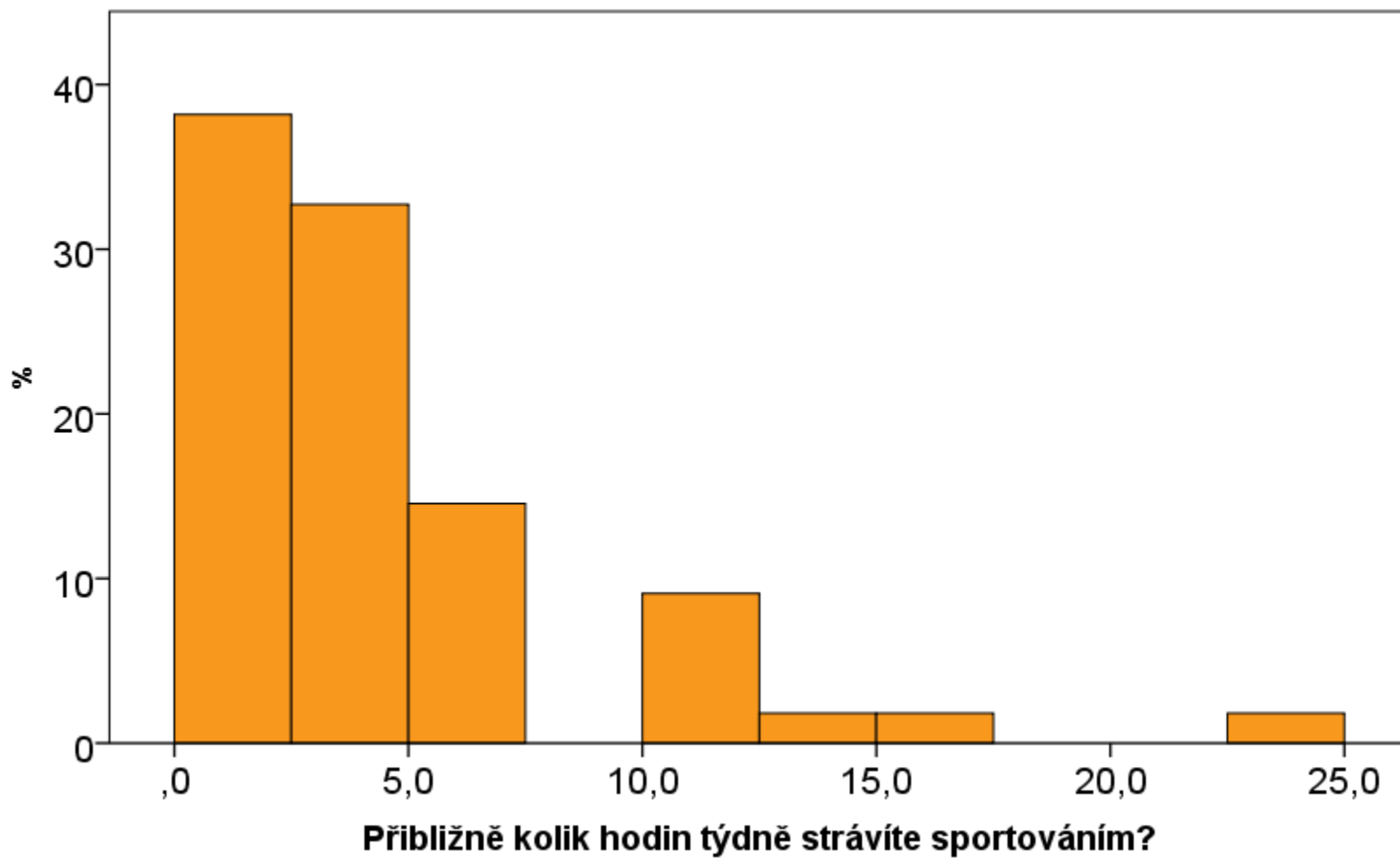
Hostogram



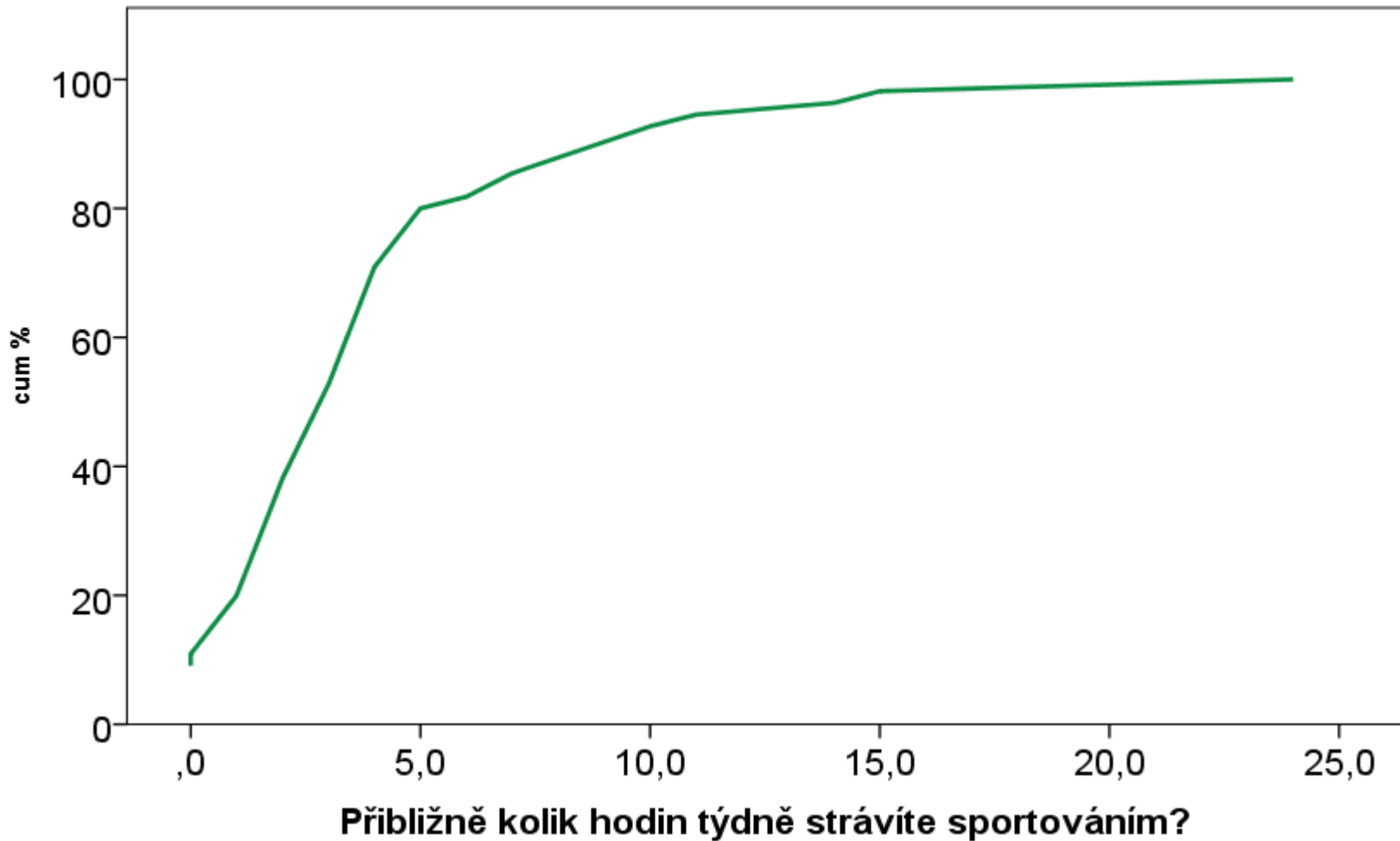
Histogram s širšími intervaly



Histogram s relativními četnostmi (%)



Kumulativní frekvenční polygon (empirická kumulativní distribuční funkce)



Číslicový histogram „stonek a list“

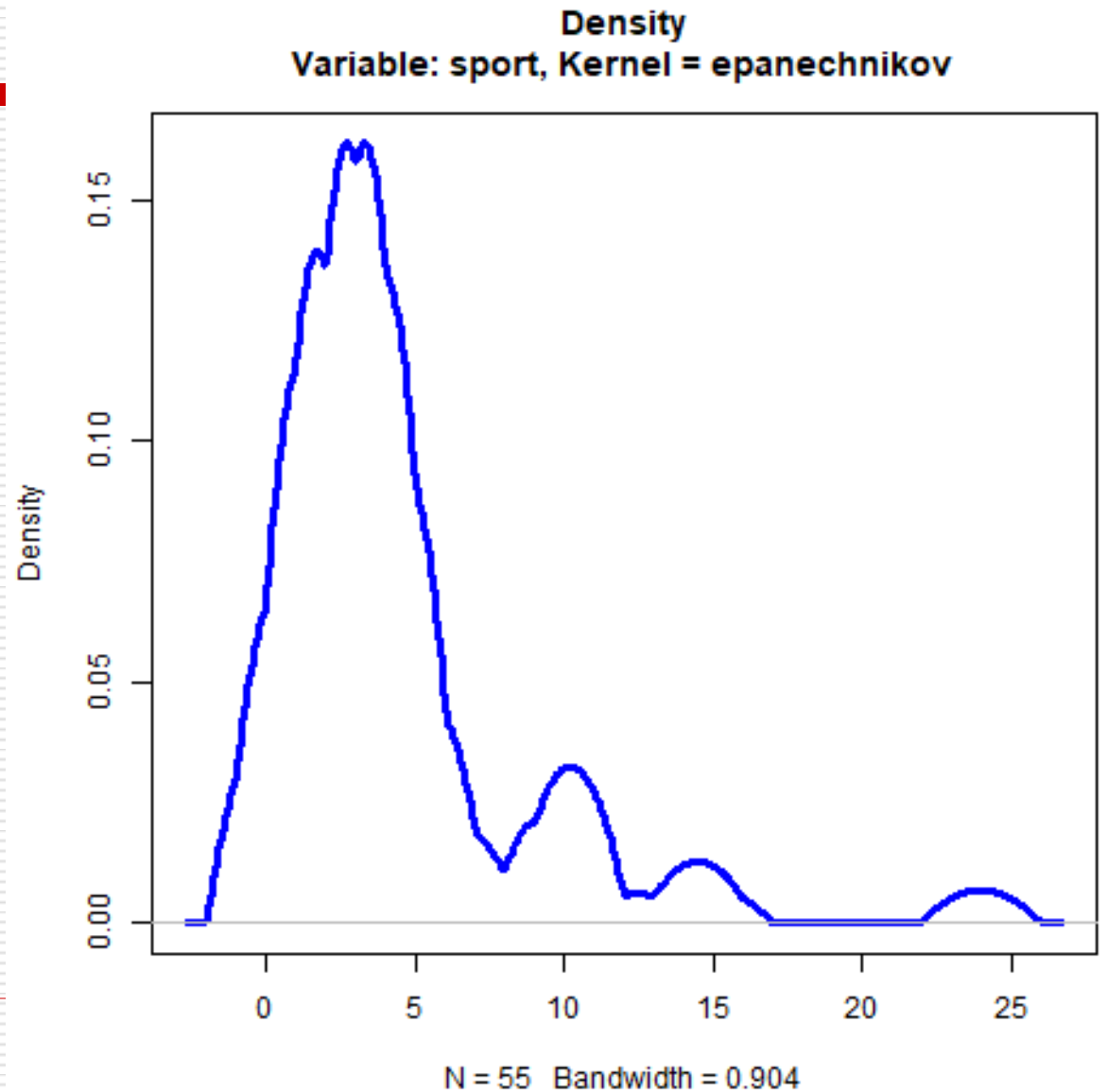
Přibližně kolik hodin týdně strávíte sportováním? Stem-and-Leaf Plot

Frequency	Stem &	Leaf
6,00	0 .	000000
5,00	1 .	00000
10,00	2 .	0000000000
8,00	3 .	00000000
10,00	4 .	0000000000
5,00	5 .	00000
1,00	6 .	0
2,00	7 .	00
8,00	Extremes	(>=10)

Stem width: 1

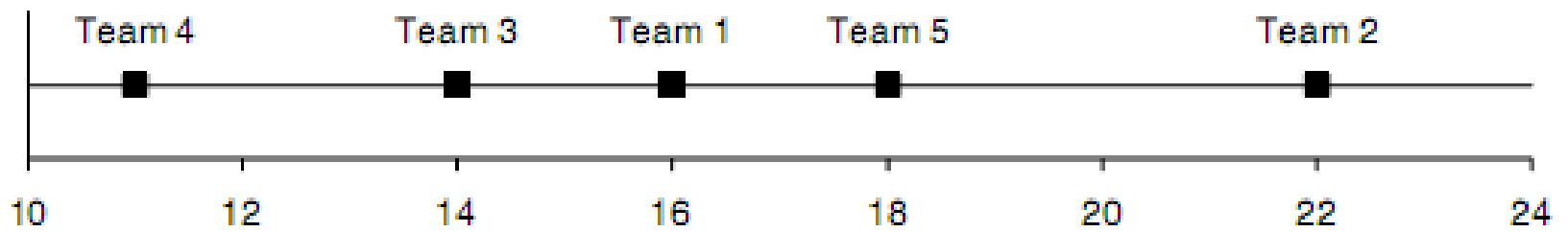
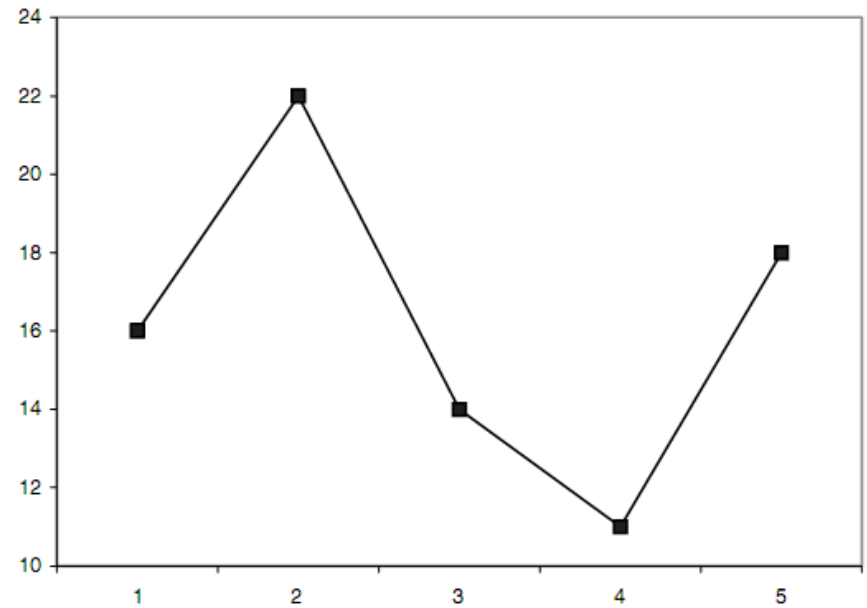
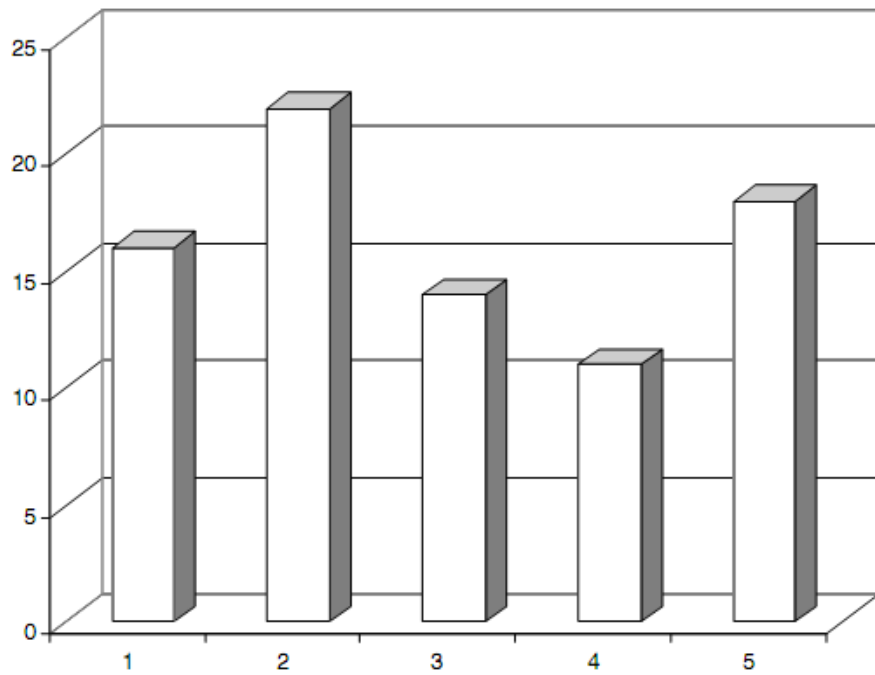
Each leaf: 1 case(s)

Kernel density plot



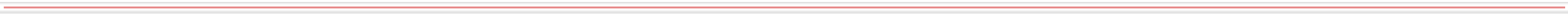
„Férové“ zobrazení dat

- 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.
 - Nepropojujeme 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

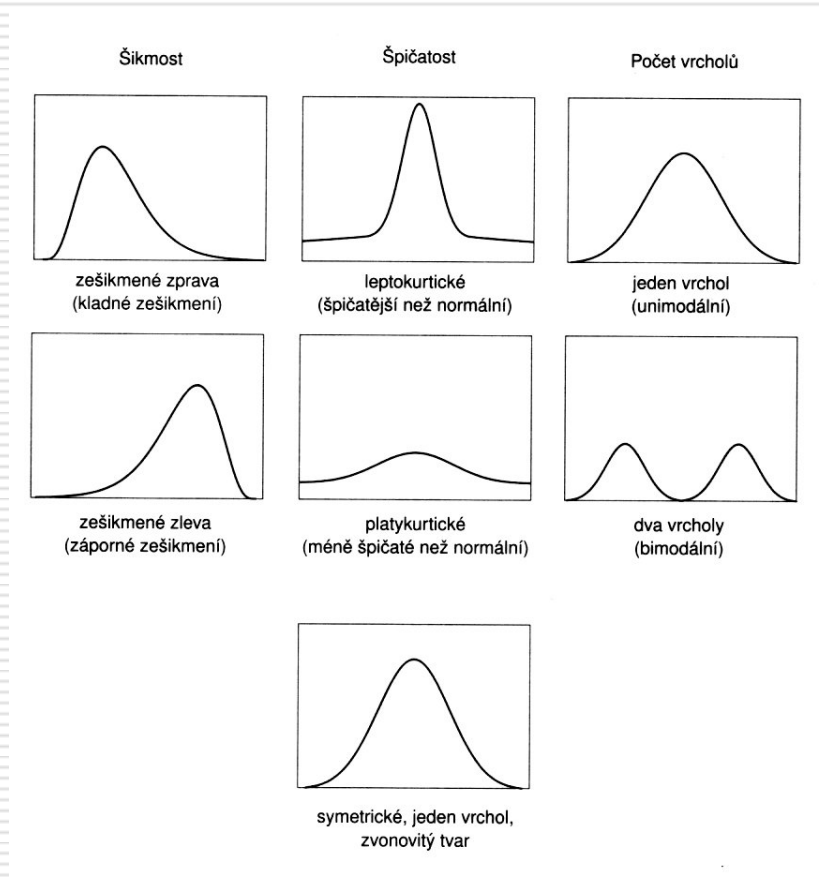
- ❑ Měřené jevy jsou nějak rozděleny do kategorií (intervalů) a tyto kategorie jsou různě „populární“ – četné.
- ❑ Četnosti u reálných ordinálních a vyšších proměnných obvykle nebývají **distribučovány** nahodile – jejich rozdělení zobrazené histogramem má popsateľný tvar.



- ❑ **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



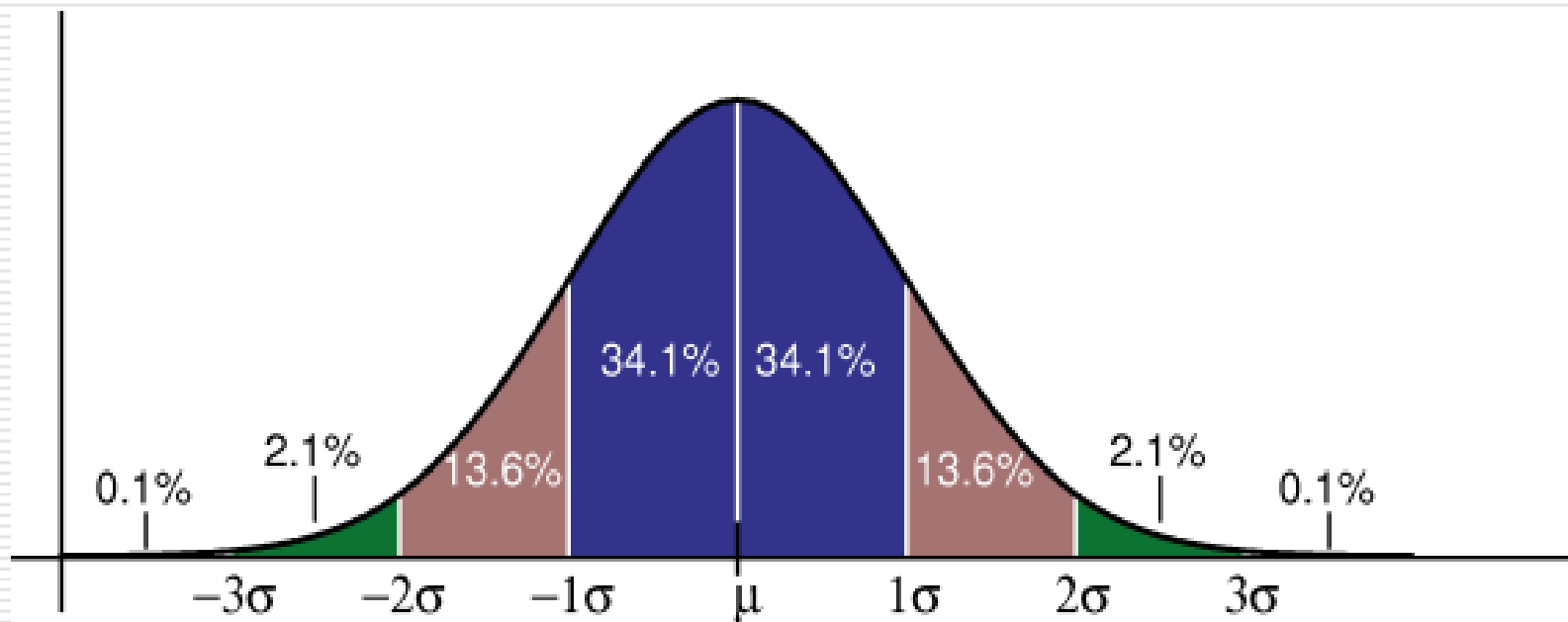
Parametrický popis rozložení

- **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“
 - $p_k = p^k(1-p)^{n-k} \binom{n}{k}$ kde n = počet hodů, k = počet hlav
 p = pravděpodobnost „hlavy“
 - **BINOMICKÉ** rozložení pro diskrétní proměnné
- Normální rozložení

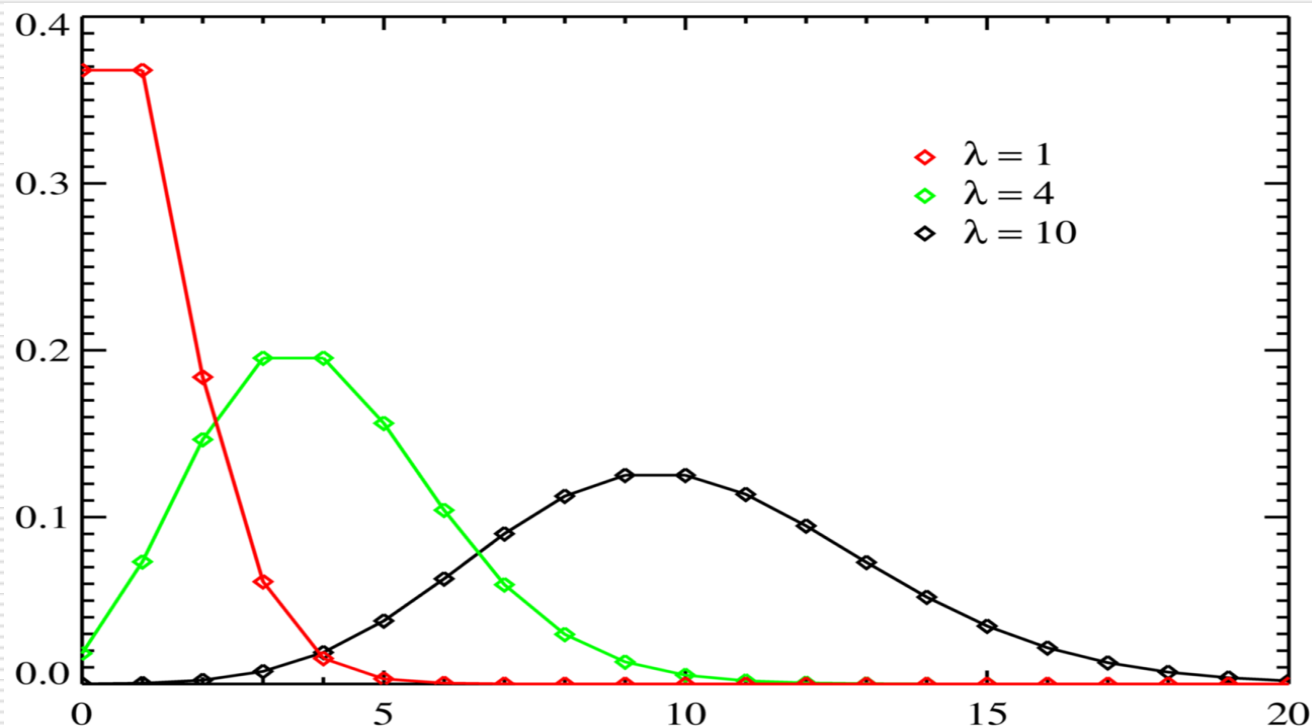
Normální (Gaussovo) rozložení



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.

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.

Rozložení

- Znamky 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**.