Biostatistika

jarkovsky@iba.muni.cz

Přednáška 1

Organizační informace – kódy předmětů

- Bi5040 Biostatistika - základní kurz (tato přednáška)
  - Bi5040c Biostatistika – cvičení (nepovinný – samostatné cvičení na PC)

- ASTAp Biostatistika – přednáška (tato přednáška)
  - ASTAc Biostatistika – cvičení (povinný – samostatné cvičení na PC)

- BMBS051 Biostatistika-základní kurz (tato přednáška)

- BLBS051p + BLBS051c – Biostatistika (sloučené, tato přednáška)

Organizační informace – poznámka k cvičení Bi5040c a ASTAc

- Cvičení biostatistiky probíhá pro každou seminární skupinu jednou za dva týdny v délce dvou hodin
- Každá seminární skupina absolvuje během semestru 6 cvičení – přesné termíny zašlou vyučující
- Materiály ke kurzu budou s předstihem k dispozici v IS.MUNI, jejich prostudování se před cvičením vřele doporučuje
- Pro získání zápočtu je třeba:
  - Účast na alespoň 5 z 6 cvičení (větší počet oprávněných absencí bude řešen individuálně)
  - Splnění zápočtového testu na konci semestru (teoretická část + řešení příkladů na počítači)

- Cvičení není nutné pro získání zkoušky z předmětu Bi5040/ASTA, jde o rozšiřující prakticky orientovaný předmět

Organizační informace – výukové materiály

- Tato prezentace v IS.MUNI (tento semestr bude vkládána po částech, snažím se ji letos upgradovat) + prezentace a příklady ovládání SW Statistica + další souhrnné podklady
- www.matematickabiologie.cz/res/file/ucebnice/pavlik-biostatistika.pdf
- portal.matematickabiologie.cz/index.php?pg=aplikovana-analyza-klinickych-a-biologickych-dat--biostatistika-pro-matematickou-biologii
- Tabulky statistických rozdělení www.statsoft.com/Textbook/Distribution-Tables
- Libovolná základní učebnice statistiky – např.
  - https://www.amazon.com/Biostatistical-Analysis-5th-Jerrold-Zar/dp/0131008463/ref=sr_1_1?ie=UTF8&qid=1505890489&sr=8-1&keywords=zar+biostatistical+analysis
  - https://www.amazon.com/Medical-Statistics-Glance-Aviva-Petrie/dp/140518051X/ref=sr_1_sc_1?s=books&ie=UTF8&qid=1505890508&sr=1-1-spell&keywords=avive+petria
  - https://www.amazon.com/Statistics-Veterinary-Animal-Science-Petrie/dp/0470670754/ref=sr_1_sc_3?s=books&ie=UTF8&qid=1505890522&sr=1-3-spell&keywords=avive+petria

Organizační informace – software

- Software
  - Univerzitní licence na inet.muni.cz (stejný login a passwd jako do is.muni.cz)
  - Statistica – www.statsoft.com, www.statsoft.cz
  - SPSS - www.ibm.com/analytics/us/en/technology/spss/
  - R – www.r-project.org, www.rstudio.com
  - Stata - www.stata.com

Organizační informace – uzavření předmětu

- Bi5040 Biostatistika - základní kurz
- ASTAp Biostatistika – přednáška
- BMBS051 Biostatistika-základní kurz
  - Písemná zkouška (2 hodiny, povoleny materiály + nutná kalkulačka a tabulky statistických rozdělení, praktické řešení příkladů + teoretické otázky, klíčové je nalezení a popsání správného postupu, numerická správnost řešení nutná „pouze“ pro dosažení plného počtu bodů)

- Bi5040c Biostatistika – cvičení (nepovinný)
- ASTAc Biostatistika – cvičení (povinný)
  - Zápočtová písemka – bližší informace u vyučujících cvičení

- BLBS051p + BLBS051c – Biostatistika (sloučené)
  - Zjednodušená písemná zkouška (výběr z možných odpovědí, materiály povoleny)

- Předtermín zkoušky 20.12.2017, další termíny v lednu

Statistika ve vědecké praxi

Pozice statistické analýzy ve vědě a klinické praxi

Význam statistických výstupů

Anotace

- Statistická analýza biologických dat je jedním z nástrojů, s jejichž pomocí se snažíme zjistit odpovědi na naše otázky týkající se pochopení živé přírody.
- Jako každý nástroj je i statistickou analýzu nezbytné na jedné straně korektně využívat a na druhou stranu nepřeceňovat její možnosti.
- Klíčovým faktem při statistické analýze dat je nahlížení na realitu prostřednictvím vzorku a přijmutí toho, že výsledky naší analýzy jsou jen tak dobré, jak dobrý je náš vzorek.
- Reprezentativnost, nezávislost a náhodnost vzorku spolu s jeho velikostí jsou důležité faktory ovlivňující věrohodnost našich závěrů.

Life is beautiful with data analysis

Co znamená pro biologa/lékaře statistická analýza dat?

- **Matematická statistika** je vědecká disciplína na pomezí popisné statistiky a aplikované matematiky. Zabývá se teoretickým rozborem a návrhem metod získávání s analýzy empirických dat obsahujících prvek nahodilosti, tedy teorií plánování experimentů, výběrů, statistických odhadů, testování hypotéz a statistických modelů.
- **Statistika** je věda a postup jak rozvíjet lidské znalosti použitím empirických dat. Je založena na matematické statistice, která je větví aplikované matematiky.
- **Biostatistika** = aplikace statistické analýzy dat v biologickém a klinickém výzkumu
  - Nástroj pro uchopení dat našeho výzkumu
  - Nezbytné chápat principy a limitace
  - Není nutná detailní matematická znalost

- **Easy to understand, hard to master**

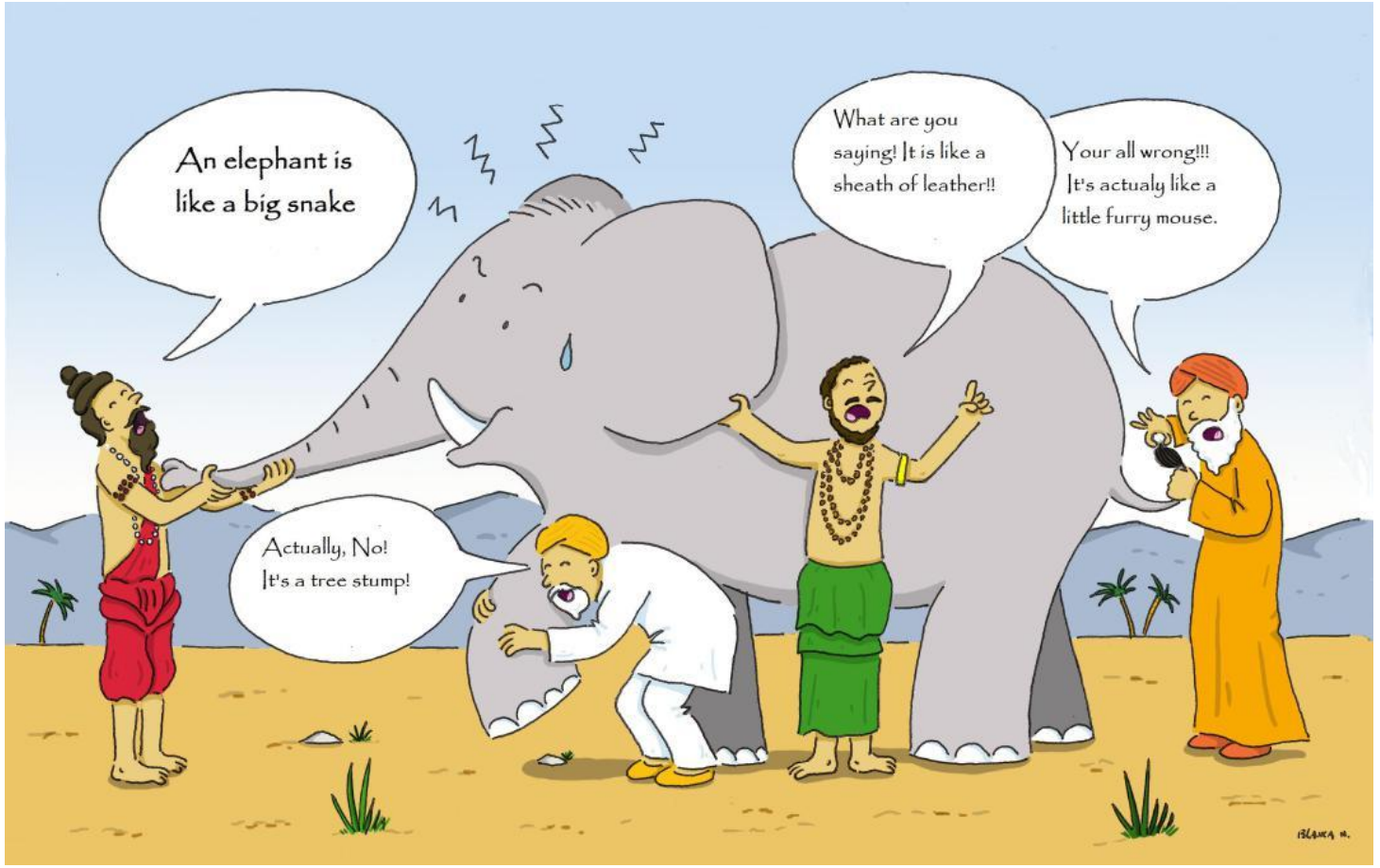
Výzkum, realita, statistika

- Výzkum je naším způsobem porozumění realitě
- Ale jak přesné a pravdivé je naše porozumění?

- **Statistika** je jedním z nástrojů umožňujícím popis a komunikaci výsledků výzkumu.
- Ale je to pouze nástroj, co je skutečně důležité jsou **data**.

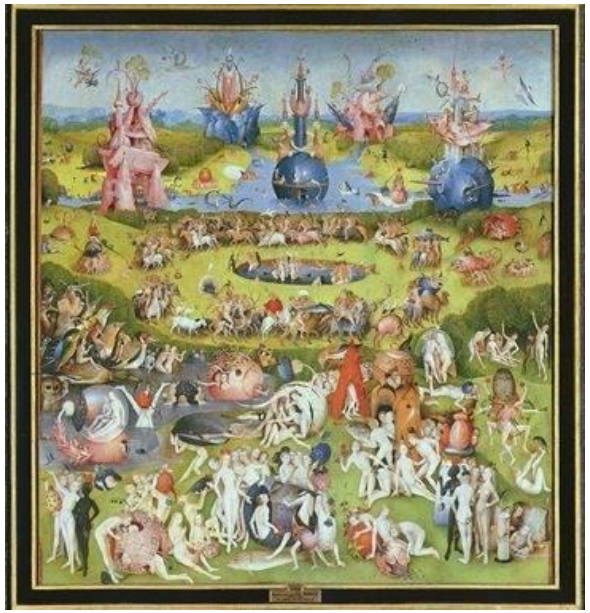
Realita a data

- Klíčovou otázkou výzkumu a následně statistické analýzy je jak dobře naše data popisují realitu
- Bez kvalitních dat není kvalitní statistiky ani kvalitního výzkumu.
- Každá chyba učiněná v úvodní fázi výzkumu se v dalších fázích znásobí a zřejmě ji již nebude možné eliminovat



Variabilita jako základní pojem ve statistice

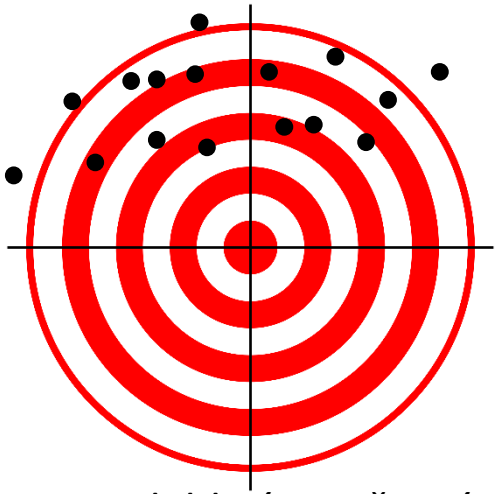
- Naše realita je variabilní a statistika je vědou zabývající se variabilitou
- Korektní analýza variabilita a její pochopení přináší užitečné informace o naší realitě
- V případě deterministického světa by statistická analýza nebyla potřebná



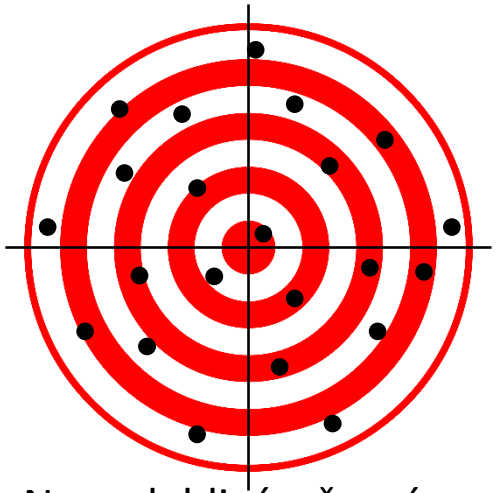
?

Spolehlivost a přesnost měření

- Kvalita dat je klíčová pro jakékoliv statistické hodnocení
- Bez spolehlivých a přesných dat není možné získat spolehlivé a přesné výsledky statistického hodnocení
- Ve statistické analýze dat musíme zohlednit jak střed měření, tak variabilitu a zamyslet se nad přesností popisu reality

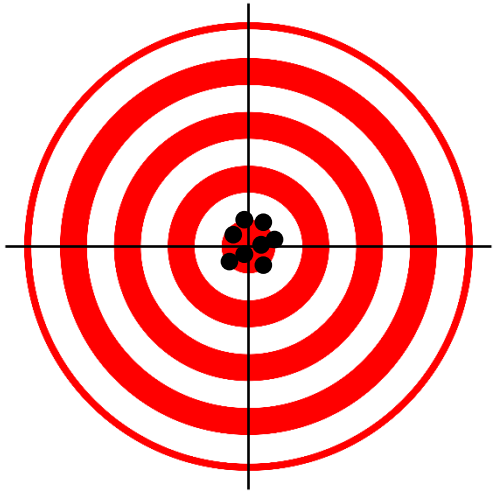


Nespolehlivý, nepřesný



Nespolehlivý, přesný

Spolehlivý, nepřesný



Spolehlivý, přesný

Variabilita a střední hodnota

- Norma = 5 gramů soli na 1 kg rýže

Nezamícháte

0g soli / 1 kg rýže

10g soli / 1 kg rýže

Průměr: 5g soli / 1 kg rýže

Vše OK !!!

Průměr není vše, je nezbytné zohlednit variabilitu

Zamícháte

5g soli / 1 kg rýže

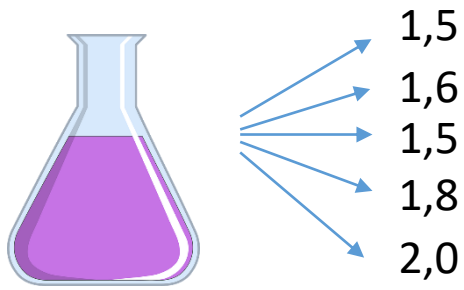
5g soli / 1 kg rýže

Průměr: 5g soli / 1 kg rýže

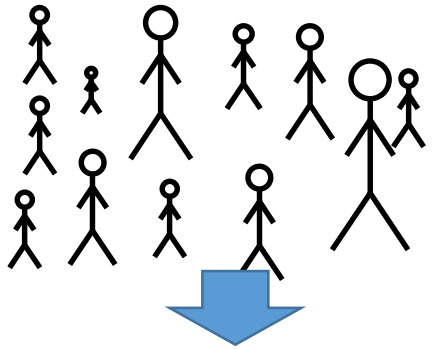
Vše OK !!!

Různé úrovně variability

Variabilita opakovaných měření

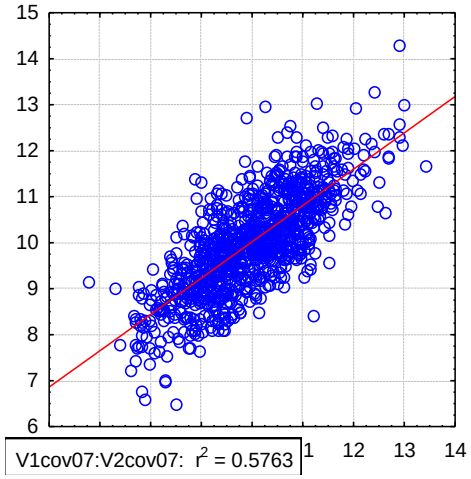


Variabilita dat v populaci

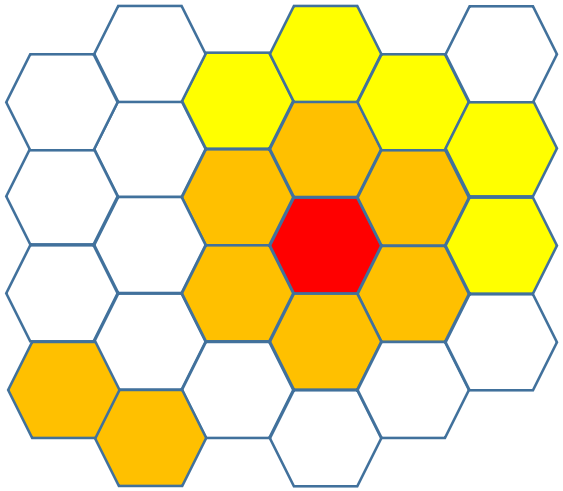


Hlavní téma kurzu

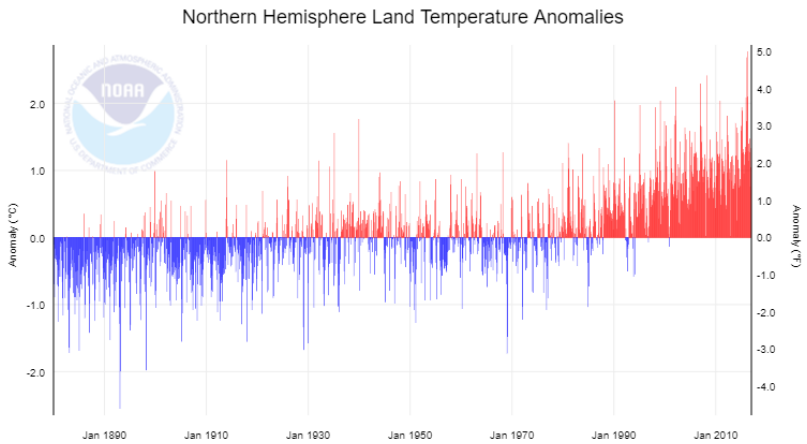
Variabilita v modelech



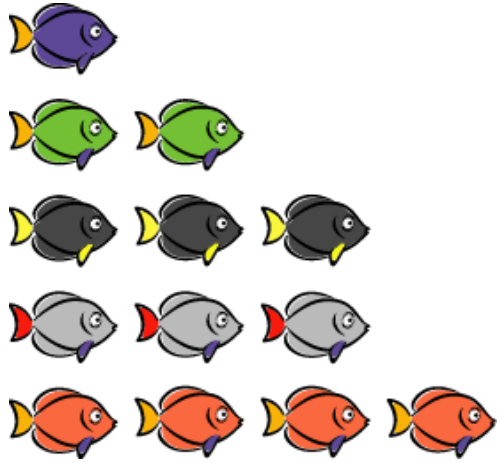
Geografická variabilita



Variabilita časových řad

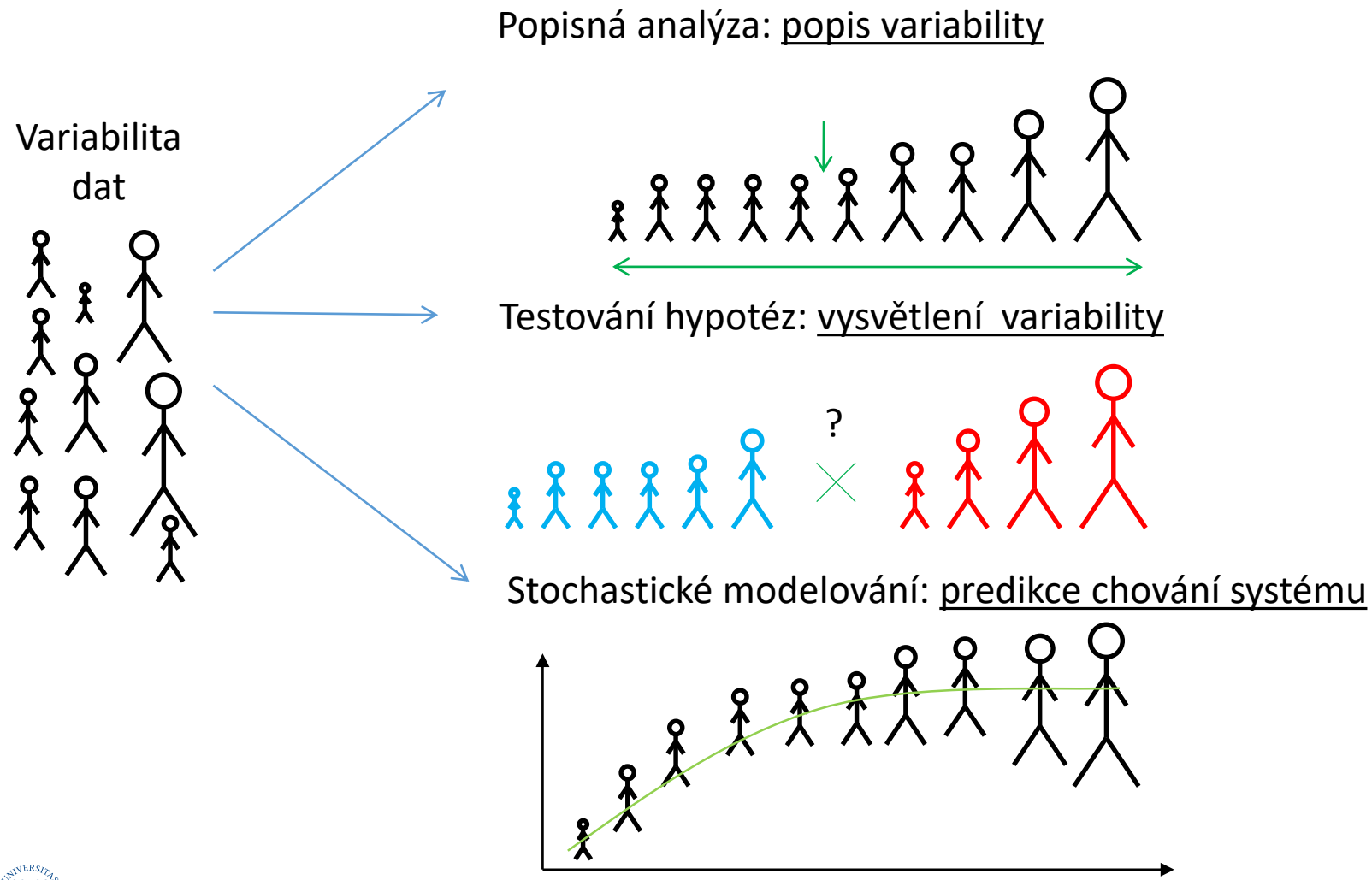


Biodiverzita



Práce s variabilitou v analýze dat

- V analýze dat existují dva hlavní přístupy k práci s variabilitou



Statistika – definice

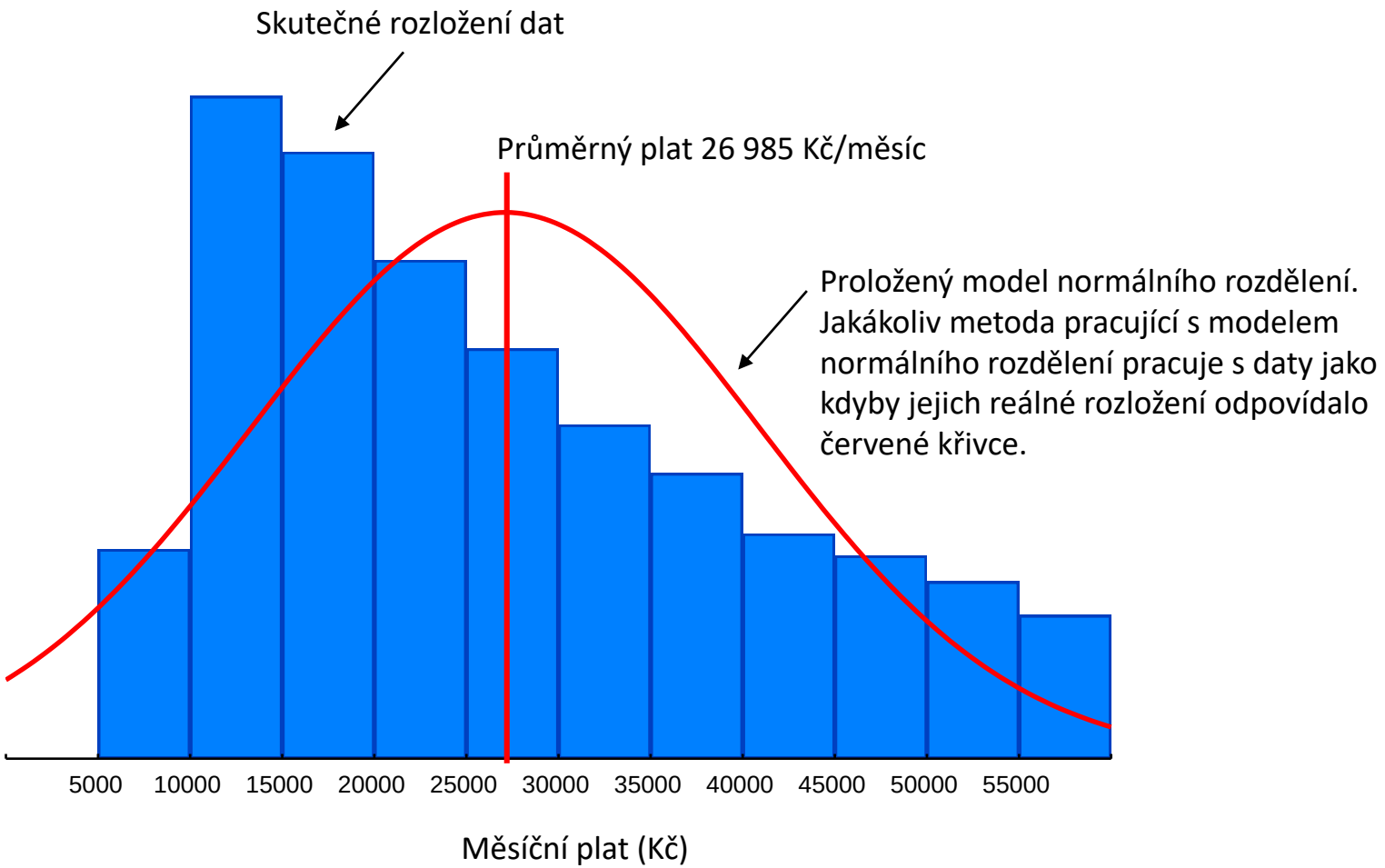
WWW.WIKIPEDIA.ORG:

Statistika je matematickou vědou zabývající se shromážděním, analýzou, interpretací, vysvětlením a prezentací dat. Může být aplikována v širokém spektru vědeckých disciplín od přírodních až po sociální vědy. Statistika je využívána i jako podklad pro rozhodování, kdy nicméně může být záměrně i nevědomky zneužita.

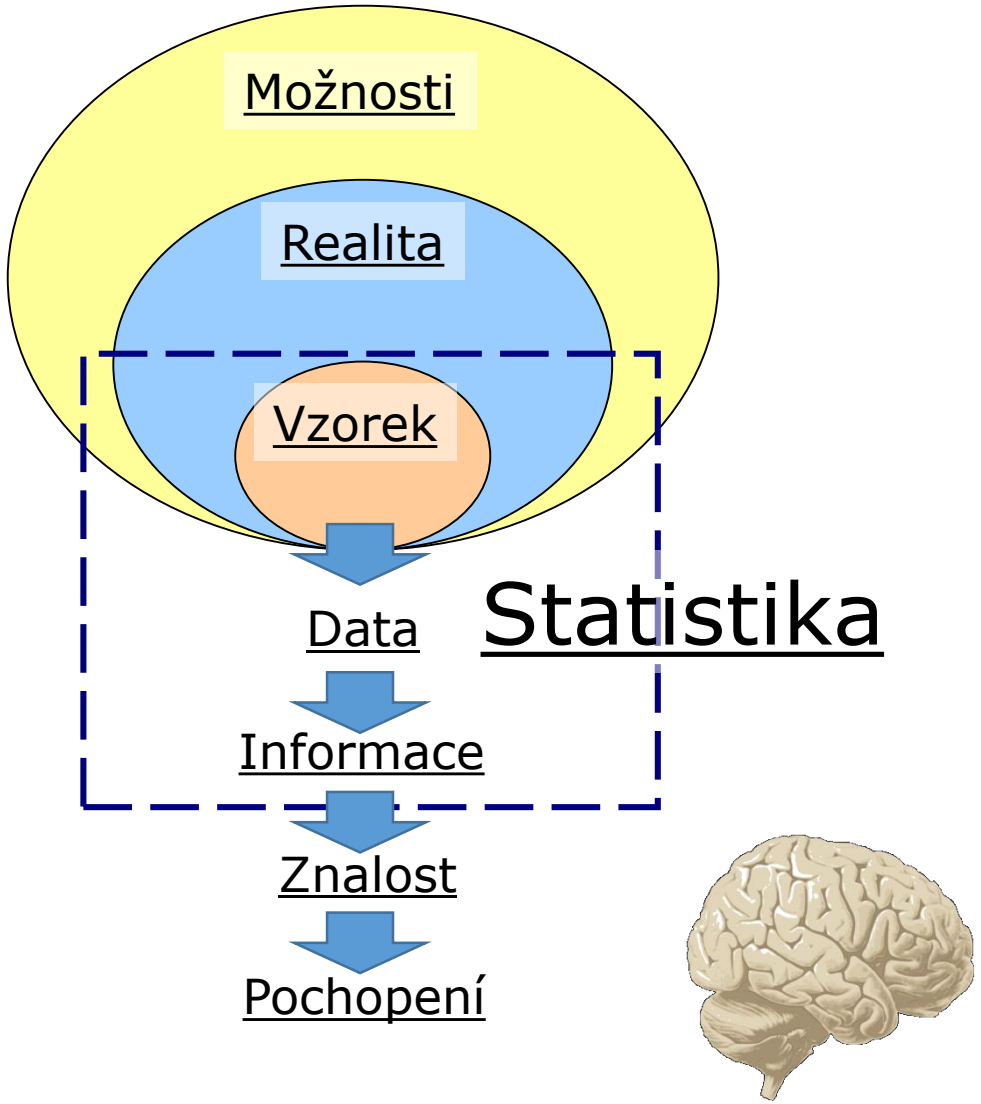
Statistika využívá matematické modely reality k zobecnění výsledků experimentů a vzorkování. Statistika funguje korektně pouze pokud jsou splněny předpoklady jejích metod a modelů.

Nesprávná aplikace modelu -> zkreslené závěry

- Různé popisné statistiky a testy jsou spjaty s různými modelovými rozděleními
- Pro správnou interpretaci je třeba ověřit shodu reálných dat s modelem
- Některé statistiky je možné vždy spočítat, ale jejich interpretace je v případě nedodržení předpokladů pouze omezená



Co může statistika říci o naší realitě?



Statistika není schopna činit závěry o jevech neobsažených v našem vzorku.

Statistika je nasazena v procesu získání informací z vzorkovaných dat a je podporou v získání naší znalosti a pochopení problému.

Statistika není náhradou naší inteligence !!!

Co musíme vědět před zahájením studie nebo experimentu?

- Cílová populace
  - Skupina objektů (pacientů, lokalit atd.) na něž je studie zaměřena

- Primární hypotézy
  - Hlavní otázka položená ve studii – odhad velikosti vzorku a design studie je vypracován vzhledem k primární hypotéze (v řadě případů nelze v reálném výzkumu formální power analýzu vypracovat, nicméně zamyšlení nad velikostí vzorku je nezbytné vždy)

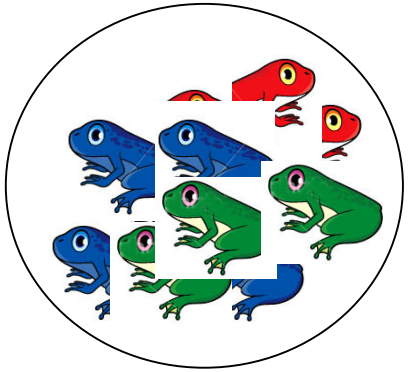
- Sekundární hypotézy
  - Vedlejší otázky, na něž by studie měla odpovědět

- Výběr adekvátní metodiky
  - Hypotézy jsou zodpovězeny prostřednictvím konkrétních proměnných (endpointů) – jejich typ (binární, kategoriální, spojité proměnné, biodiverzita, přežití, mortalita atd.) určuje výběr způsobu statistického zpracování

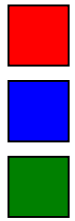
Cílová populace

- Cílová populace – klíčový pojem statistického zpracování
  - Skupina objektů o nichž se chceme něco dozvědět (např. lokality v daném povodí, laboratorní organismy v daných podmínkách, pacienti s danou diagnózou, všichni lidé nad 60 let, měření hemoglobinu v dané laboratoři)
  - Musí být definována ještě před zahájením sběru dat
  - Na cílové populaci probíhá vzorkování dat, které musí cílovou populaci dobře (reprezentativně) charakterizovat

Cílová populace



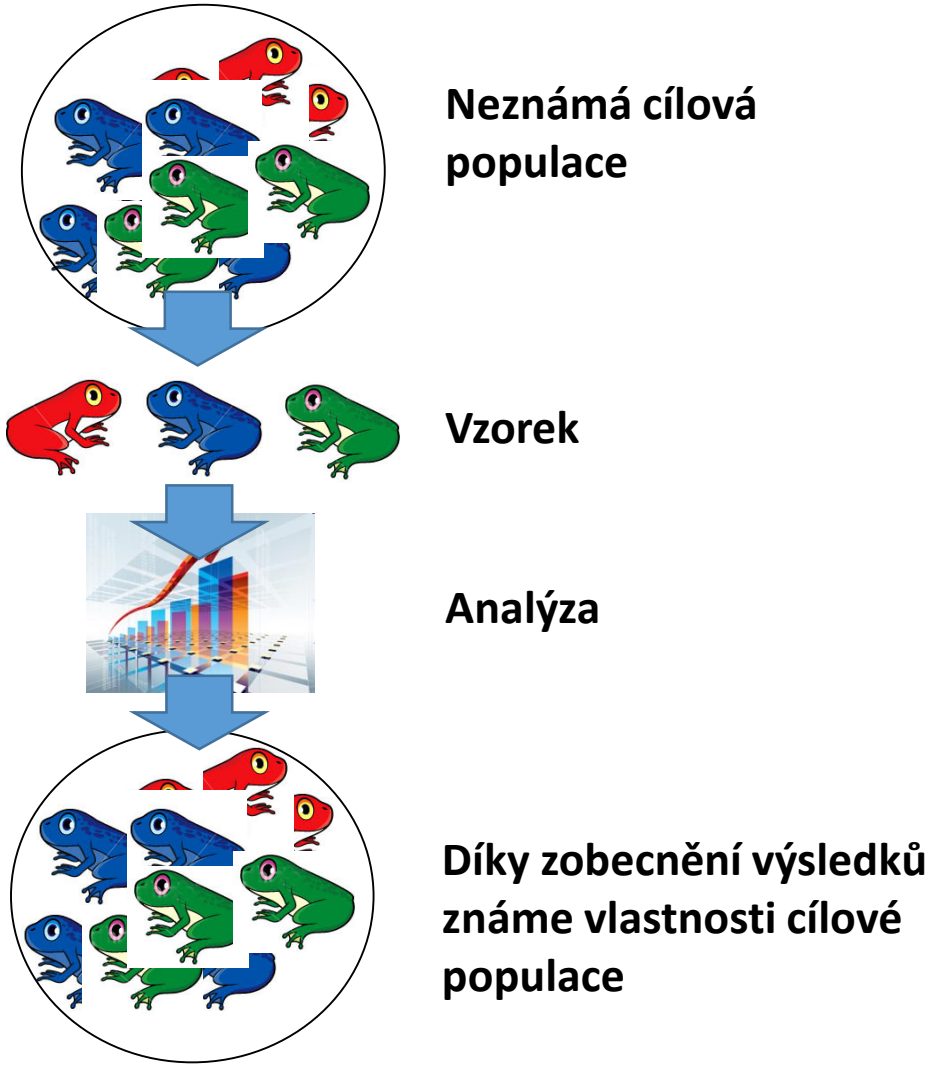
Klíčové faktory cílové populace



Design experimentu a vzorkovací plán

Vzorkování a analýza dat

Statistika a zobecnění výsledků



Neznámá cílová populace

Vzorek

Analýza

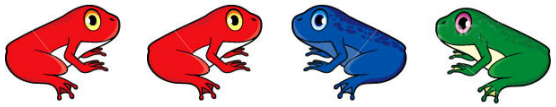
Díky zobecnění výsledků známe vlastnosti cílové populace

- Cílem analýzy není pouhý popis a analýza vzorku, ale zobecnění výsledků ze vzorku na jeho cílovou populaci
- Pokud vzorek nereprezentuje cílovou populaci, vede zobecnění k chybným závěrům

Vzorkování a jeho význam ve statistice

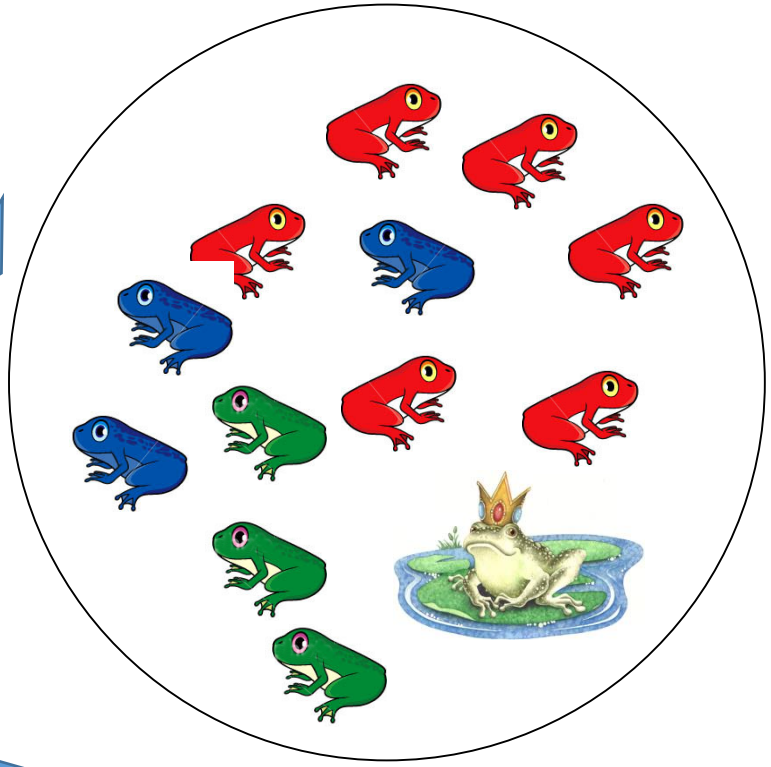
- Statistika hovoří o realitě prostřednictvím vzorku!!!
- Statistické předpoklady korektního vzorkování

- **Representativnost:** struktura vzorku musí maximálně reflektovat realitu

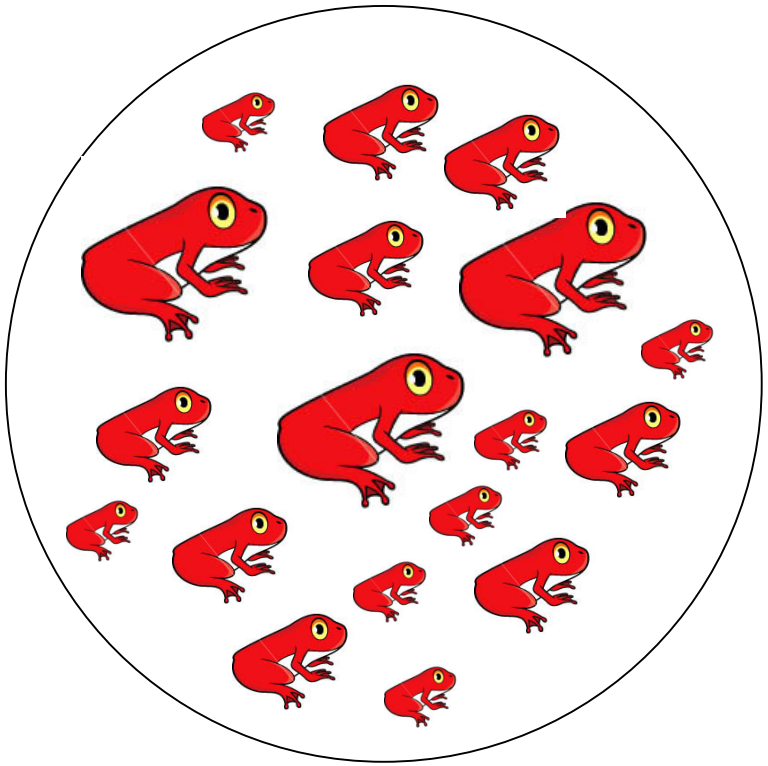


- **Nezávislost:** několikanásobné vzorkování téhož objektu nepřináší ze statistického hlediska žádnou novou informaci

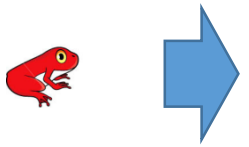
- **Náhodnost:** zajišťuje náhodný vliv zavádějících faktorů



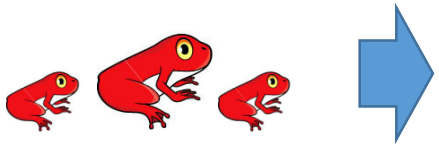
Velikost vzorku a spolehlivost statistických výstupů



- Existuje skutečné rozložení a skutečná střední hodnota měřené proměnné
- Z jednoho měření nezjistíme nic



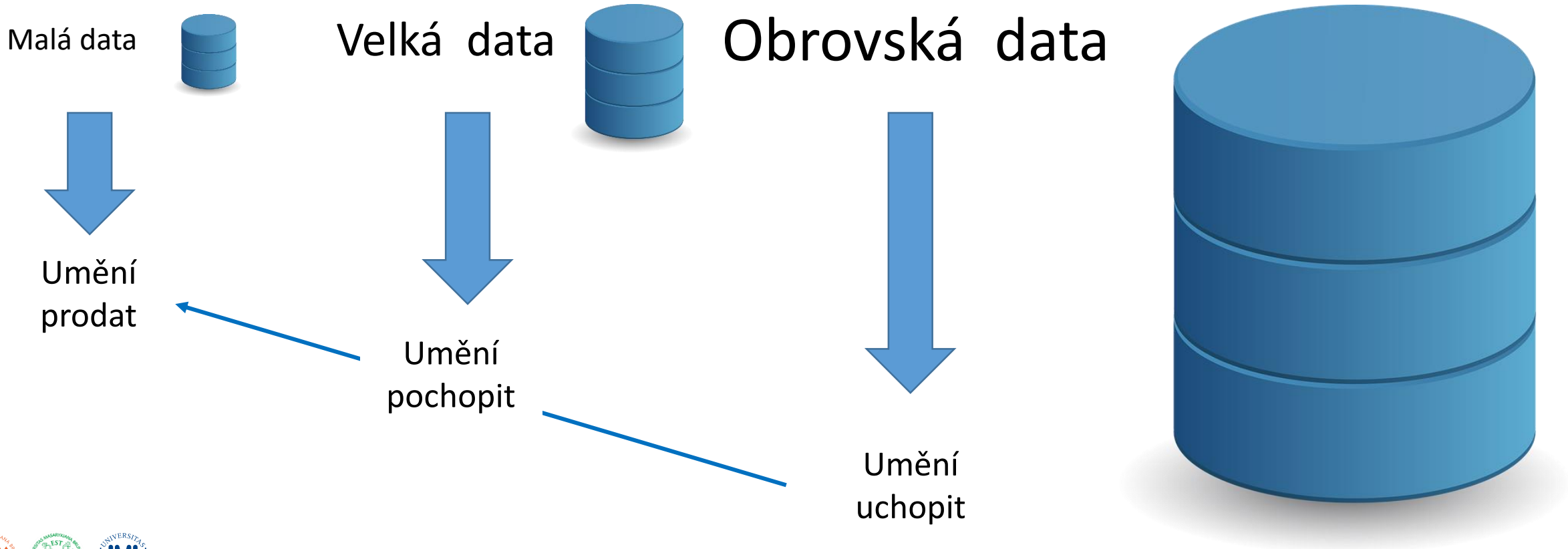
- Vzorek určité velikosti poskytuje odhad reálné hodnoty s definovanou spolehlivostí



- Vzorkování všech existujících objektů poskytne skutečnou hodnotu dané popisné statistiky, nicméně tento přístup je ve většině případech nereálný.

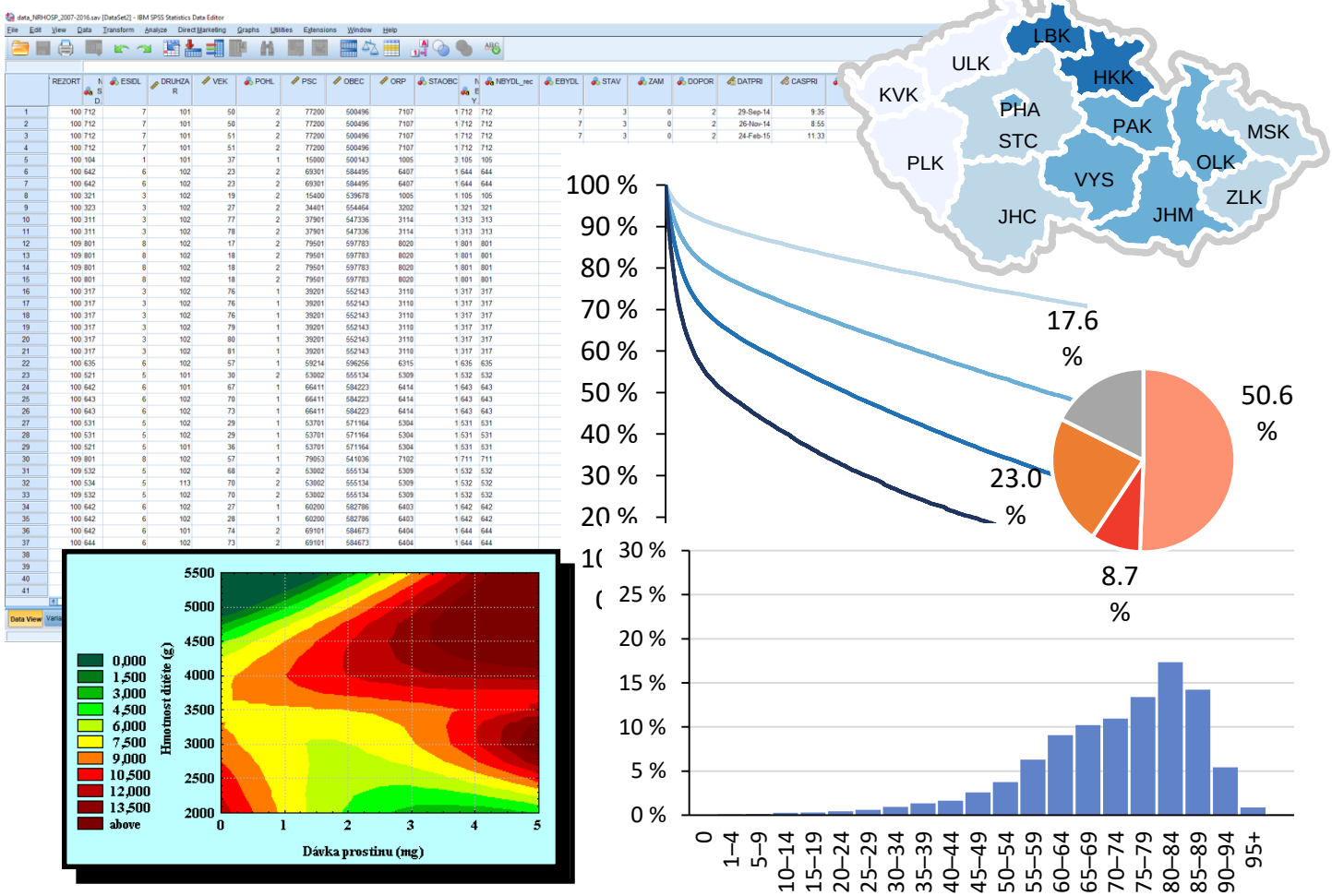
Různá velikost vzorku – různé úkoly analýzy dat

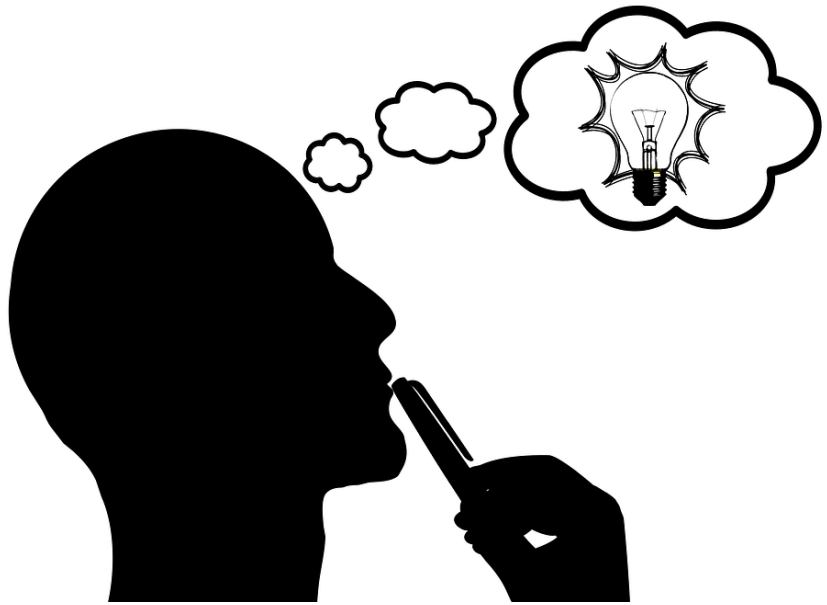
- Náročnost analýzy dat stoupá i s jejich objemem
- I u největších dat stále platí, že klíčová je schopnost data prodat = smysluplně interpretovat a prezentovat



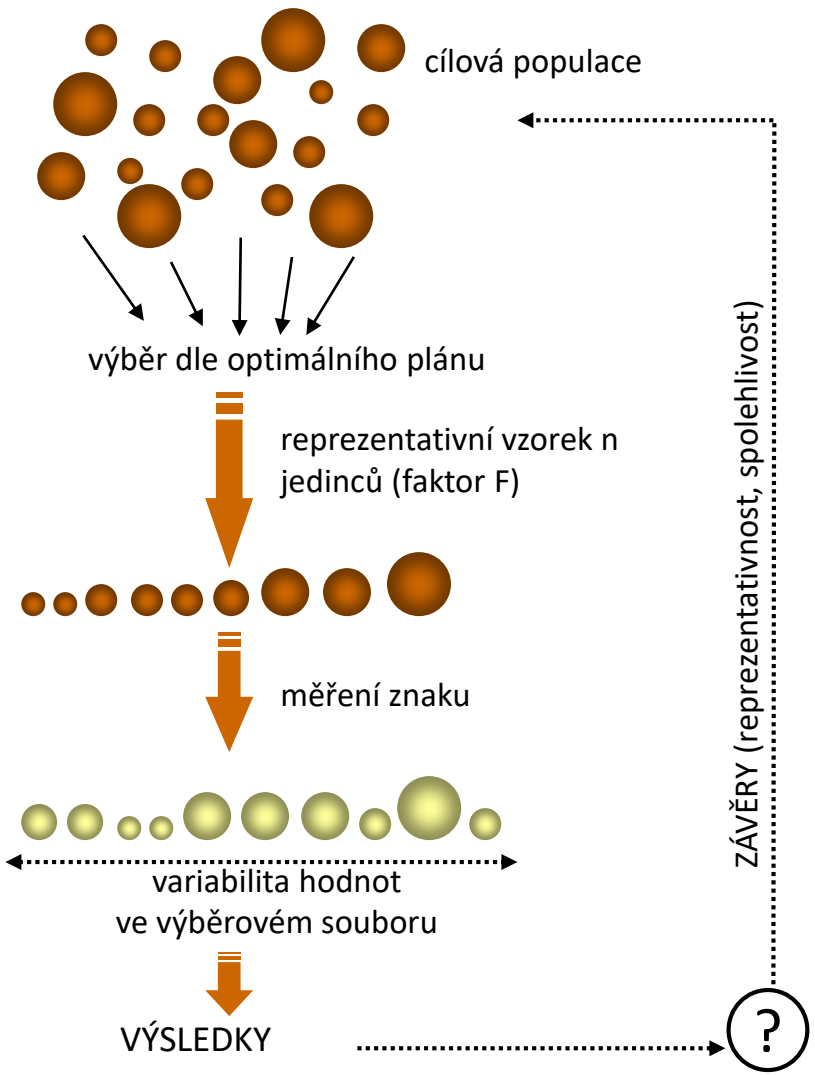
Přístup biostatistiky

- Schopnost: vidět data – komunikovat – interpretovat - prodávat





Experimentální design: nezbytná výbava biologa



Účel analýzy: Popisný

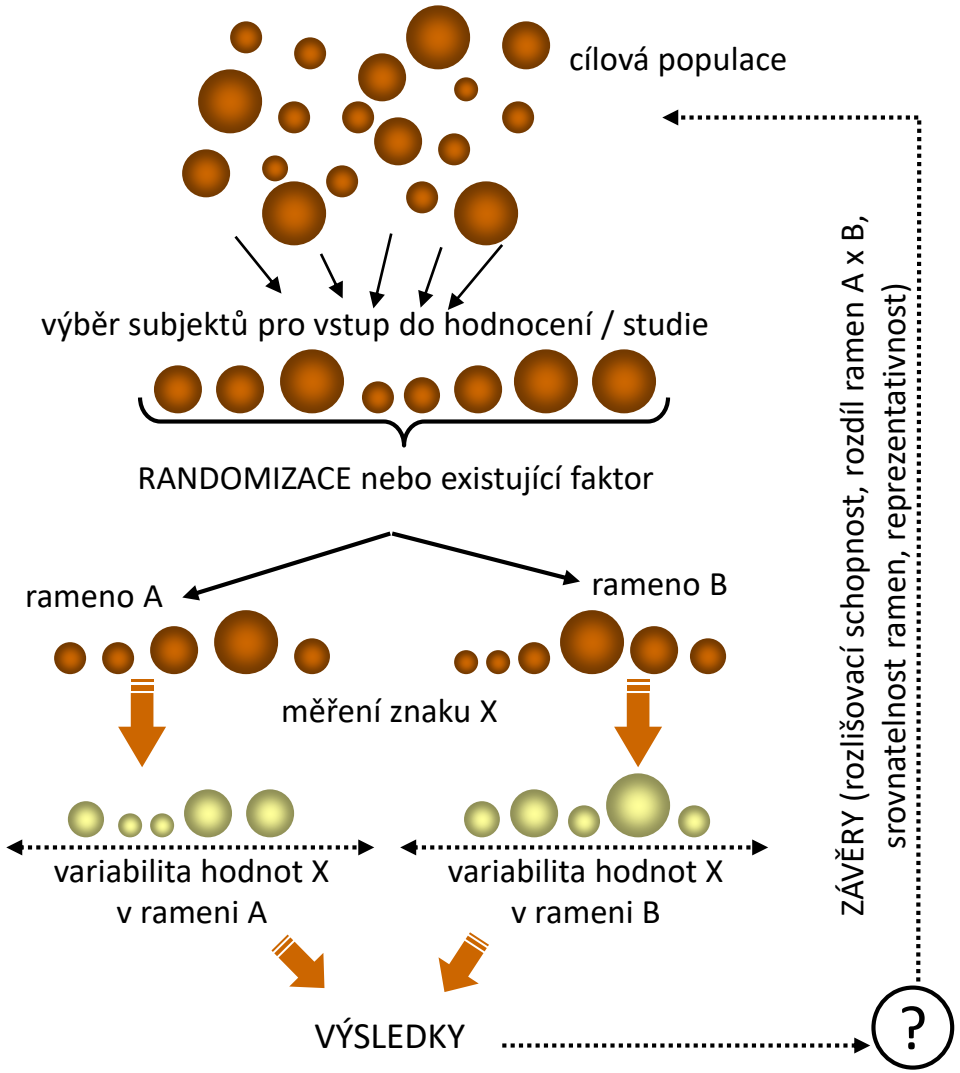
?

Reprezentativnost

Spolehlivost

Přesnost

Experimentální design: nezbytná výbava biologa



Účel analýzy: Srovnávací (2 skupiny)

?

Reprezentativnost

Srovnatelnost

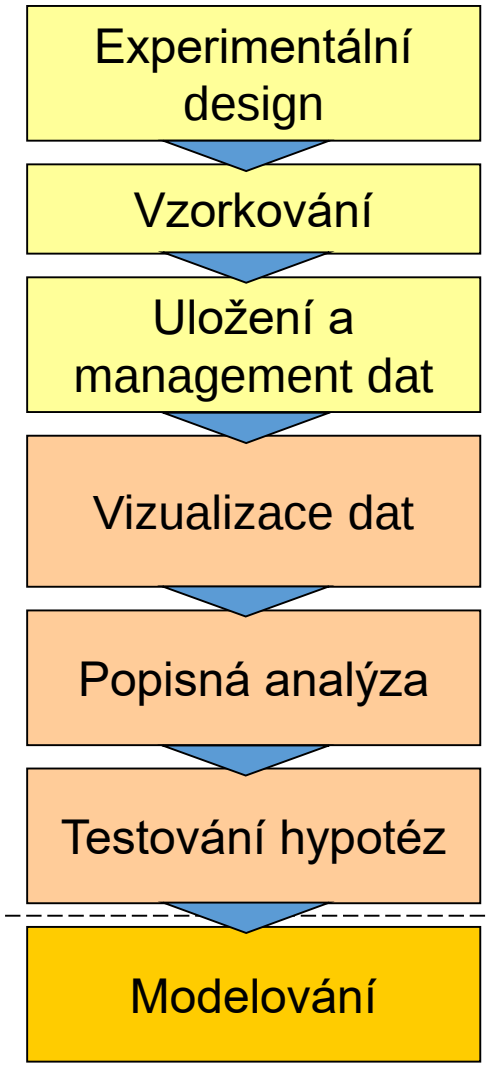
Spolehlivost

Přesnost

… analyzovaný znak cílové populace (X)

… jiný významný faktor charakterizující cílovou populaci (F)

Obecné schéma využití statistické analýzy



Jak velký vzorek je nezbytný pro statisticky relevantní výsledky? Klíčová stratifikační kritéria cílové populace.

Vzorkovací plán zabezpečující náhodnost a reprezentativnost vzorku.

Uložení dat ve vhodné formě a jejich vyčištění předcházející vlastní analýze je klíčovým krokem statistické analýzy.

Grafická inspekce dat je nezbytným krokem analýzy vzhledem ke schopnosti lidského mozku primárně akceptovat obrazová data. Poskytne vhled do dat, představu o jejich rozložení, vazbách proměnných apod.

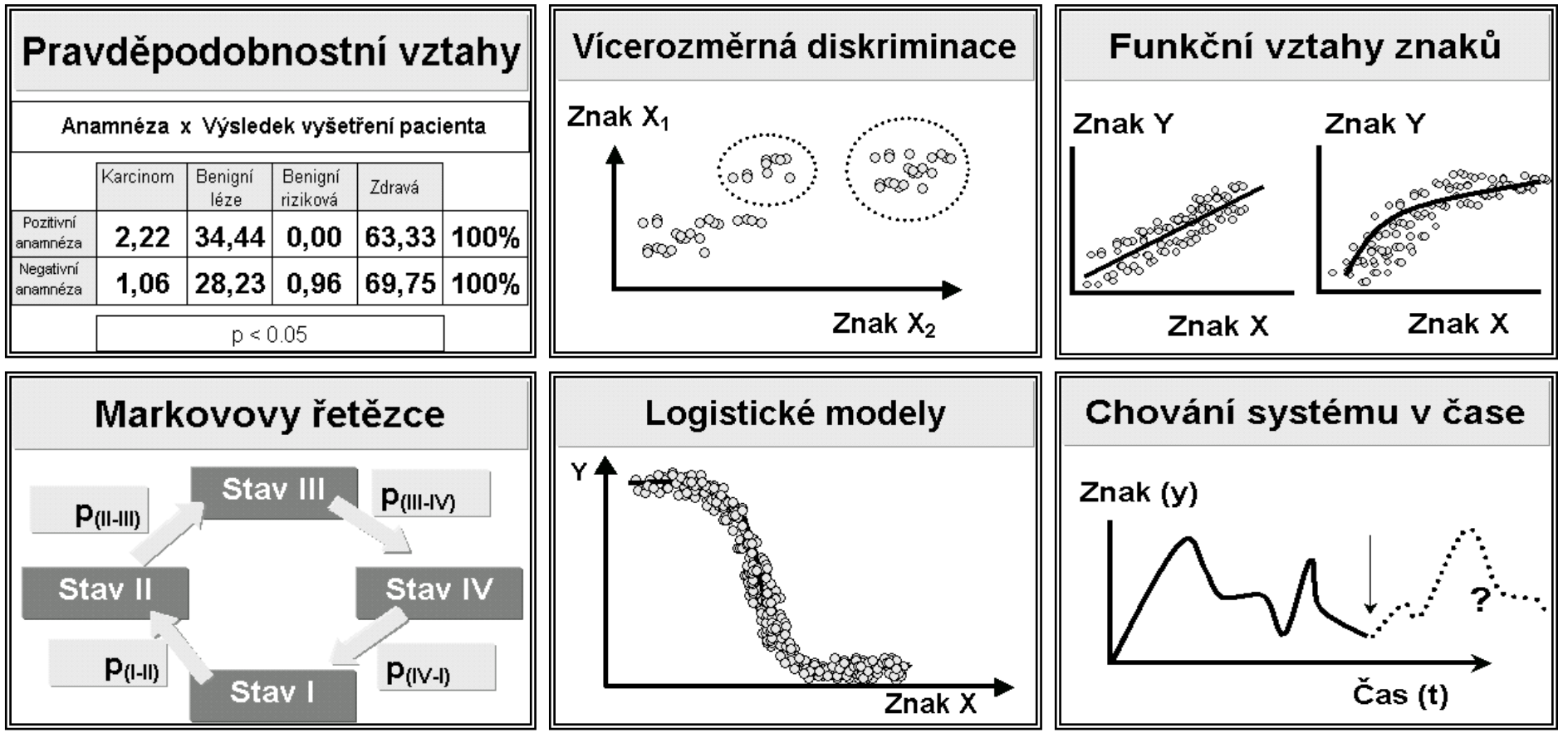
Popisná analýza umožňuje vyhodnotit srovnáním s existující literaturou realističnost naměřených rozsahů dat.

Testování vazeb mezi různými proměnnými s cílem navzájem vysvětlit jejich variabilitu a tím přispět k pochopení řešeného problému.

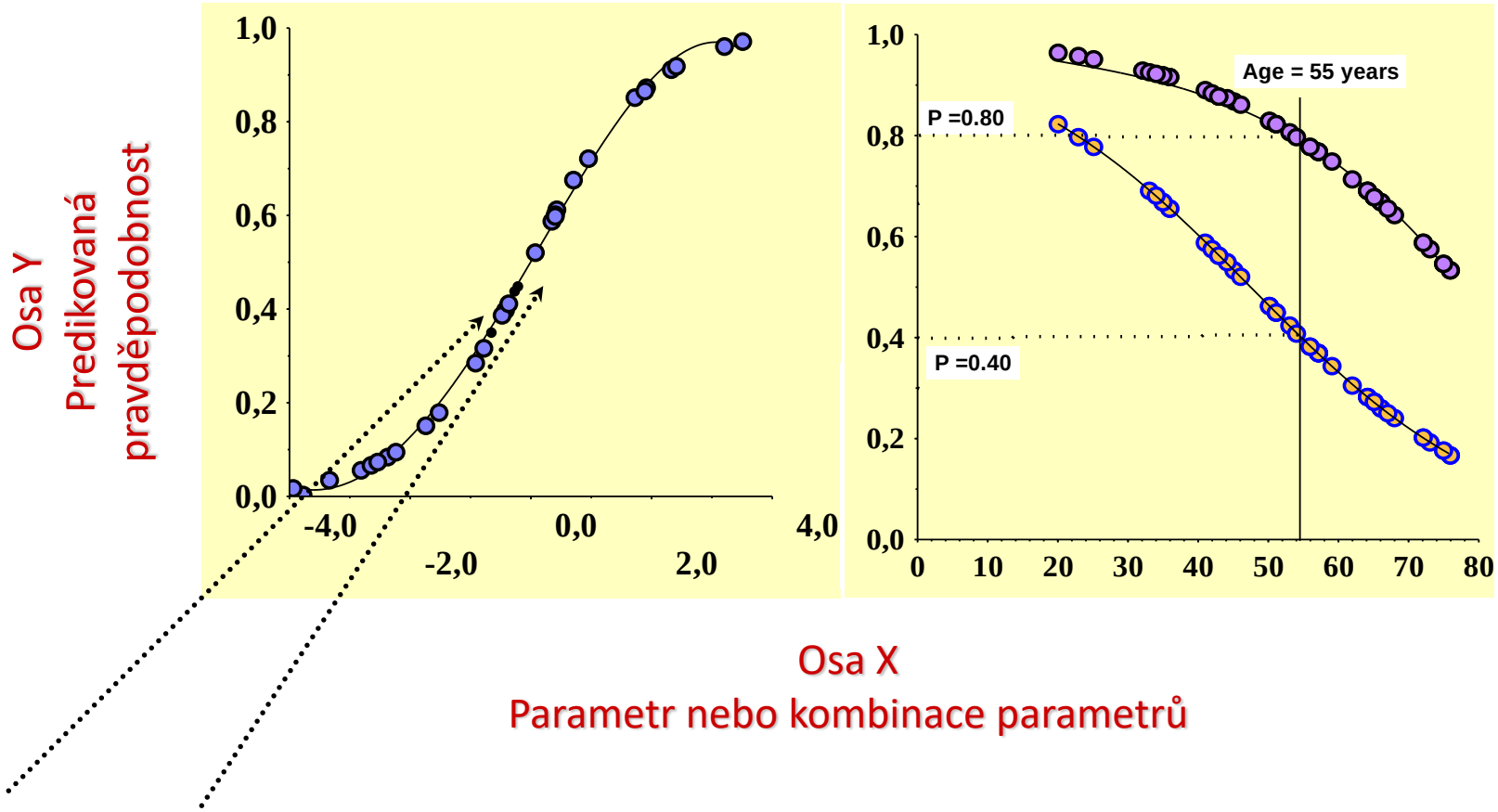
Možným vyvrcholením analýzy je využití získaných znalostí a pochopení problému k vytvoření prediktivních modelů.

Stochastické modelování: predikce neurčitých jevů

- Prospektivně – modelově - postihuje chování jevů při respektování variability



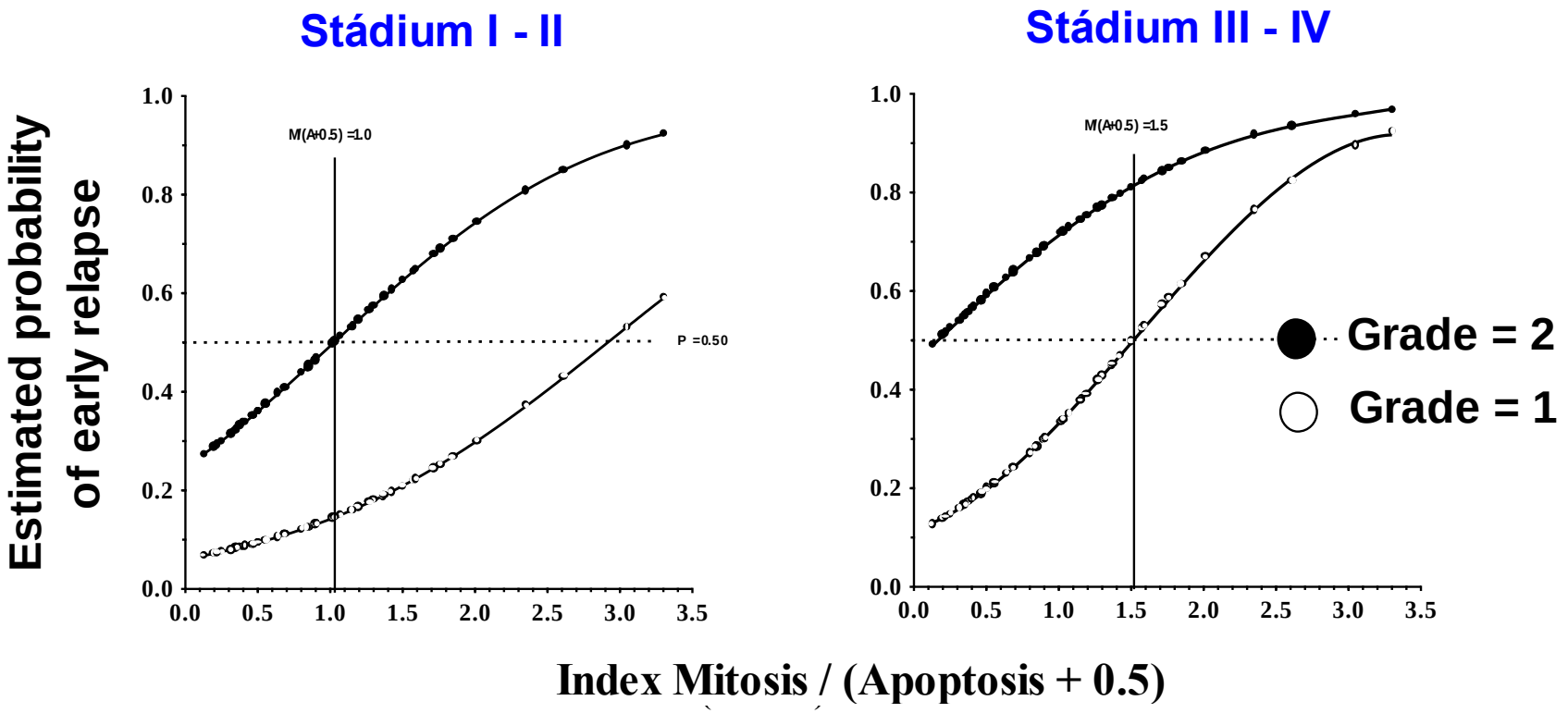
Stochastické modelování: predikce neurčitých jevů



Data konkrétních objektů k přímému hodnocení

Stochastické modelování: predikce neurčitých jevů

- Schopnost: vytvářet prakticky využitelné nástroje



Přednáška 2

Klíčové principy biostatistiky

Zkreslení, reprezentativnost, srovnatelnost, spolehlivost významnost

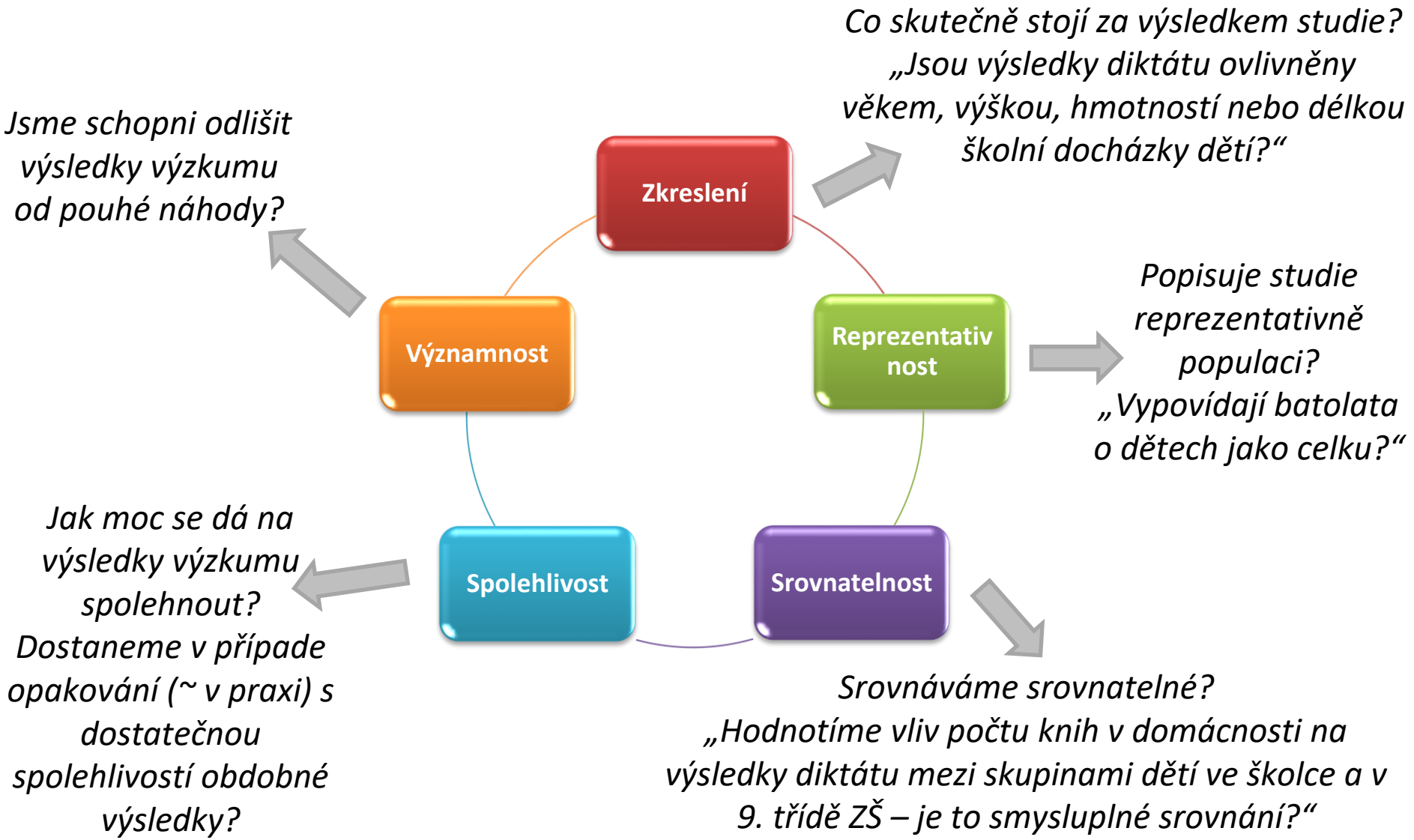
Anotace

- Ve statistické analýze biologických a klinických dat musíme vždy nad prováděným výzkumem a jeho výsledky přemýšlet v kontextu 5 klíčových principů biostatistiky.

- Zkreslení – skutečně vidíme to co si myslíme, že vidíme?
- Reprezentativnost – vypovídá naše analýza o skupině objektů, která nás zajímá?
- Srovnatelnost – co ve skutečnosti v analýze srovnáváme?
- Spolehlivost – jak spolehlivé jsou naše výsledky, dají se zopakovat?
- Významnost – jak moc je pravděpodobné, že pozorujeme výsledky pouhé náhody?

- Zanedbání těchto principů může vést k chybné interpretaci výsledků.

Klíčové principy biostatistiky



Klíčové principy – zkreslení

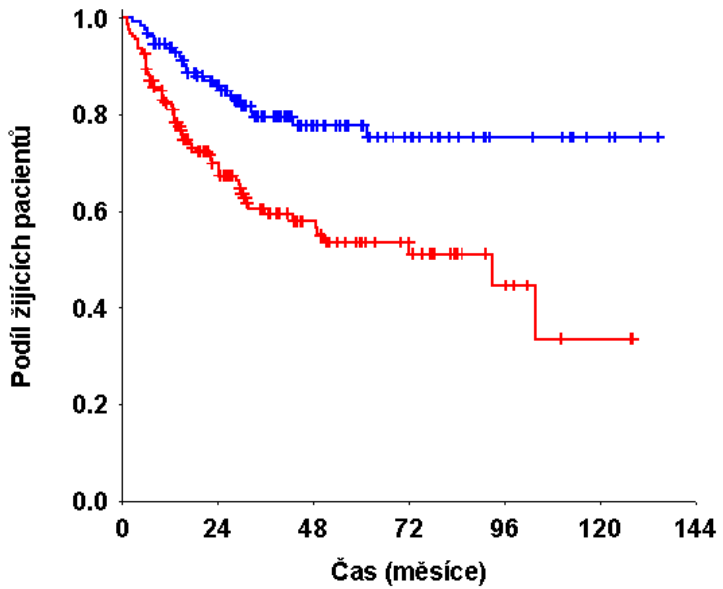
- V jakémkoliv hodnocení se snažíme vyhnout zkreslení výsledků („biased results“) – tedy zkreslení výsledků jinými faktory než těmi, které jsou cíli výzkumu.

- Statistické srovnání není nikdy 100% spolehlivé, existuje náhoda a tedy i pravděpodobnost chybného úsudku – to nelze ovlivnit.

- Chceme použít adekvátní metody pro odstranění vlivů, které by zkreslily výsledky a nebyly přitom náhodné (např. zastoupení pohlaví, nadmořská výška).

Klíčové principy – zkreslení

- Co způsobuje rozdíl v saprobním znečištění vodního toku?
- Co způsobuje rozdíl v naměřených biochemických ukazatelích?
- Čím by mohl být způsoben pozorovaný rozdíl v 10letém přežití pacientů?



Léčba?

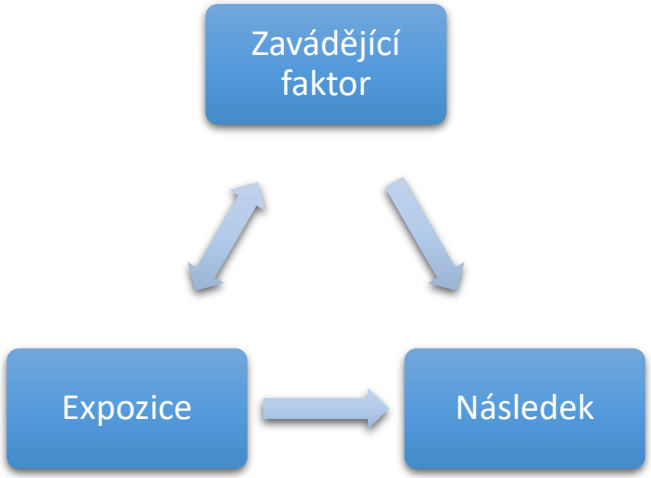
Nějaký prognostický faktor?

Stadium nemoci?

Věk?

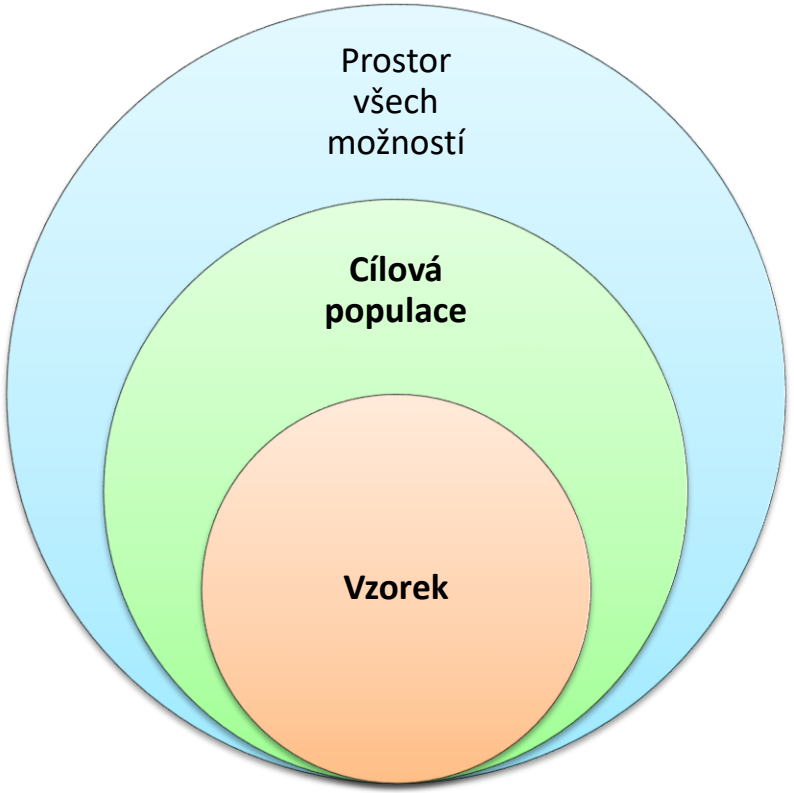
Klíčové principy – zkreslení

- Pojem zavádějící faktor
- Pro zavádějící faktor současně platí, že
  - přímo nebo nepřímo ovlivňuje sledovaný následek,
  - je ve vztahu se studovanou expozicí ,
  - není mezikrokem mezi expozicí a následkem.

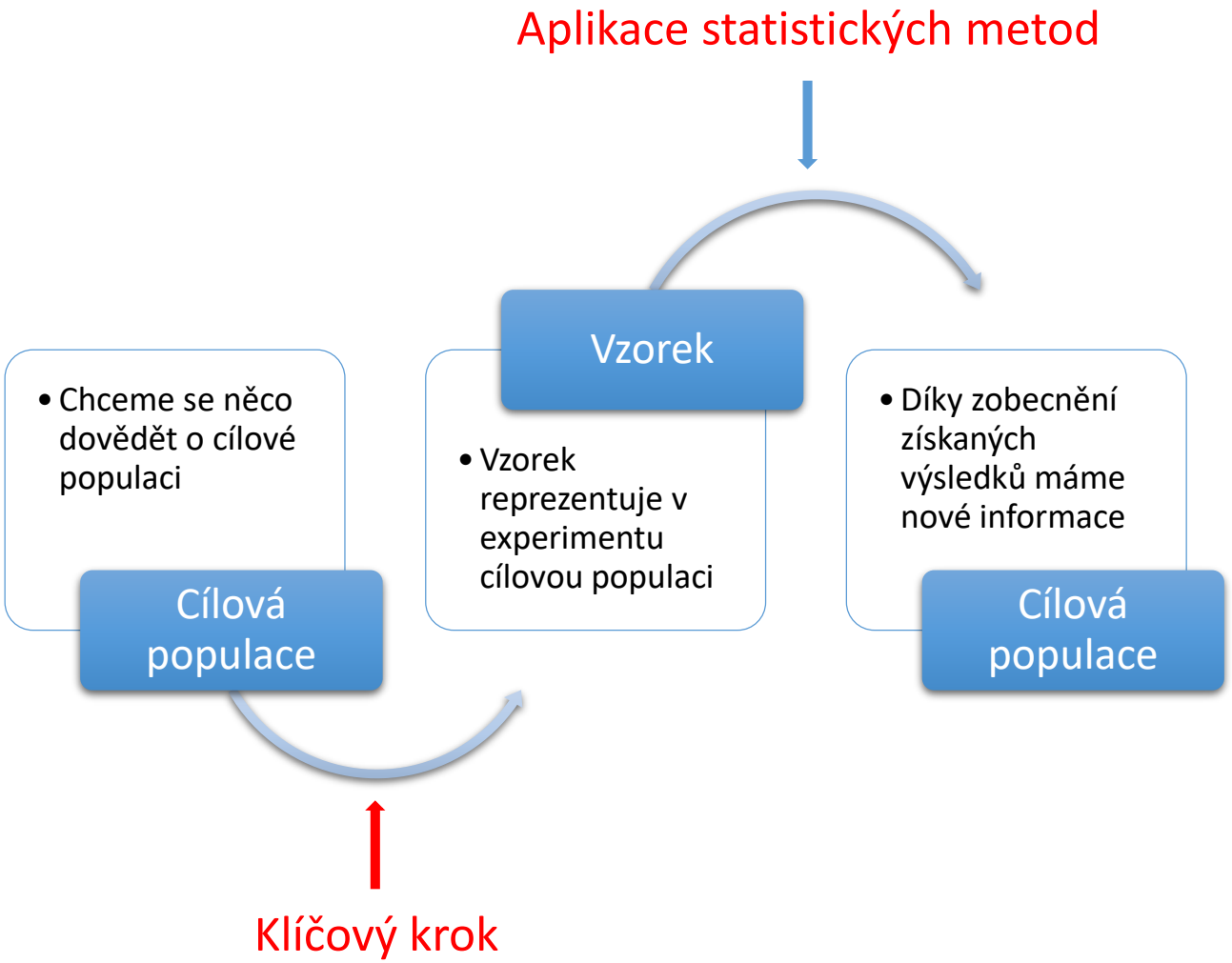


Klíčové principy – reprezentativnost

- Pojem cílová populace – skupina subjektů, o které chceme zjistit nějakou informaci.
- Pojem experimentální vzorek – podskupina cílové populace, kterou „máme k dispozici“.
  - Musí odpovídat svými charakteristikami cílové populaci.
  - Chceme totiž zobecnit výsledky na celou cílovou populaci.
  - Souvislost s náhodným výběrem.



Klíčové principy – reprezentativnost



Klíčové principy – srovnatelnost

- Korektní výsledky při srovnávacích analýzách lze získat pouze při srovnávání srovnatelného.
- V striktně kontrolovaných studiích je srovnatelnost zajištěna randomizací.
- U studií bez randomizace je nutné se tématu srovnatelnosti skupin věnovat.
- Metody adjustace, matching, propensity scores.

Klíčové principy – spolehlivost

- Ve většině studií nás zajímá kvantifikace sledovaného efektu nebo charakteristiky, obecně náhodné veličiny, ve formě jednoho čísla, bodového odhadu.

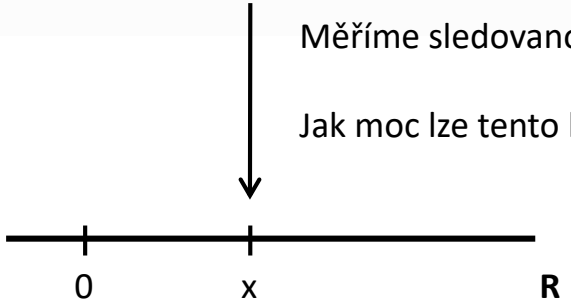
- Bodový odhad je však sám o sobě nedostatečný.

- Je nutné ho doplnit intervalovým odhadem, který odpovídá pravděpodobnostnímu chování sledované veličiny, tedy odpovídá určité spolehlivosti výsledku.

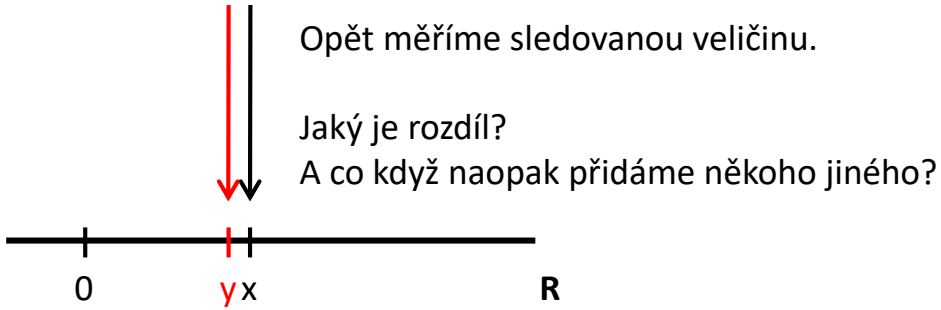
Klíčové principy – spolehlivost

Měříme sledovanou veličinu a následně spočítáme odhad.

Jak moc lze tento bodový odhad zobecnit na cílovou populaci?

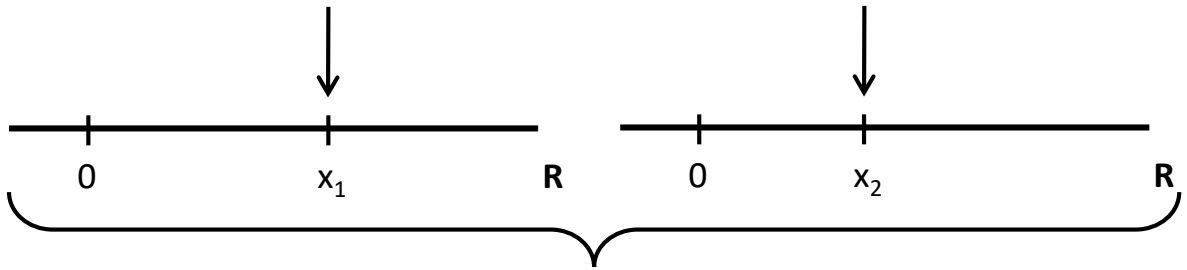


Klíčové principy – spolehlivost



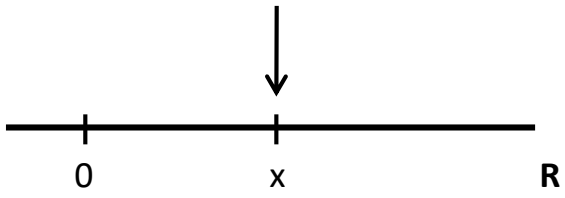
Klíčové principy – spolehlivost

Výběr číslo 1

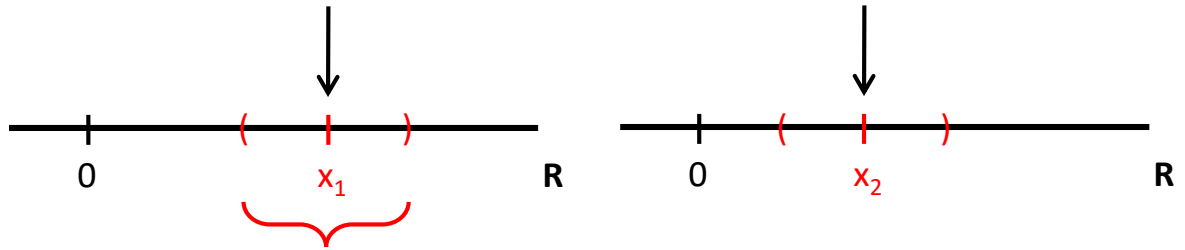


Výběr číslo 2

Celá cílová populace



Pracujeme-li s výběrem z cílové populace, je třeba na základě variability pozorovaných dat spočítat tzv. interval spolehlivosti pro bodový odhad.



Interval spolehlivosti na základě výběru číslo 1.

Umíme-li „změřit“ celou cílovou populaci, nepotřebujeme interval spolehlivosti, protože jsme schopni odhadnout sledovaný parametr přesně – v praxi je tato situace nereálná.

Klíčové principy – významnost

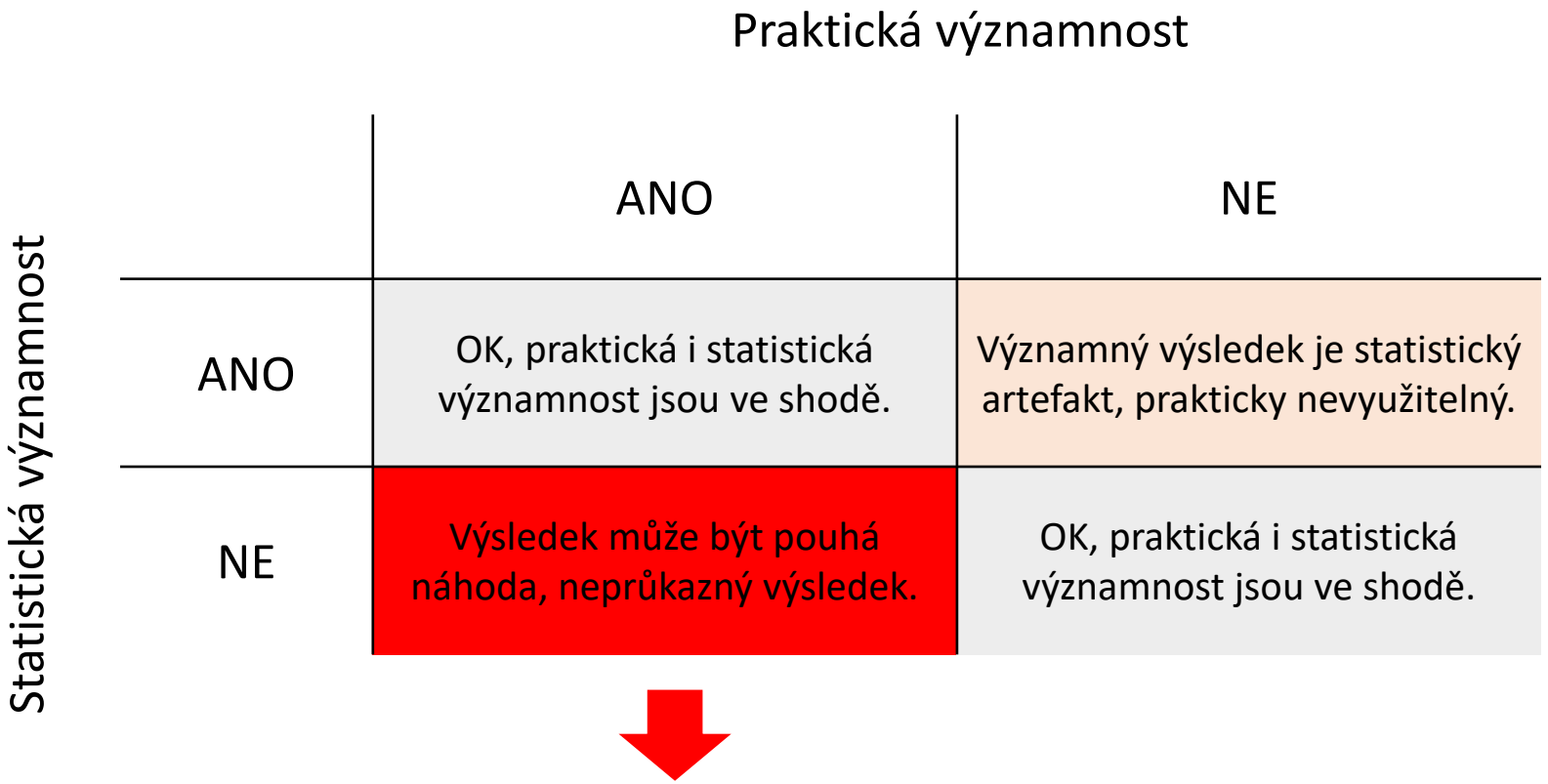
- Analytické výsledky studie nemusí odpovídat realitě a skutečnosti. Statistická významnost jednoduše nemusí znamenat příčinný vztah!

- Statistická významnost pouze indikuje, že pozorovaný rozdíl není náhodný (ve smyslu stanovené hypotézy).

- Stejně důležitá je i praktická významnost, tedy významnost z hlediska lékaře nebo biologa.

- Statistickou významnost lze ovlivnit velikostí vzorku.

Klíčové principy – významnost



Statisticky nevýznamný výsledek neznamená, že pozorovaný rozdíl ve skutečnosti neexistuje! Může to být způsobeno nedostatečnou informací v pozorovaných datech!

Příprava dat

Klíčový význam korektního uložení získaných dat

Pravidla pro ukládání dat

Čištění dat před analýzou

Anotace

- Současná statistická analýza se neobejde bez zpracování dat pomocí statistických software.
- Předpokladem úspěchu je správné uložení dat ve formě „databázové“ tabulky umožňující jejich zpracování v libovolné aplikaci.
- Neméně důležité je věnovat pozornost čištění dat předcházející vlastní analýze.
- Každá chyba, která vznikne nebo není nalezeno ve fázi přípravy dat se promítne do všech dalších kroků a může zapříčinit neplatnost výsledků a nutnost opakování analýzy.

DATA – ukázka uspořádání datového souboru

Parametry, znaky, charakteristiky, proměnné

Záznamy

Pacient	Clovek	aLeu	aTy%	aSe%	aNeu%	aLy%	aTy	aSe	aNeu	aLy	aHtc	aCLsk	aCLNeus	aCLOZ	aCLNeuO
		cell.10^6/	%	%	%	%	cell.10^6/	cell.10^6/	cell.10^6/	cell.10^6/	%	mV.s.10^3	mV.s.10^3	mV.s.10^3	mV.s.10^3
3	1	4									33	72		32	
4	2	7,6	8	58	66	24	0,6	4,4	5,0	1,8	33	95	19	48	10
8	3	4	3	52	55	40	0,1	2,1	2,2	1,6	22	77	35	33	15
11	4	6,1	5	59	64	35	0,3	3,6	3,9	2,1	33	103	26	49	13
12	5	6,9	3	85	88	9	0,2	5,9	6,1	0,6	37	81	13	45	7
14	6	5,9	15	55	70	19	0,9	3,3	4,1	1,1	32	137	33	61	15
16	7	8	18	75	93	7	1,4	6,0	7,4	0,6	34	151	20	59	8
20	8	9,6	3	72	75	23	0,3	6,9	7,2	2,2	40	77	11	38	5
21	9	6	10	67	77	19	0,6	4,0	4,6	1,1	32	120	26	52	11
22	10	3,3	4	55	59	39	0,1	1,8	2,0	1,3	28	81	42	24	12
37	11	3,8	10	60	70	30	0,4	2,3	2,7	1,1	32	111	42	29	11
38	12	6,4	2	76	78	17	0,1	4,9	5,0	1,1	25	366	73	115	23
39	13	6,8	1	57	58	39	0,1	3,9	3,9	2,7	20	234	59	71	18
49	14	8,5	7	67	74	26	0,6	5,7	6,3	2,2	30	156	25	108	17
51	15	9,3	7	57	64	35	0,7	5,3	6,0	3,3	35	129	21	23	4
52	16	2,2	10	56	66	34	0,2	1,2	1,5	0,7	33	46	30	12	8
55	17	9,9	3	78	81	10	0,3	7,7	8,0	0,1	30	189	24	140	18
56	18	5	2	80	82	13	0,1	4,0	4,1	0,7	26	101	25	54	13
6	1	8,8	11	72	83	12	1,0	6,3	7,3	1,1	44	268	36,6	145	19,9
9	2	9,2	2	66	68	28	0,2	6,1	6,3	2,6	42	168	26,9	76	12,2
13	3	10,0	7	83	90	8	0,7	8,3	9,0	0,8	54	181	20,1	81	9
15	4	9,6	1	75	76	23	0,1	7,2	7,3	2,2	45	343	47	124	16,9
17	5	6,0									45	40		21	

Datová tabulka a její možné problémy

Jednoznačné ID nezbytné pro identifikaci a případné propojení do dokumentace.

Sloupec nesmí obsahovat kombinaci textu a čísel.

Chybně uvedeno datum.

ID	Pohlaví	Věk	Výška	Zařazen	Alergie	TKD/TKS
9	M	53	177	13.9.2001	N	80/120
14	M	41	167	10.9.2001	N	75/119
19	M	52	182	14.90.2001	N	91/145
22	M	26	193	17.9.2001	A	78/130
23	MM	53	neznámo	17.9.2001	N	80/120
29	M	23	197	4.10.2001	0	75/119
30	M	58	158	4.10.2001	N	91/145
32	Z	198	45	5.10.2001	N	78/130
33	Z	51	191	5.10.2001	1	80/120
34	M	44	169	5.10.2001	1	75/119
35	Z	22	0	5.10.2001	N	91/145
38	M	42	163	5.10.2001	A	78/130

Překlep v názvu kategorie, při zpracování dat se chová jako nová kategorie.

Je třeba uvádět v samostatných sloupcích pro diastolický a systolický tlak.

Nereálné odlehlé hodnoty, pravděpodobně prohozen věk a výška.

Uvedena 0 zřejmě namísto chybějící hodnoty, je třeba ponechat prázdnou buňku.

Kombinace dvou možných kategorizací (0/1 nebo N/A), je třeba si vybrat jednu z nich.

Zásady pro ukládání dat

- Správné a přehledné uložení dat je základem jejich pozdější analýzy
- Je vhodné rozmyslet si předem jak budou data ukládána
- Pro počítačové zpracování dat je nezbytné ukládat data v tabulární formě
- Nejvhodnějším způsobem je uložení dat ve formě databázové tabulky
  - Každý sloupec obsahuje pouze jediný typ dat, identifikovaný hlavičkou sloupce
  - Každý řádek obsahuje minimální jednotku dat (např. pacient, jedna návštěva pacienta apod.)
  - Je nepřípustné kombinovat v jednom sloupci číselné a textové hodnoty
  - Komentáře jsou uloženy v samostatných sloupcích
  - U textových dat nezbytné kontrolovat překlepy v názvech kategorií
  - Specifickým typem dat jsou datumy u nichž je nezbytné kontrolovat, zda jsou datumy uloženy v korektním formátu

- Takto uspořádaná data je v tabulkových nebo databázových programech možné převést na libovolnou výstupní tabulku
- Pro základní uložení a čištění dat menšího rozsahu je možné využít aplikací MS Office

Vizualizace dat

Typy grafické vizualizace

Rizika desinterpretace grafického zobrazení dat

Anotace

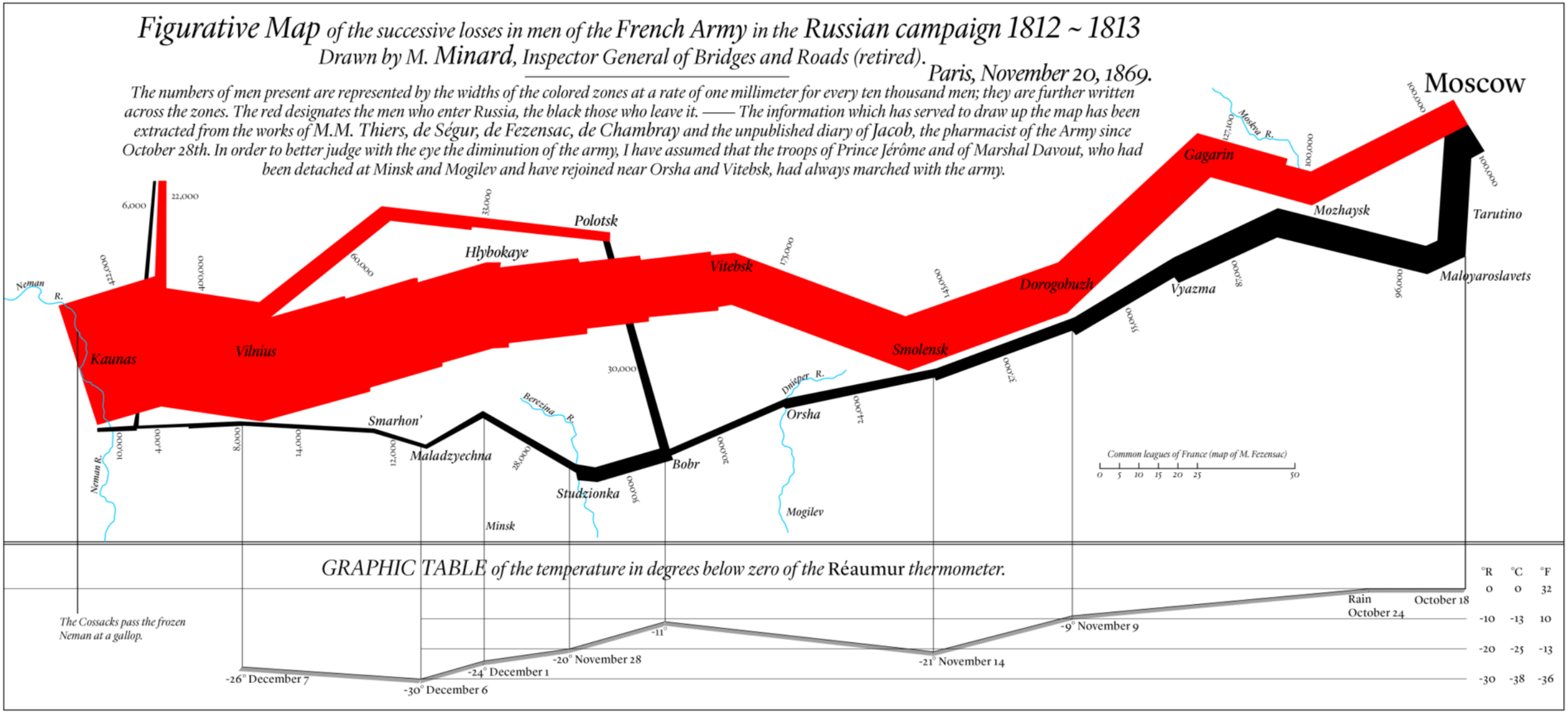
- Prvním krokem v analýze dat je jejich vizualizace.
- Různé typy dat nám umožňující získání představy o rozložení dat, zastoupení kategorií i vztazích proměnných navzájem.
- Prostřednictvím vizualizace získáváme vhled do dat a začínáme vytvářet hypotézy o zákonitostech panujících mezi proměnnými v hodnoceném souboru dat.

V čem vytvářet grafy

- Nejrůznější software – nejrůznější možnosti
  - MS Office – základní grafy, snadná editovatelnost, lze invenčně upravit, snadná replikovatelnost výměnou dat
  - R – různé knihovny (např. ggplot) – vyšší vstupní investice, nejrůznější typy grafů, automatizace
  - SPSS, Statistica – rychlá tvorba velkého množství grafů, mnoho typů grafů

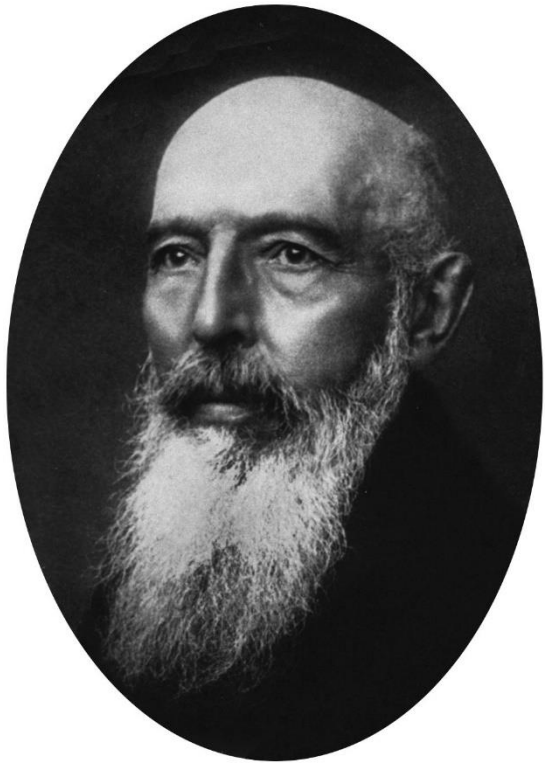
- Kritéria
  - Výběr různých typů grafů
  - Snadnost editace a úpravy vzhledu
  - Snadná replikovatelnost/automatizace/rychlost tvorby grafů

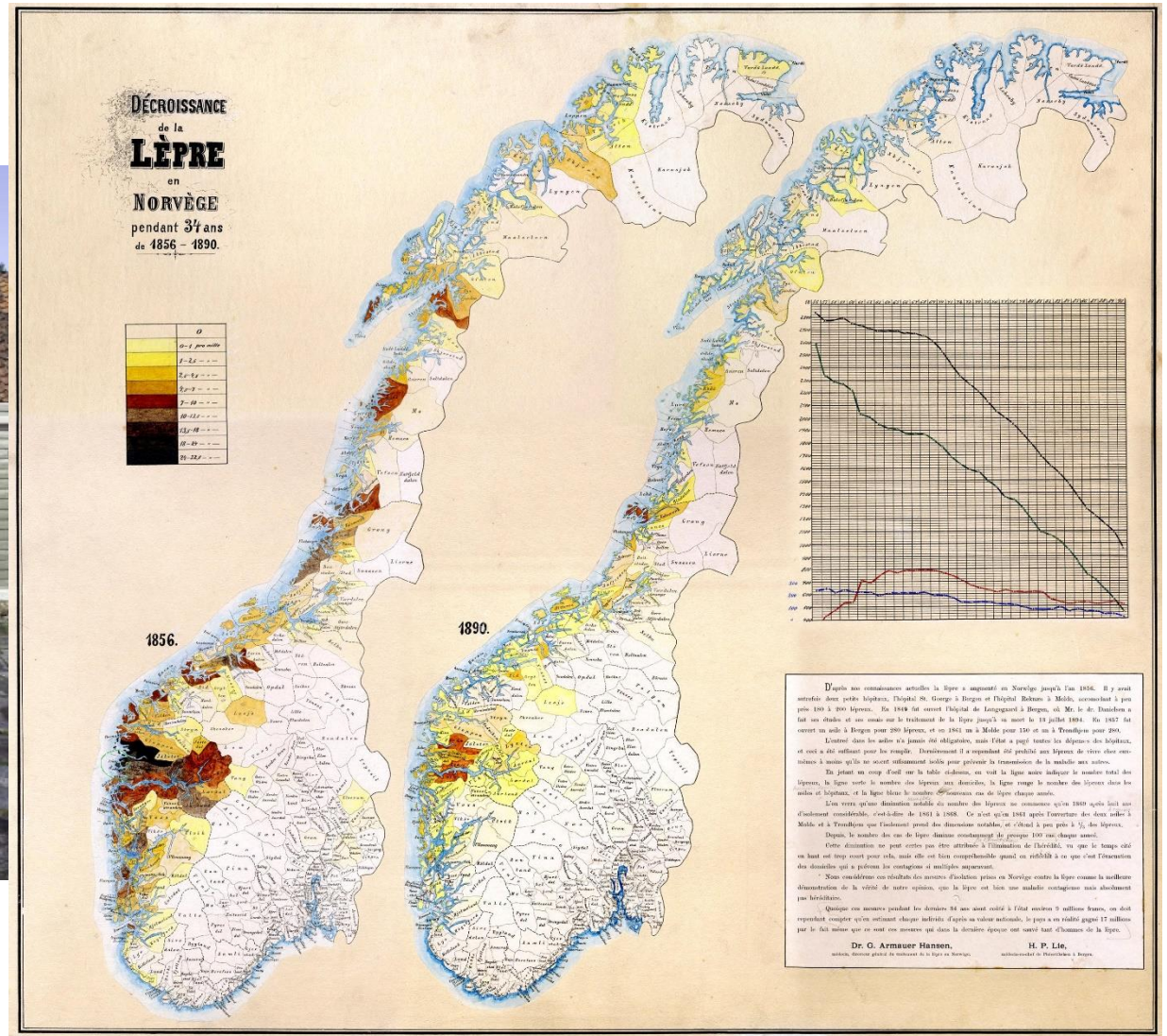
Slavné grafy: Charles Joseph Minard – Napoleonovo tažení do Ruska



Slavné grafy: Eradikace lepry v Norsku

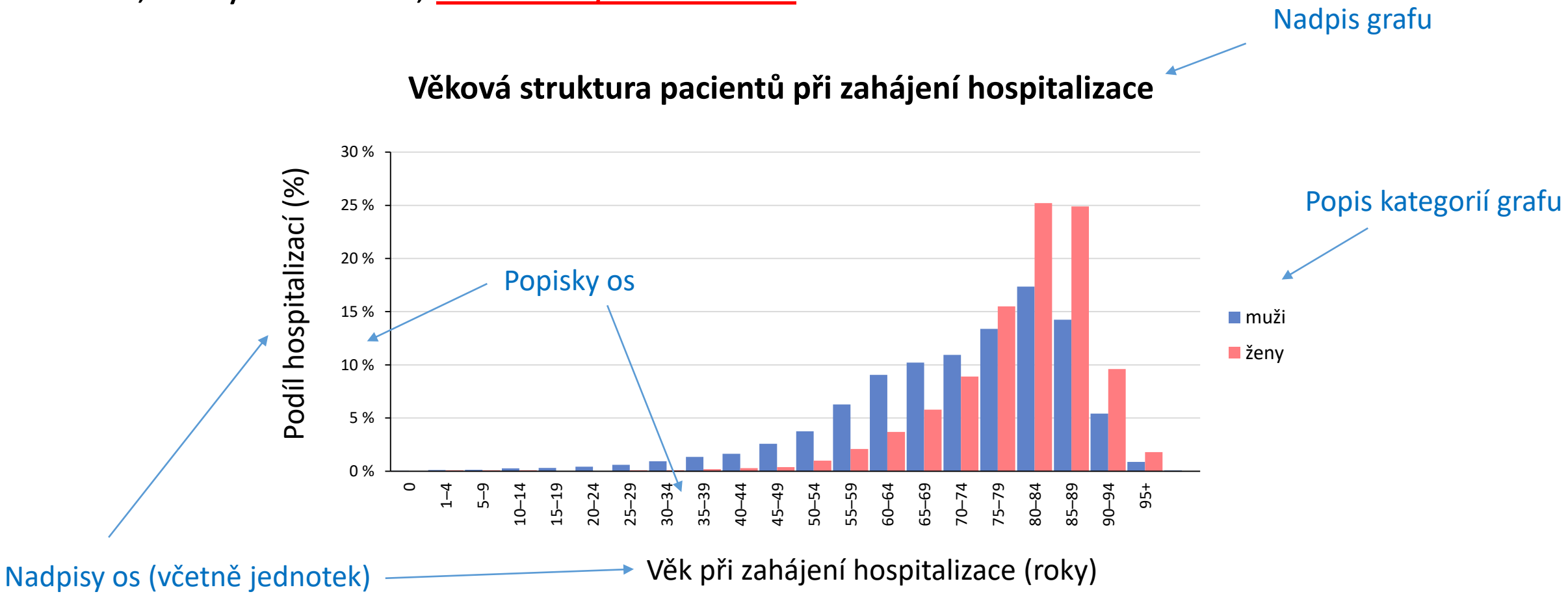
- 1856 – národní registr lepry v Norsku založen v Bergenu -> analýza získaných dat -> opatření k eradikaci lepry v Norsku
- Gerhard Armauer Hansen





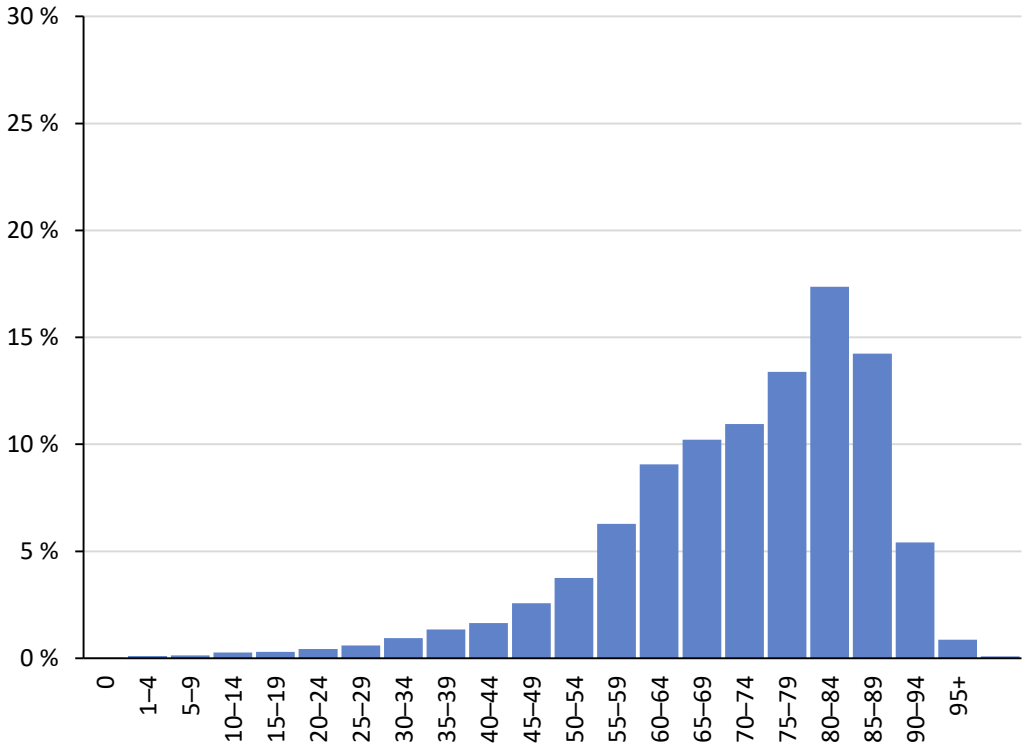
Co nesmí chybět na grafu

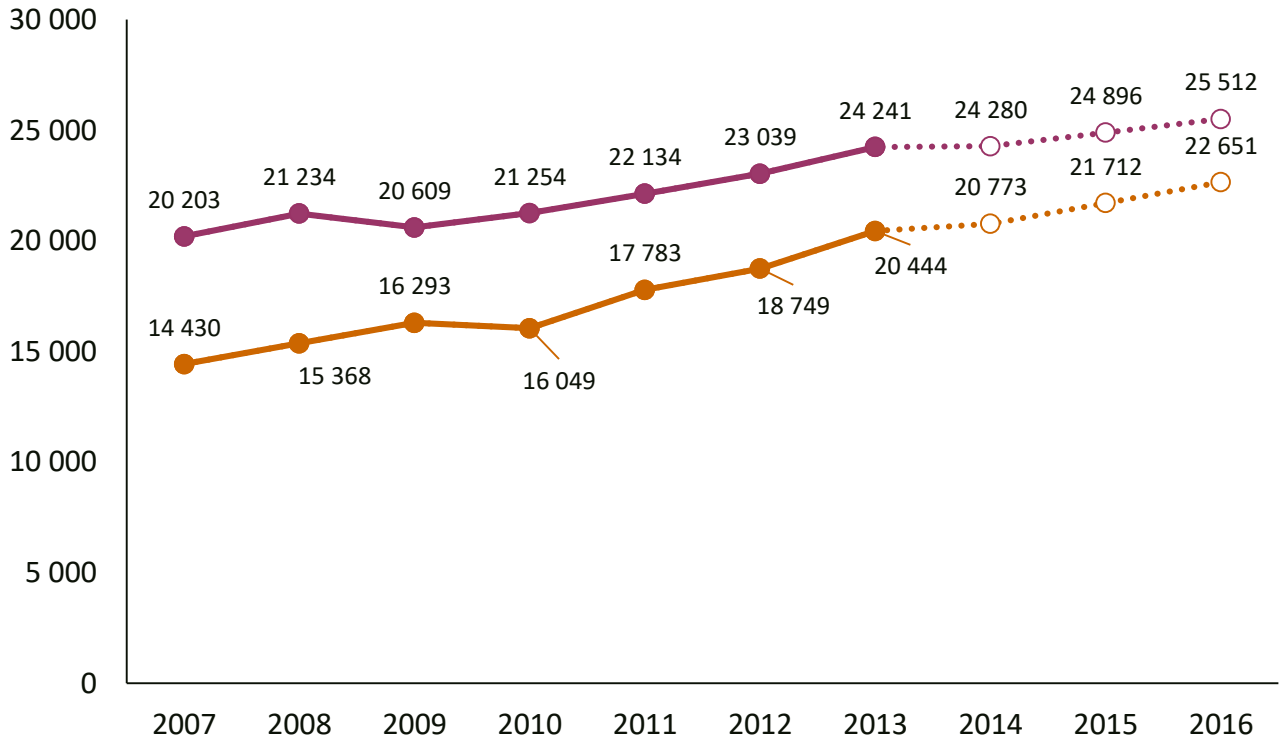
- Každý graf musí být jednoznačně popsán – self explained
- Graf, který nic neříká, nemá smysl kreslit !!!



Sloupcové a čárové grafy

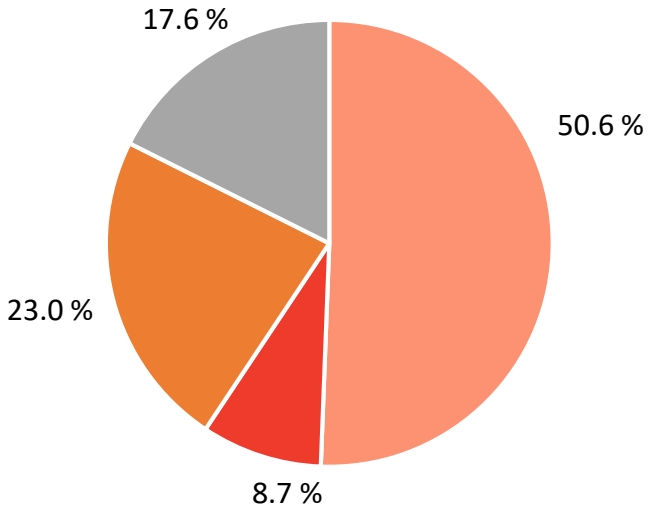
- Jednoduchá tvorba, vizualizace absolutních hodnot nebo procent

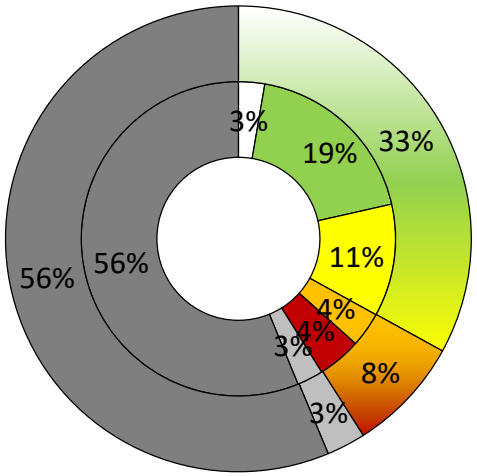


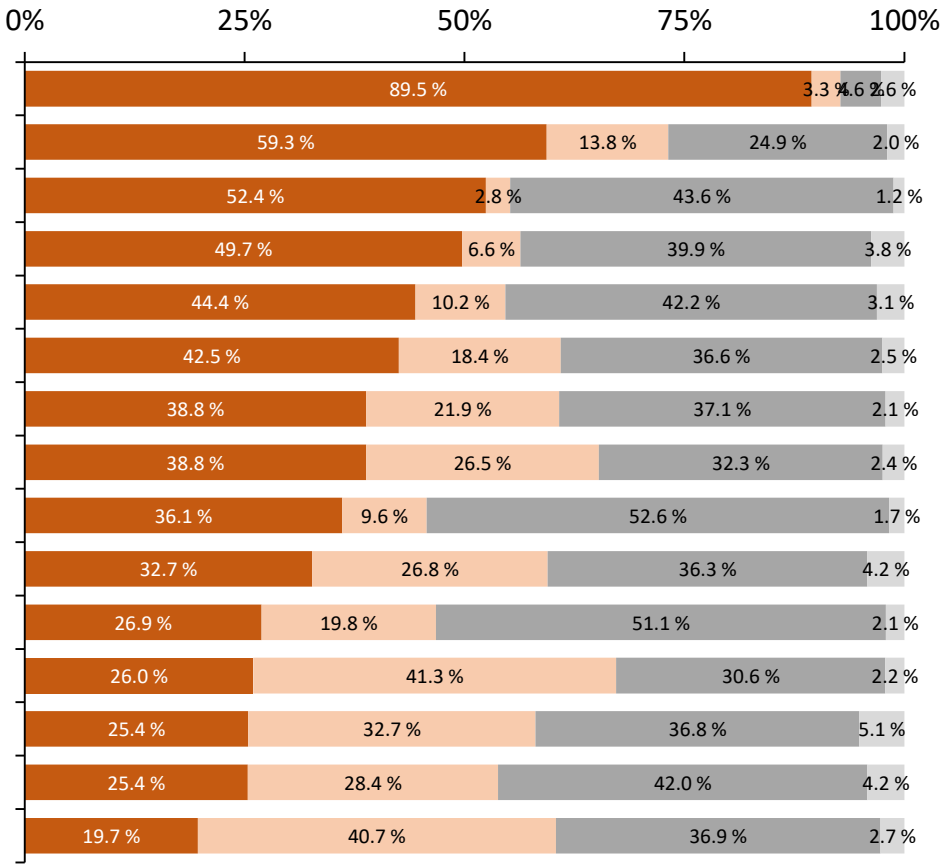


Koláčové a páskové grafy

- Jednoduchá tvorba, vizualizace procent

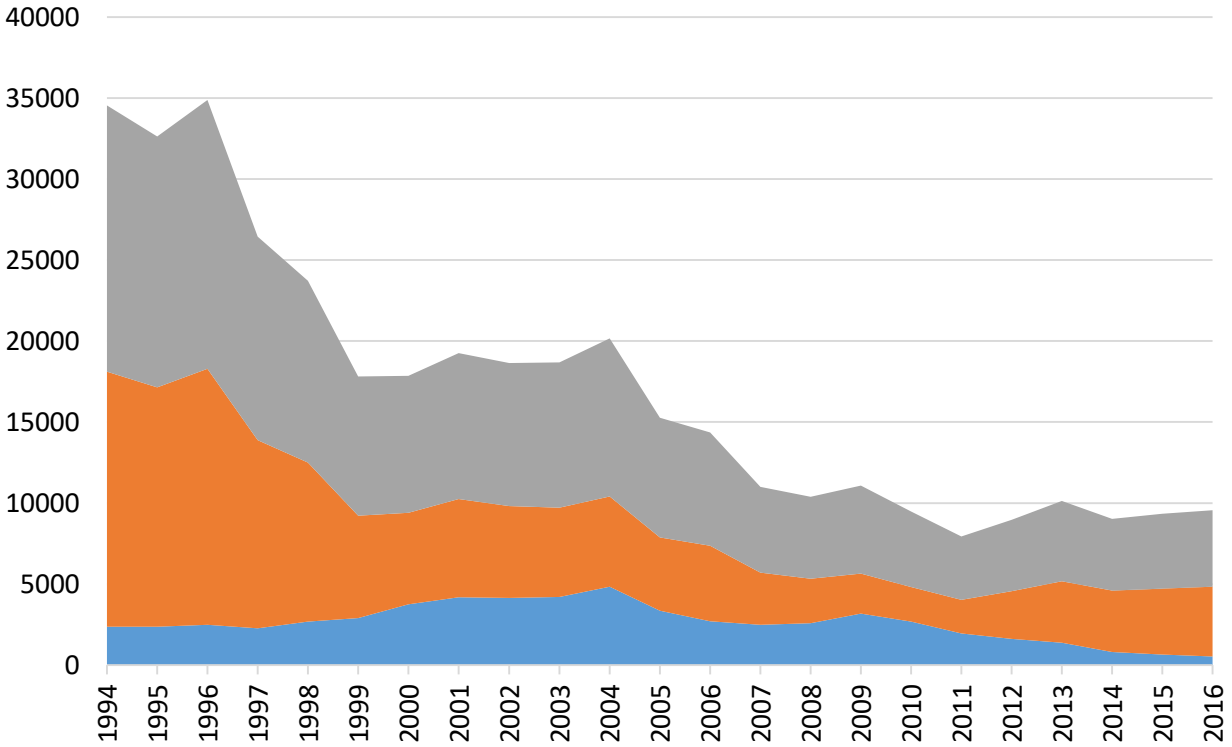


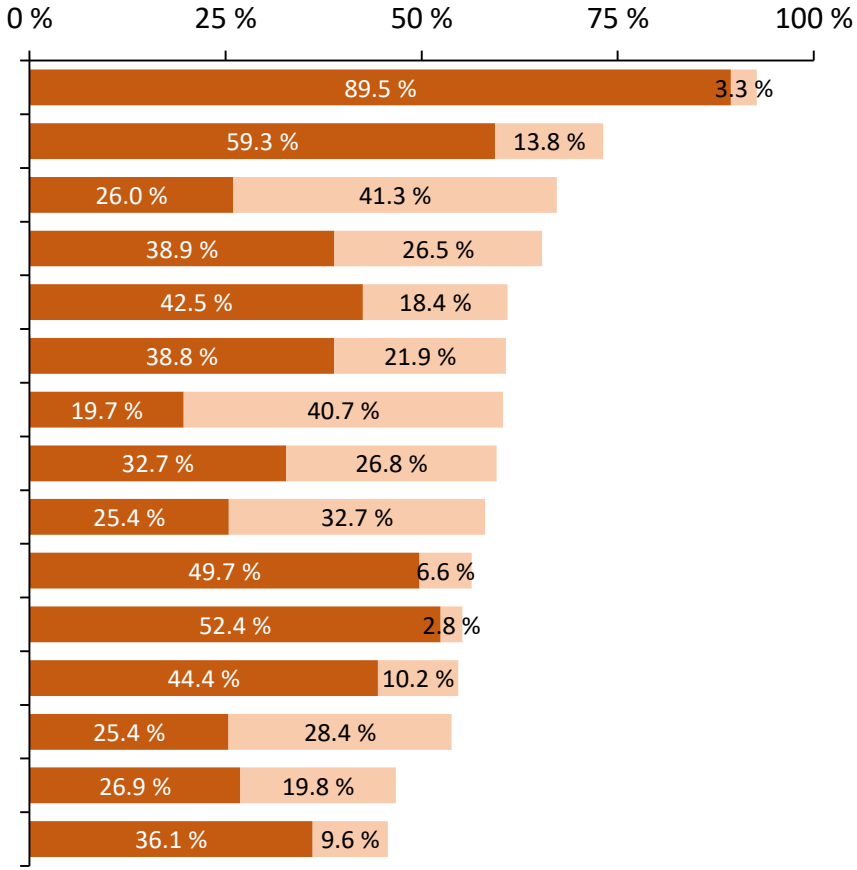




Skládané grafy

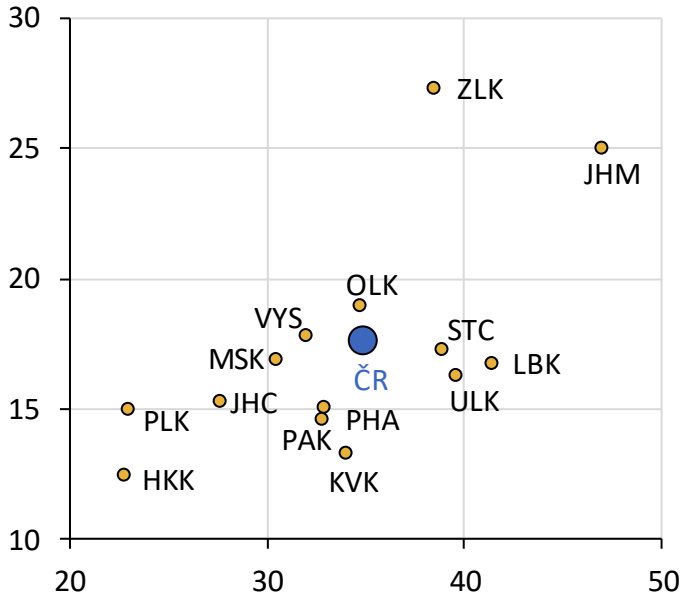
- Kumulativní zobrazení více informací

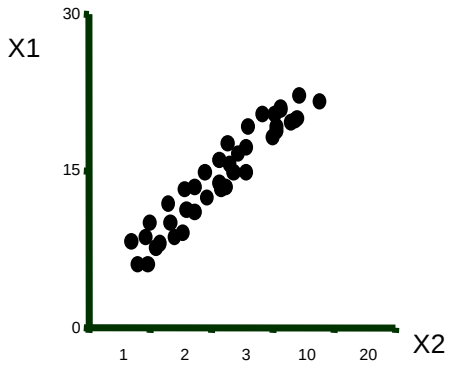


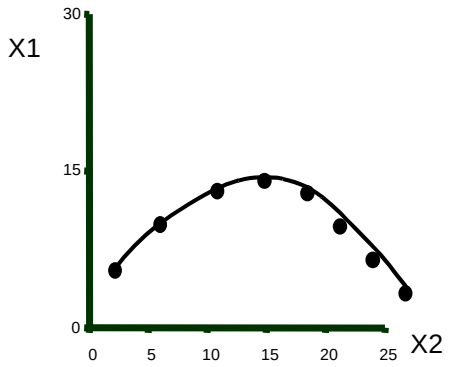


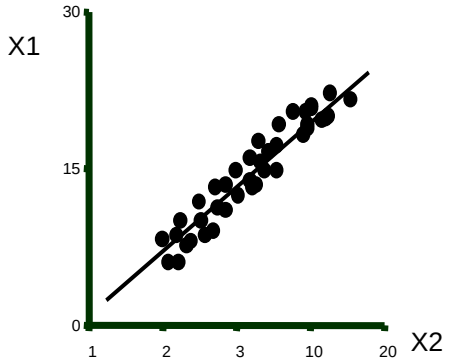
XY graf (scatter plot)

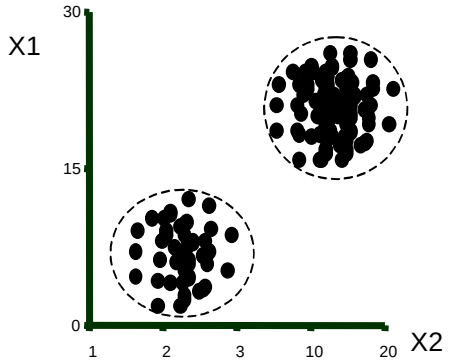
- Popis vztahu dvou spojitých proměnných
- Možnost kategorizace a popisu bodů
- Prokládání modelů do grafů
- Základní graf pro prohlídku dat před korelační a regresní analýzou





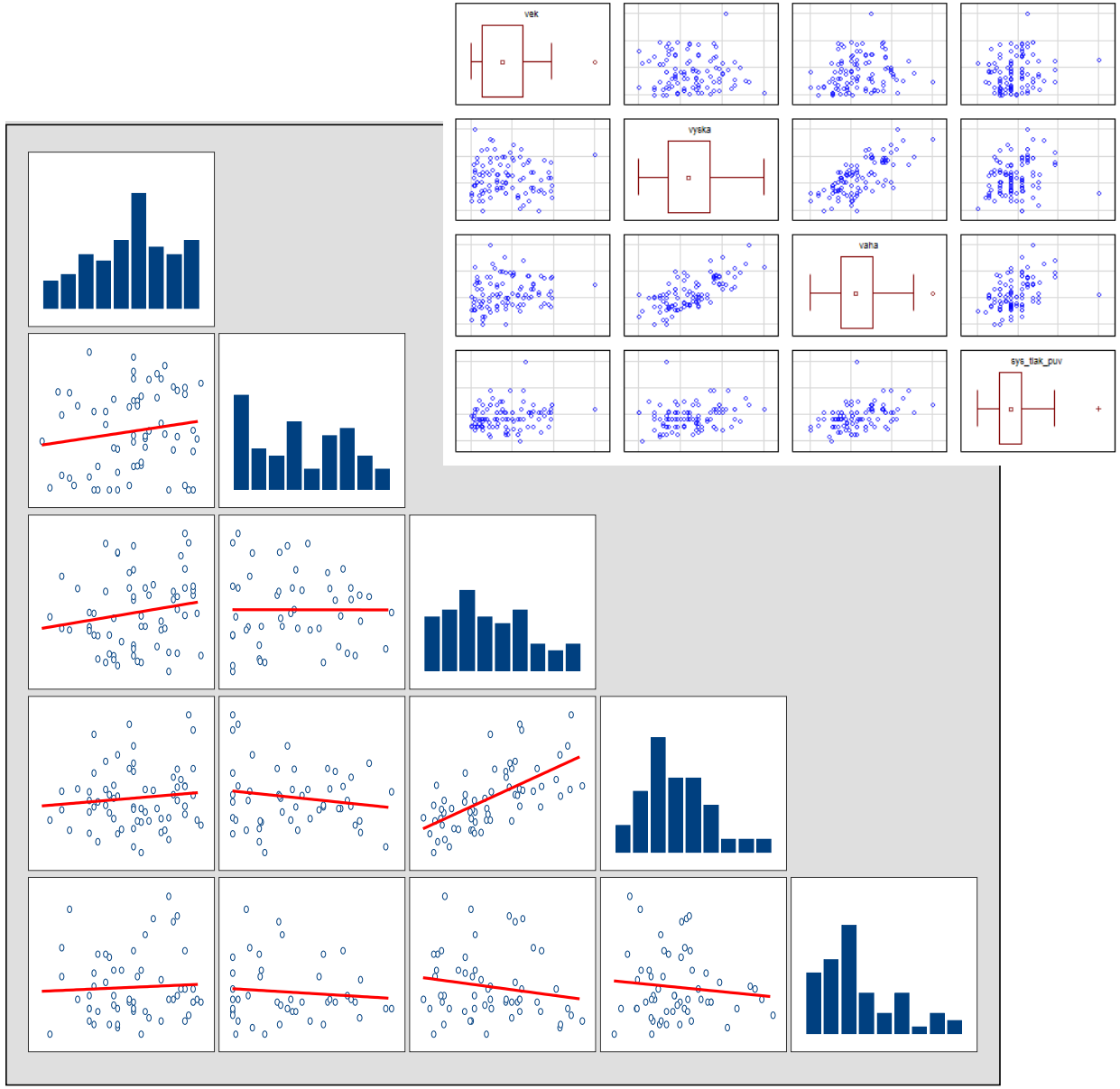






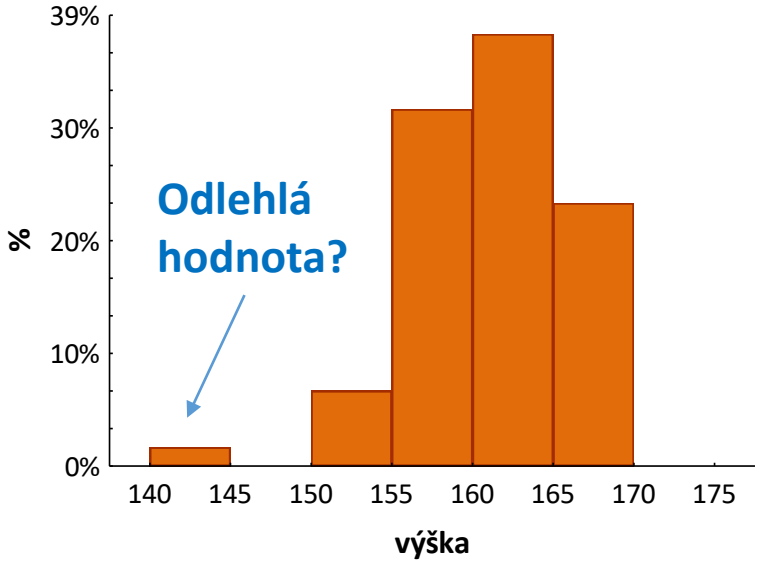
Maticový graf

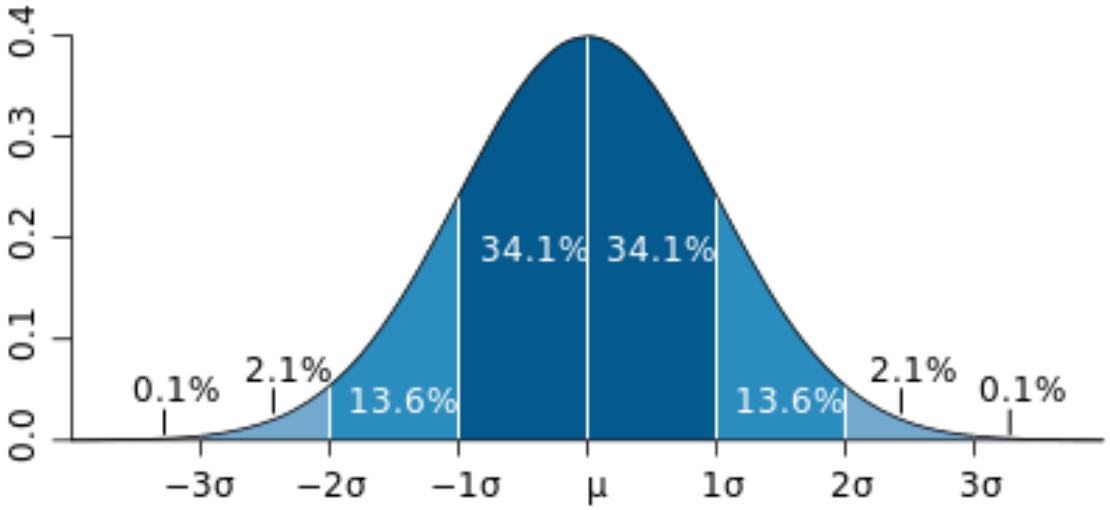
- Rozšíření xy grafů ve statistických SW
- Současná vizualizace rozložení hodnot (diagonála) a vzájemných vztahů většího počtu spojitých proměnných
- Různé varianty
  - Sada proměnných každý s každým
  - Dvě sady proměnných proti sobě
  - Doplnění o výpočet korelačních koeficientů
- Základní nástroj vizualizace před vícerozměrnou analýzou



Histogram

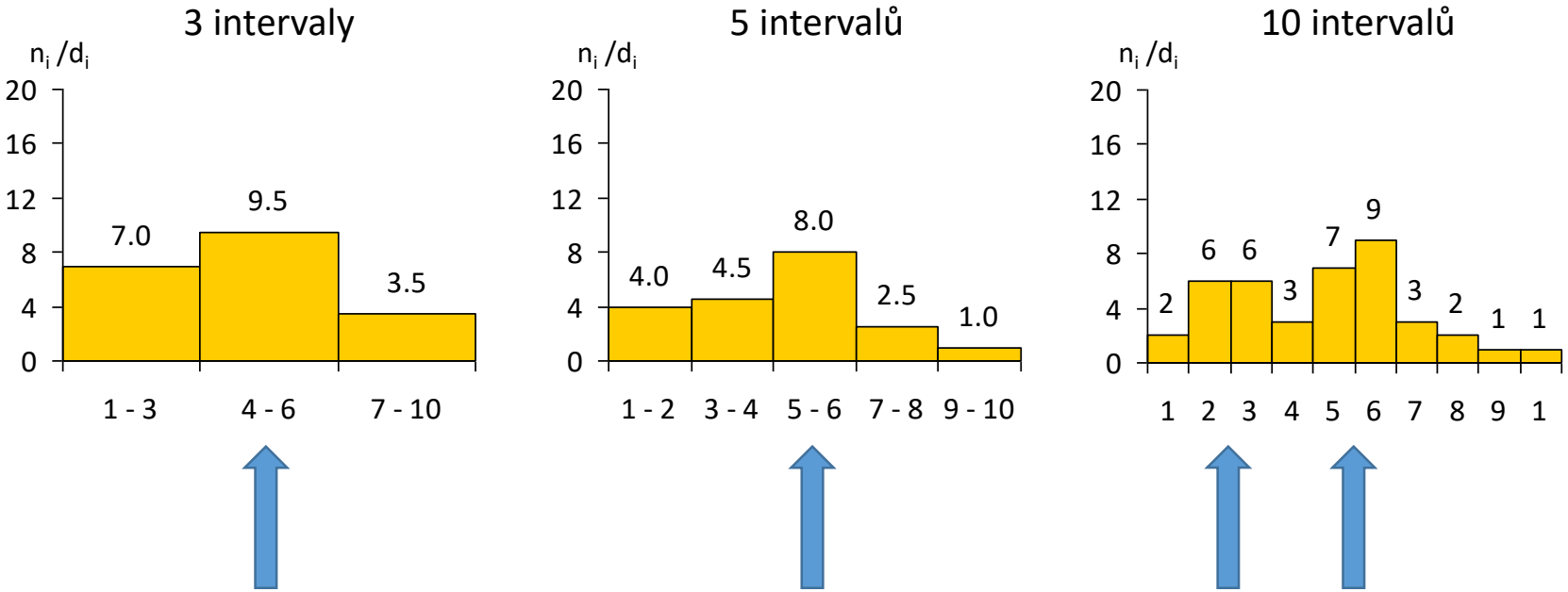
- Graf sumarizující rozložení hodnot spojitých proměnných, úzce spjat s teorií statistických rozdělení
- V klasické formě podobný (ale nikoliv totožný) se sloupcovým grafem
- V praxi se pod názvem histogram často skrývá sloupcový graf (přípustné pokud nevede k dezinterpretaci dat)
- Jeden ze základních grafů pro posouzení rozložení dat





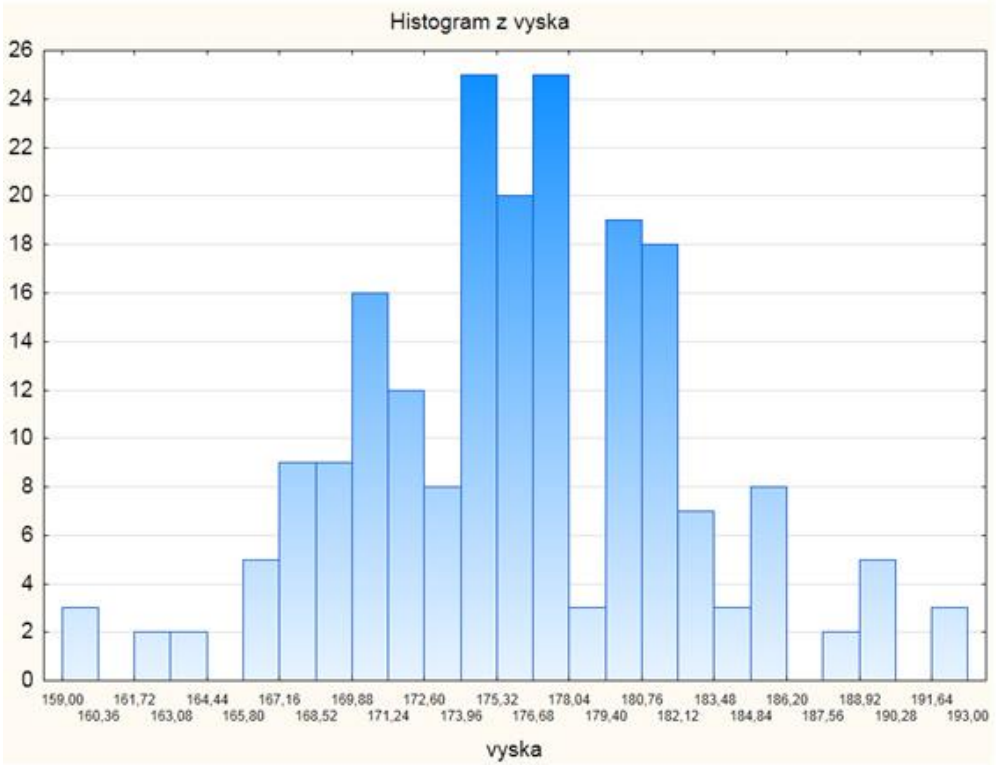
Histogram: vliv kategorizace dat

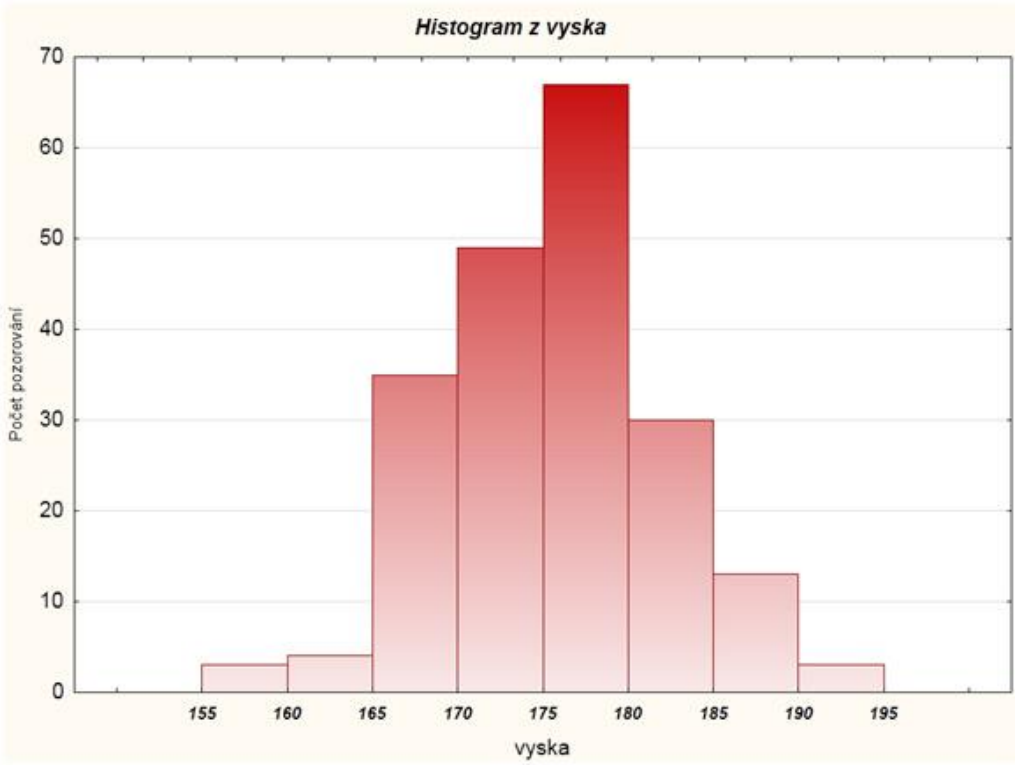
- Počtem zvolených intervalů v histogramu rozhodujeme o tom, jak bude vypadat. Při malém počtu můžeme přehlédnout důležité prvky v datech, při velkém zase může být informace roztříštěná.



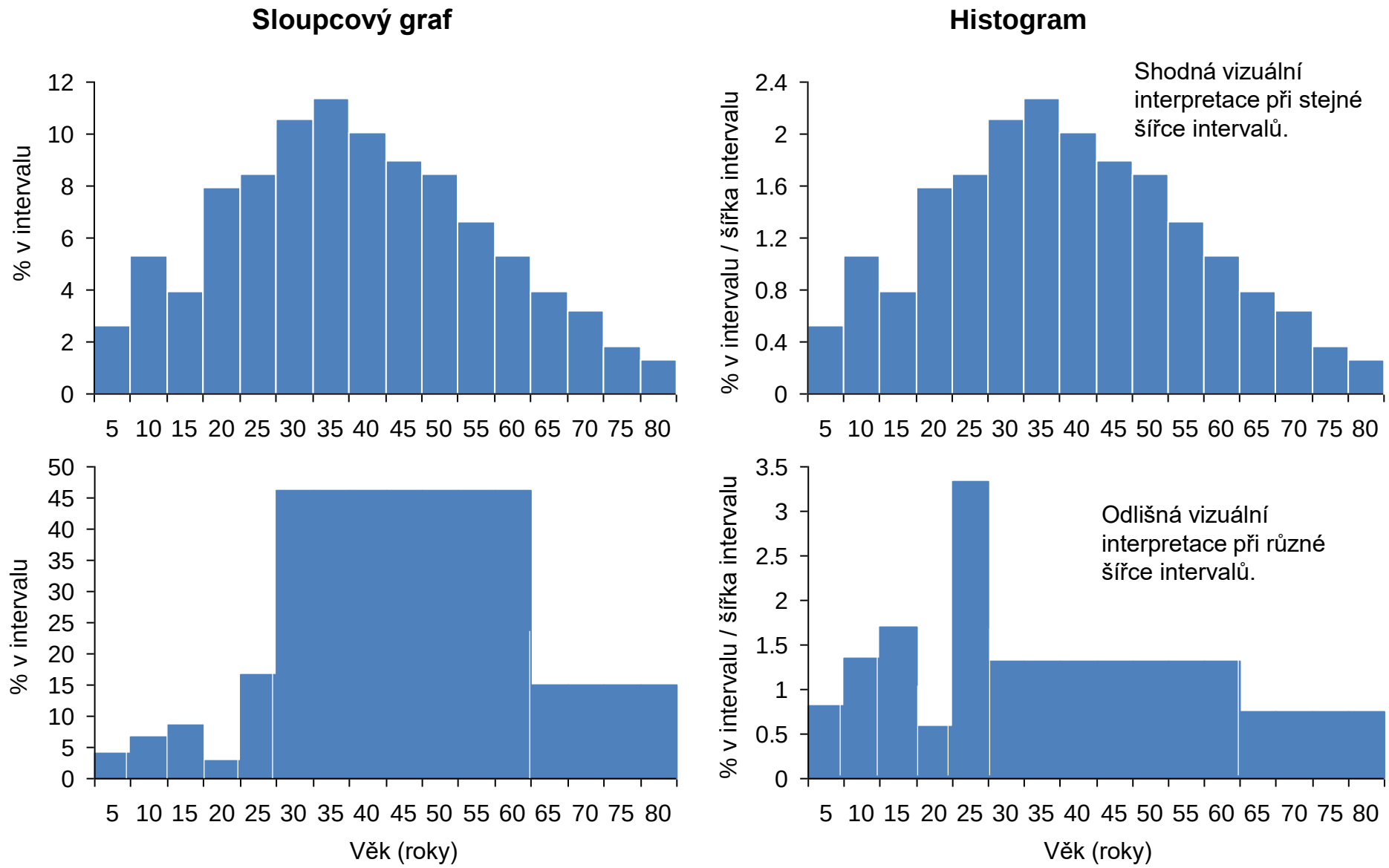
Histogram: vliv kategorizace dat

- Výběr počtu kategorií – důležitý pro interpretaci
- Ruční nebo automatický výběr – různé algoritmy (závisí na velikosti vzorku a variabilitě dat)



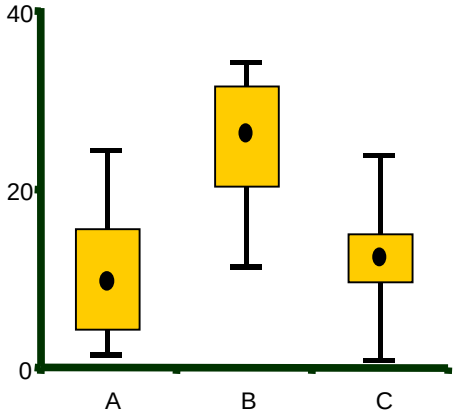


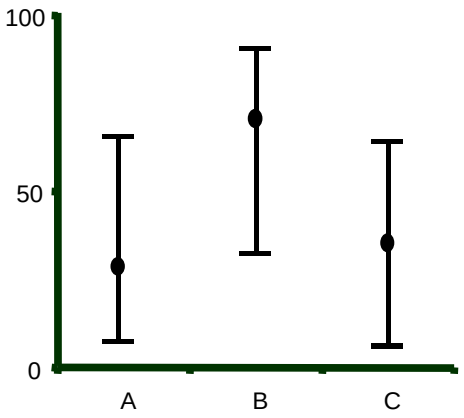
Histogram a sloupcový graf

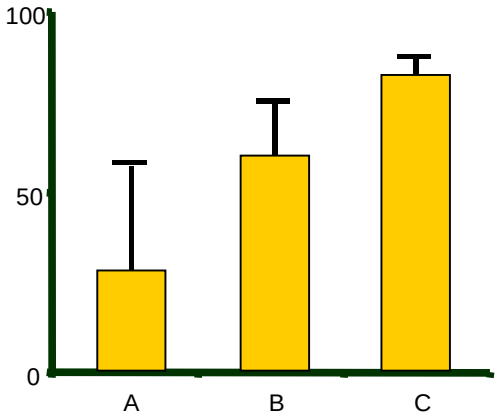


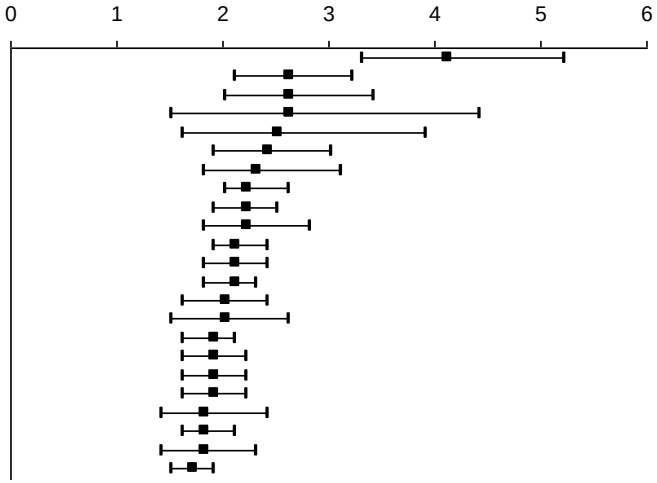
Krabicový graf – box and whisker plot: co to je?

- V analýze dat oblíbený typ grafu umožňující jednoduché srovnání více skupin objektů a hodnocení rozložení dat
- Nejběžnější pro popis spojitých dat, ale využitelný pro libovolné typy dat, které lze popsat střední hodnotou a variabilitou (procenta, regresní koeficienty, odds ratia, risk ratia, hazard ratia atd.)
- Obrovské množství variant

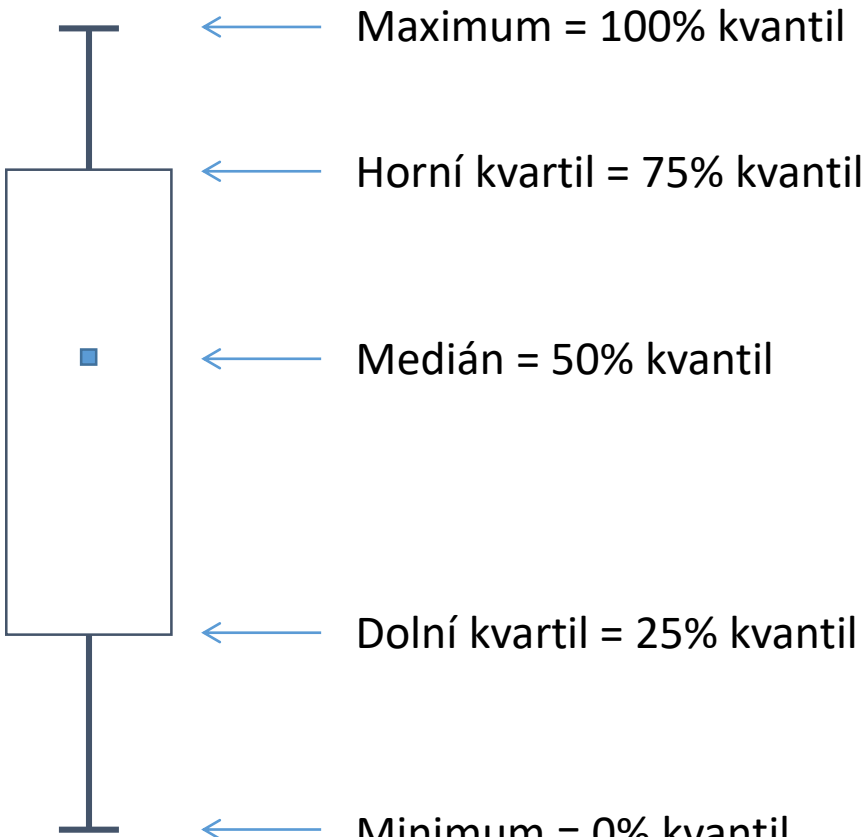








Krabicový graf – box and whisker plot: příklad jedné možné varianty

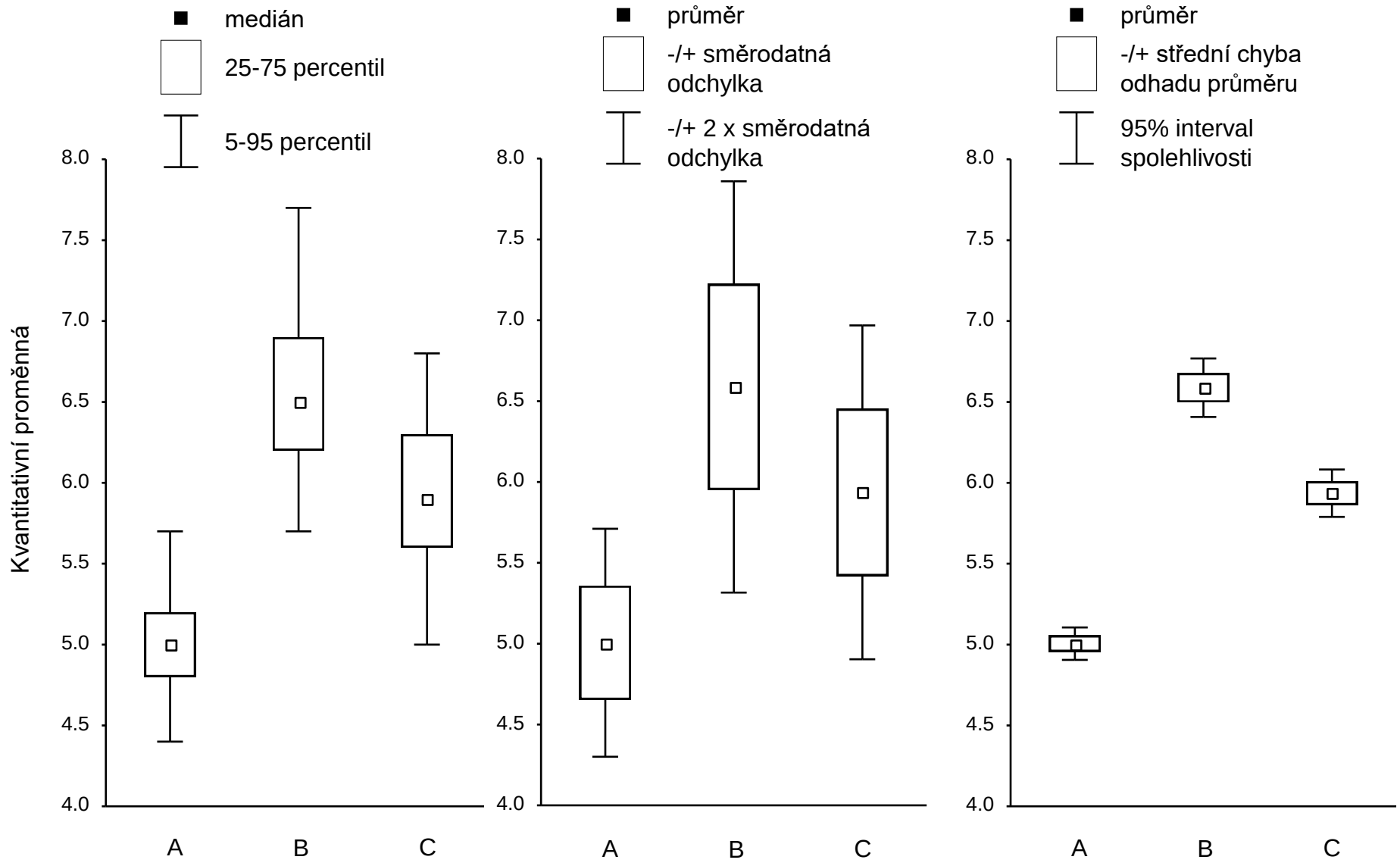


Jednotlivé body grafů mohou obsahovat libovolné popisné statistiky – průměry, směrodatné odchylky, intervaly spolehlivosti, odds ratia, hazard ratia atd.

Počet datových bodů v grafu může být od tří do např. devíti.

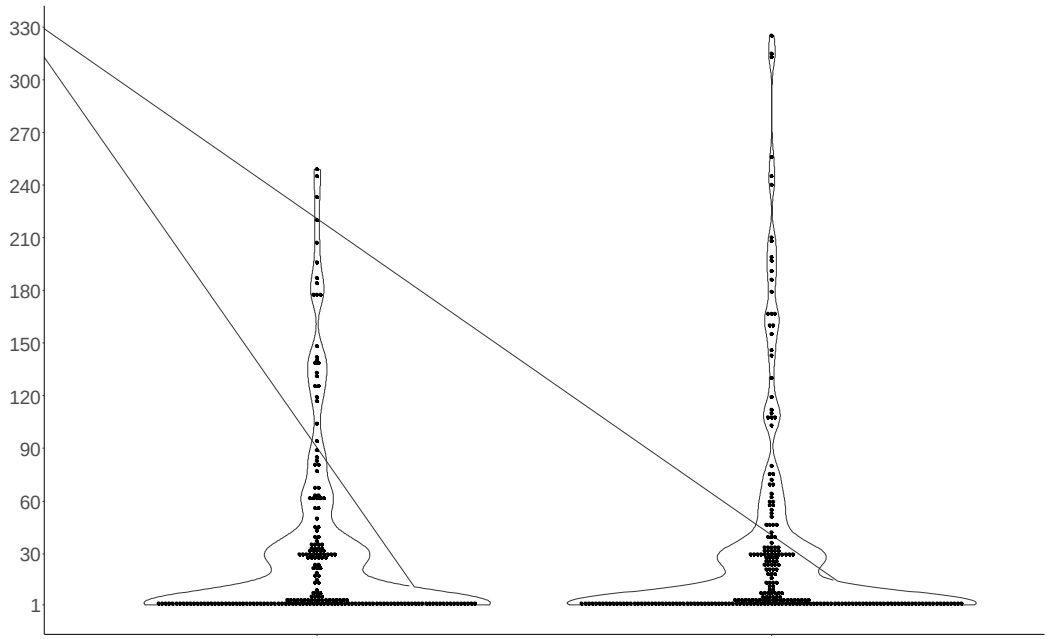
Box and whisker plot a jeho různé varianty I

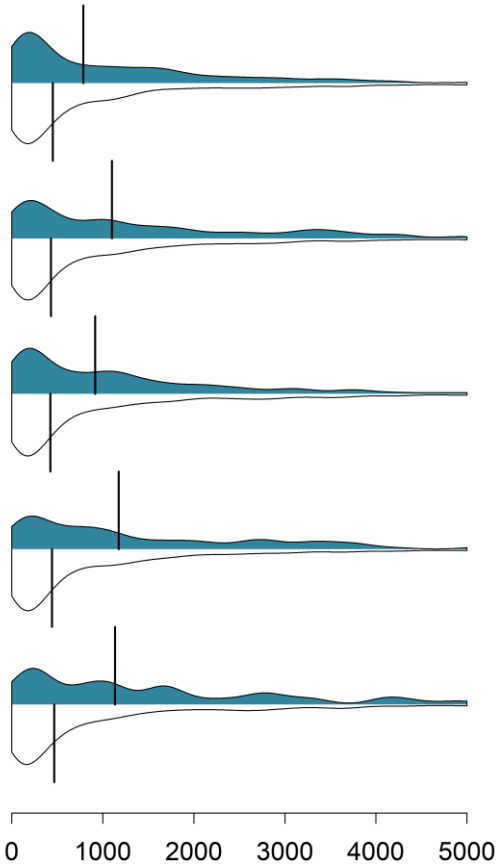
- Je nezbytné číst popisky
- Různé varianty grafu mohou mít zcela jinou interpretaci

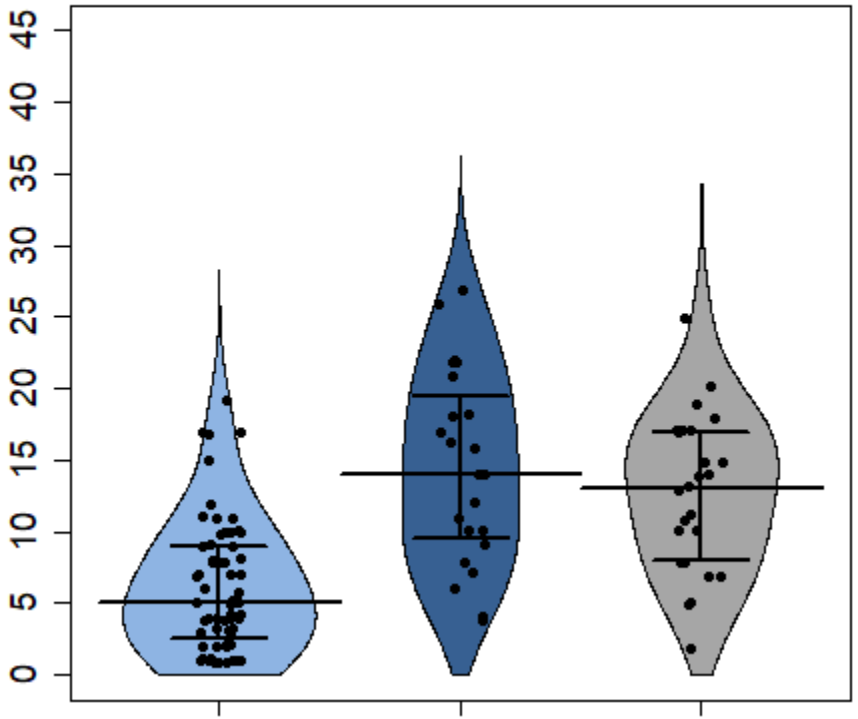


Box and whisker graf a jeho různé varianty II: Violin plot a Beanplot

- Kombinace histogramu a box plotu nebo tečkového grafu
- K dispozici v R – např. knihovny beanplot a ggplot2

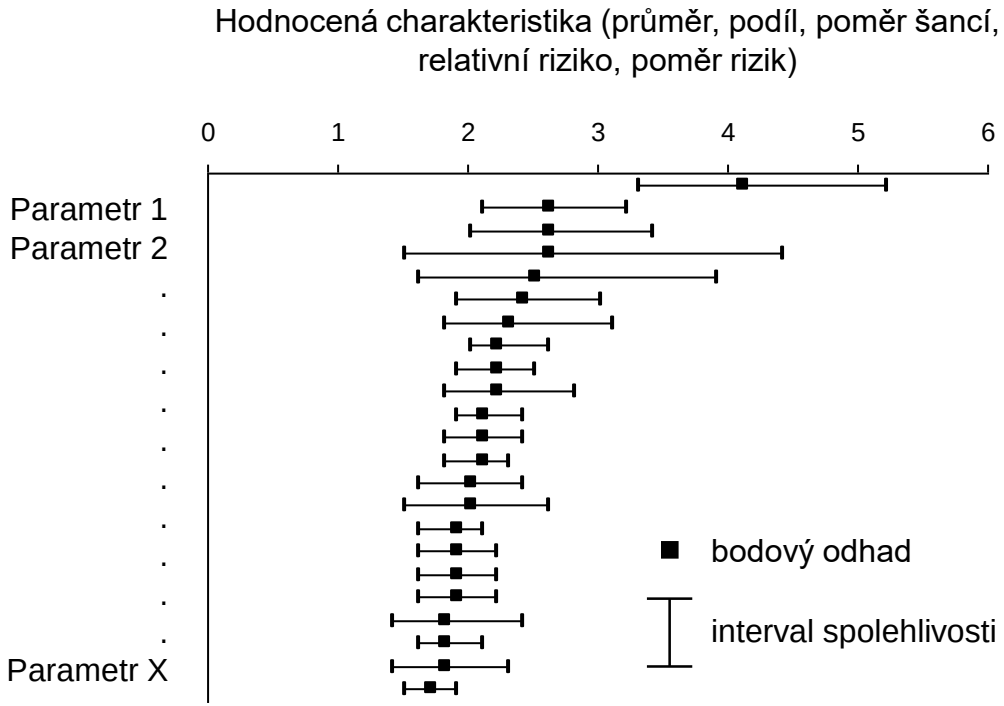




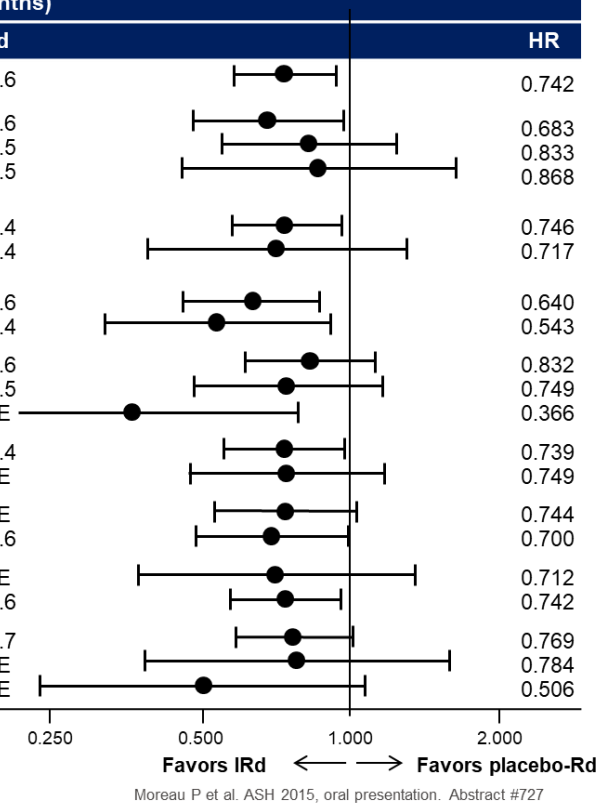


Box and whisker graf a jeho různé varianty III: Forest plot

- Varianta box and whisker plotu
- Často používaná pro zobrazení regresních koeficientů nebo odds/risk/hazard ratií



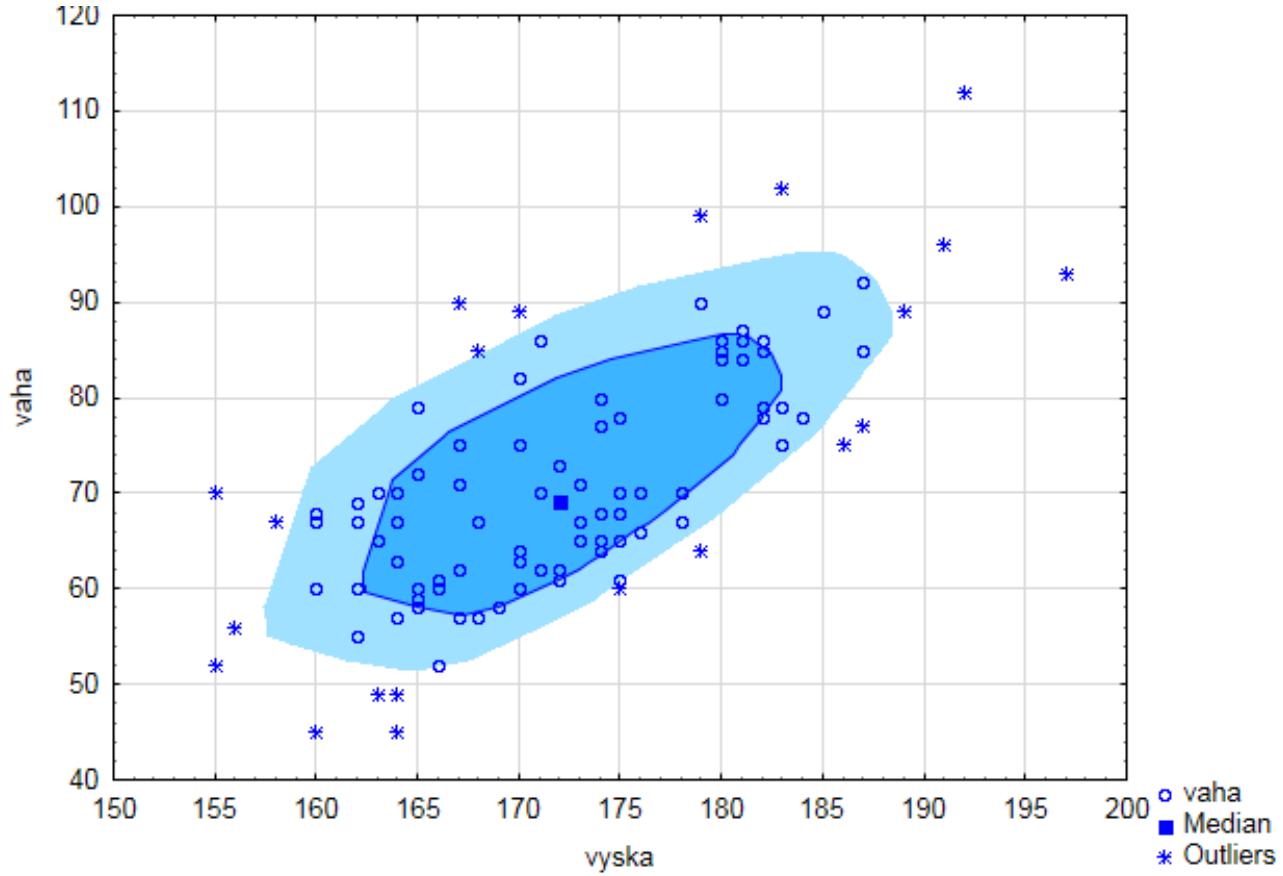
		N		Median PFS (months)		
Variable	Subgroup	Placebo-Rd	IRd	Placebo-Rd	IRd	HR
All patients	ALL	362	360	14.7	20.6	0.742
Age (yrs)	≤65	176	168	14.1	20.6	0.683
	>65-75	125	145	17.6	17.5	0.833
	>75	61	47	13.1	18.5	0.868
ISS stage (stratification factor)	I or II	318	314	15.7	21.4	0.746
	III	44	46	10.1	18.4	0.717
Cytogenetic risk	Standard-risk	216	199	15.6	20.6	0.640
	High-risk	62	75	9.7	21.4	0.543
Number of prior therapies	1	217	224	15.9	20.6	0.832
	2	111	97	14.1	17.5	0.749
	3	34	39	10.2	NE	0.366
Proteasome inhibitor	Exposed	253	250	13.6	18.4	0.739
	Naive	109	110	15.7	NE	0.749
Prior IMiD therapy	Exposed	204	193	17.5	NE	0.744
	Naïve	158	167	13.6	20.6	0.700
Refractory to last prior therapy	Yes	55	59	NE	NE	0.712
	No	307	301	14.1	20.6	0.742
Relapsed or refractory	Relapsed	280	276	15.6	18.7	0.769
	Refractory	40	42	13.0	NE	0.784
	Ref & rel	42	41	13.1	NE	0.506



Moreau P et al. ASH 2015, oral presentation. Abstract #727

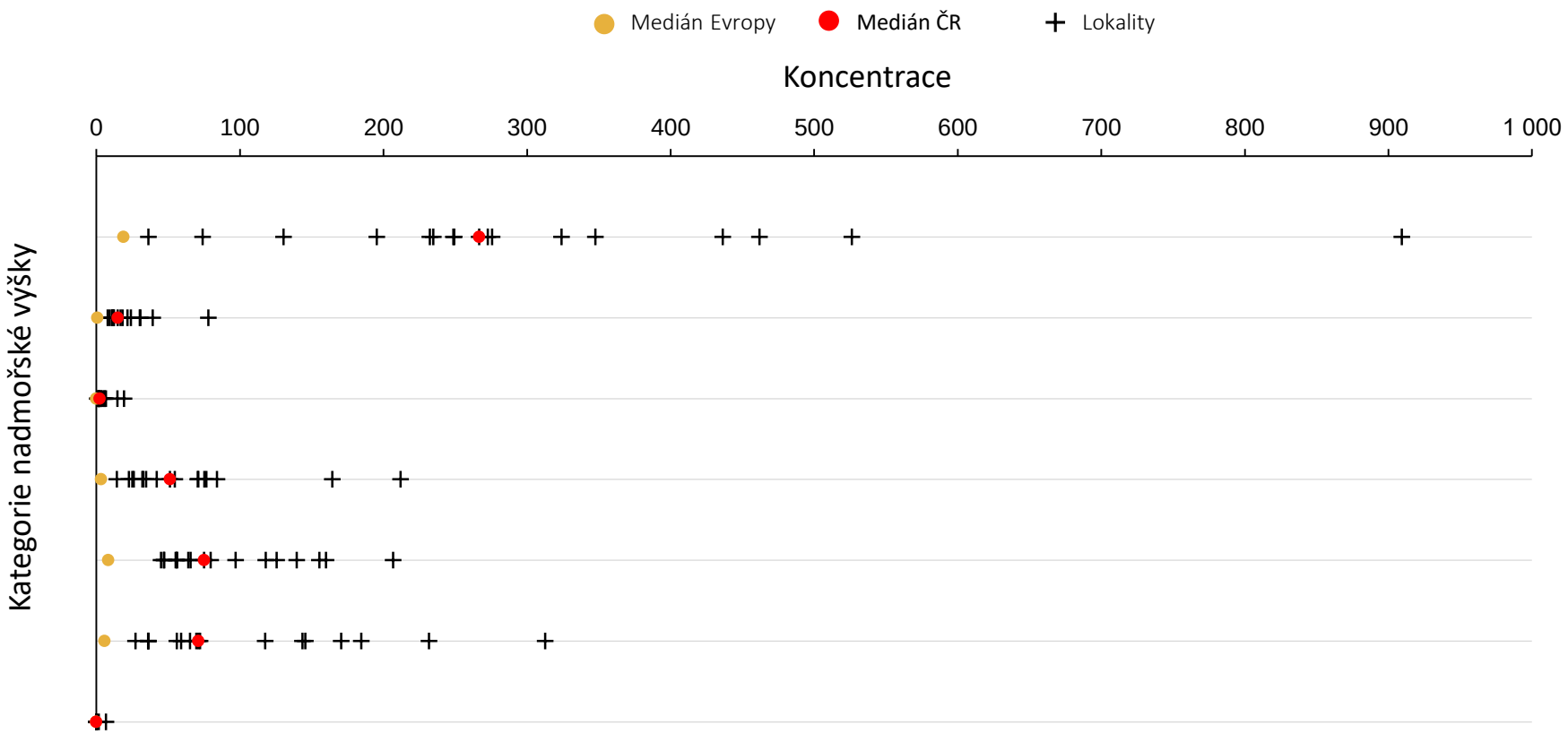
Box and whisker graf a jeho různé varianty IV: Bagplot

- Bagplot = „bivariate boxplot“ (tzn. „dvourozměrný krabicový graf“)



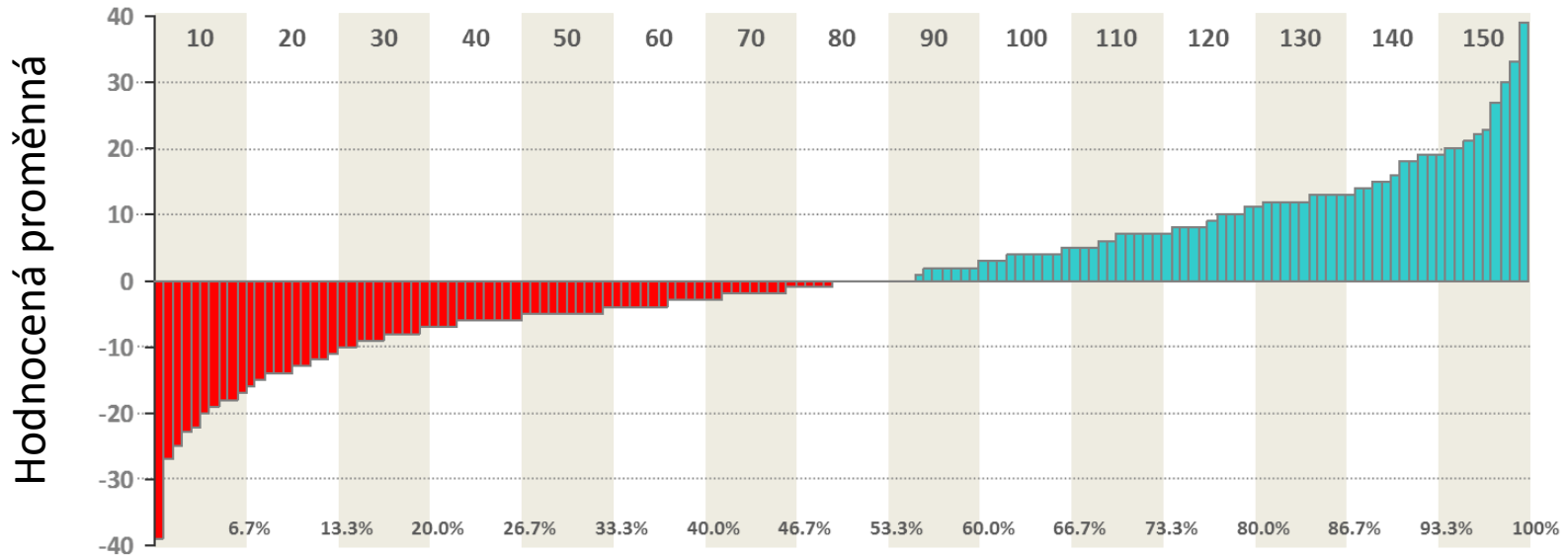
Invenční využití jednoduchých grafů: Korálkový graf

- Lze vytvořit z XY grafu v MS Office
- Velké množství informace na malé ploše



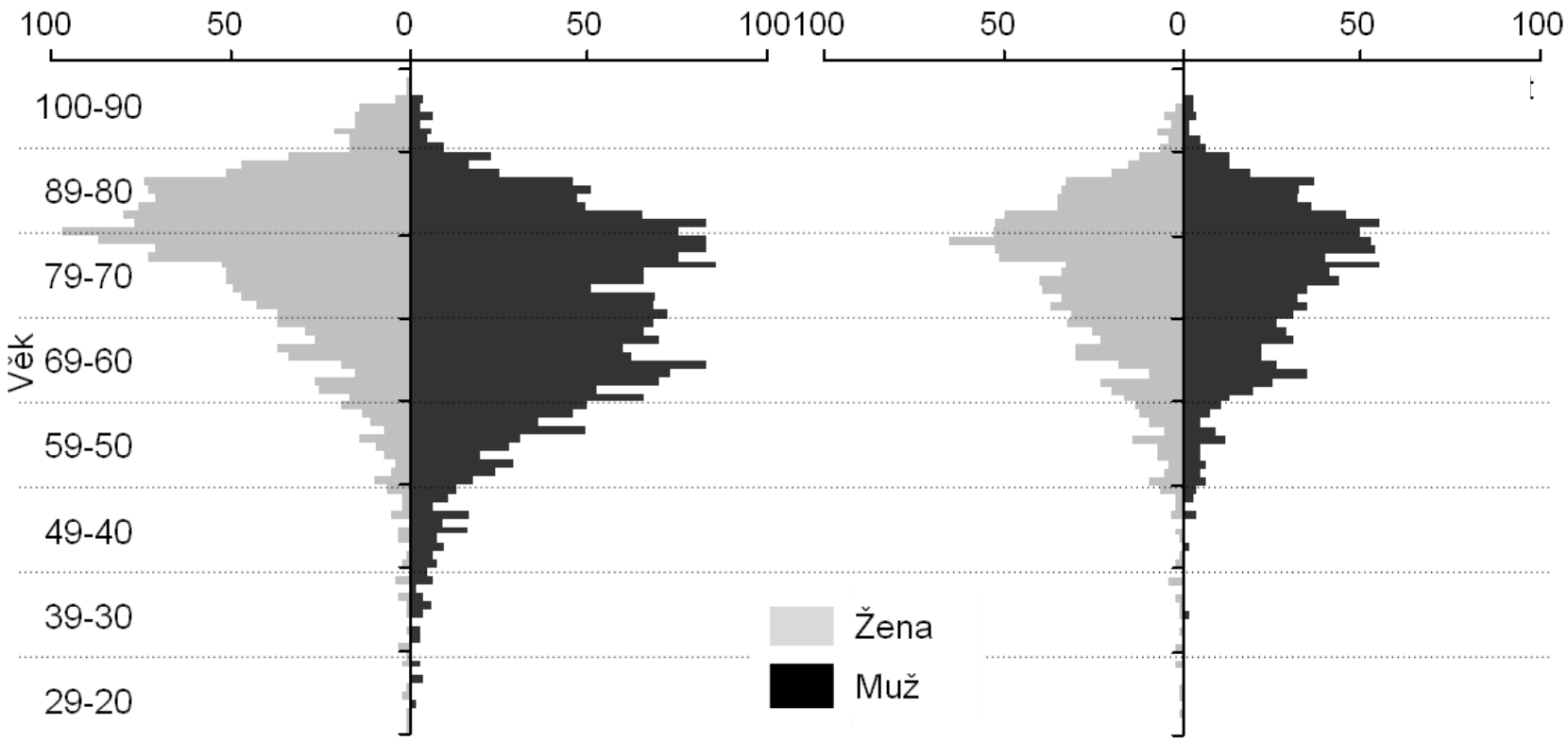
Invenční využití jednoduchých grafů: Waterfall plot

- Vizualizace výsledků individuálních objektů, často u proměnných popisujících změny
- Hodnoty jsou v grafu seřazeny dle velikosti
- Může být doplněn o hodnoty norem, procenta objektů v kategoriích normy apod.



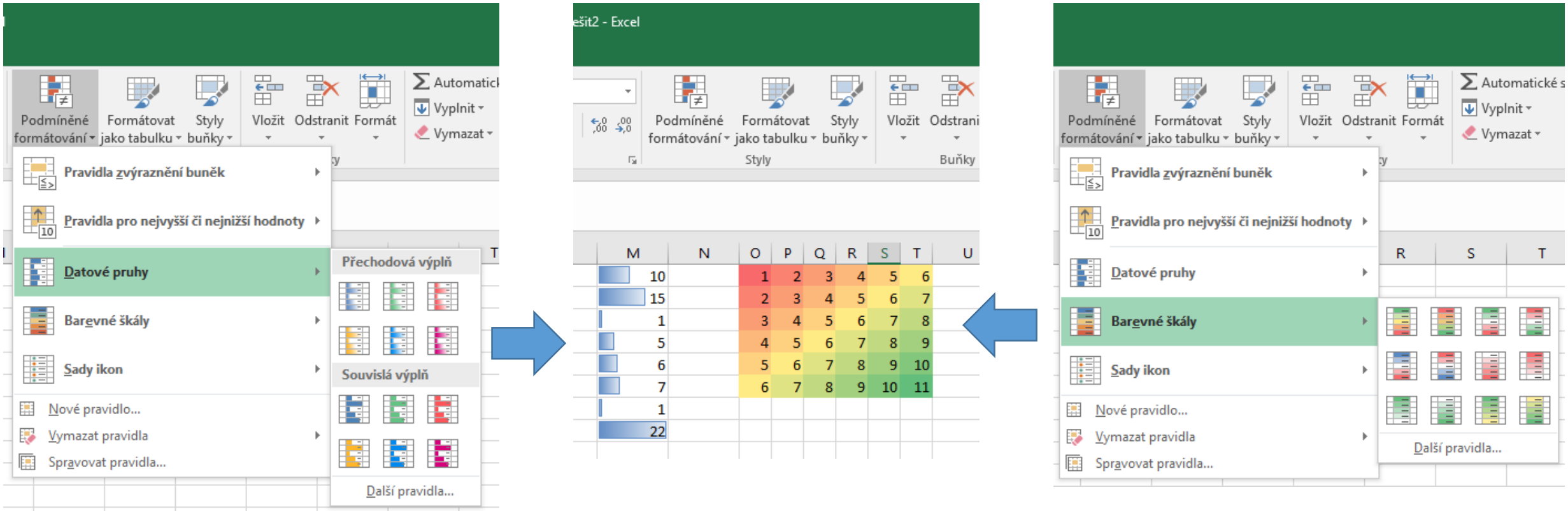
Invenční využití jednoduchých grafů: Demografická pyramida

- Jednoduchý ležatý sloupečkový graf
- Atraktivní vizualizace pro srovnání dvou skupin objektů



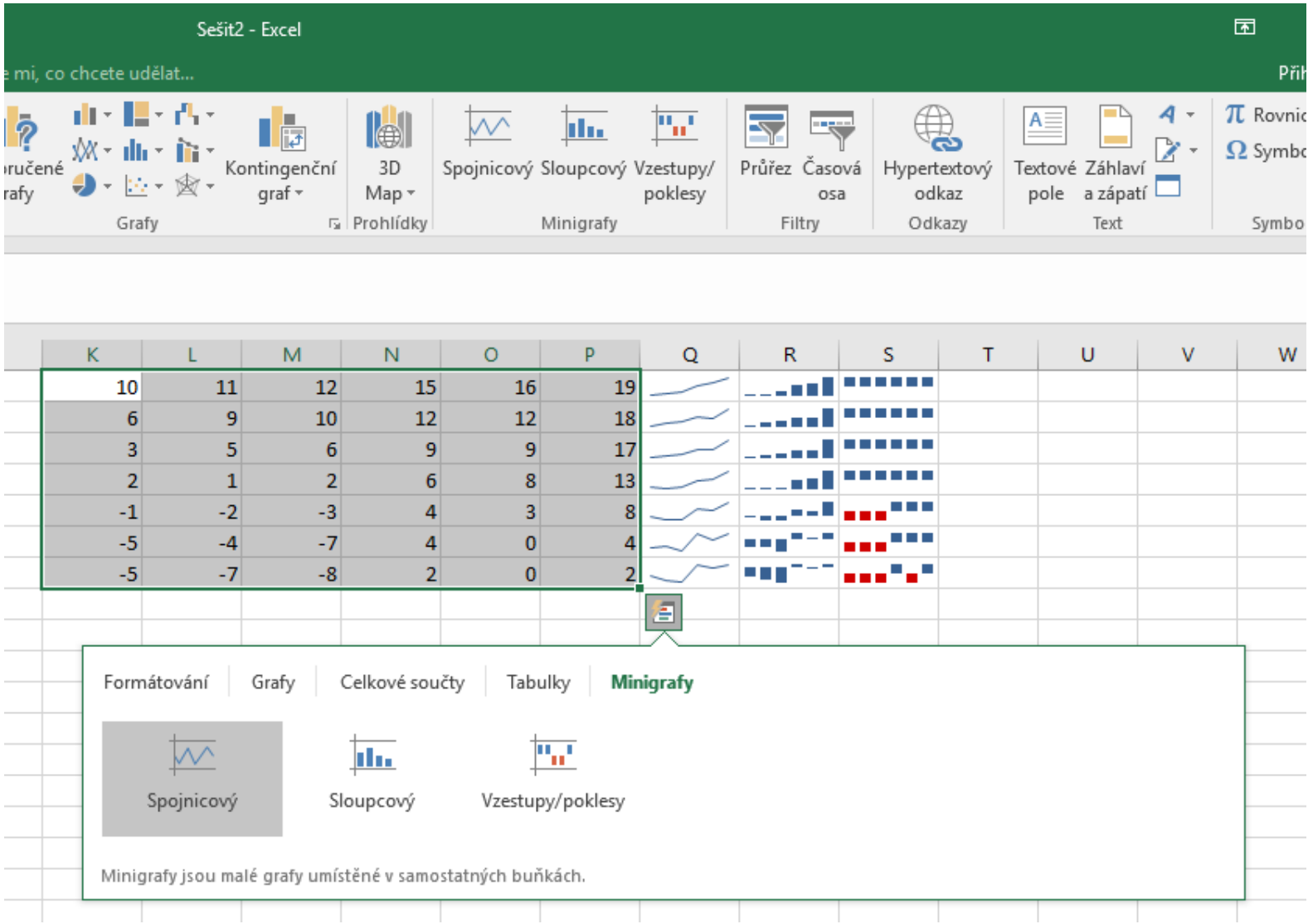
Excel – podmíněné formátování jako grafy

- Pro zpřehlednění excelových tabulek je možné využít grafické prvky v jeho buňkách
- Datové pruhy a barevné škály



Excel – grafy v buňkách

- Pro zpřehlednění excelových tabulek je možné využít grafické prvky v jeho buňkách
- Několik typů grafů umožňujících vizualizovat v jedné buňce datové řady
- Základní možnosti editace os a vzhledu



Heatmapa

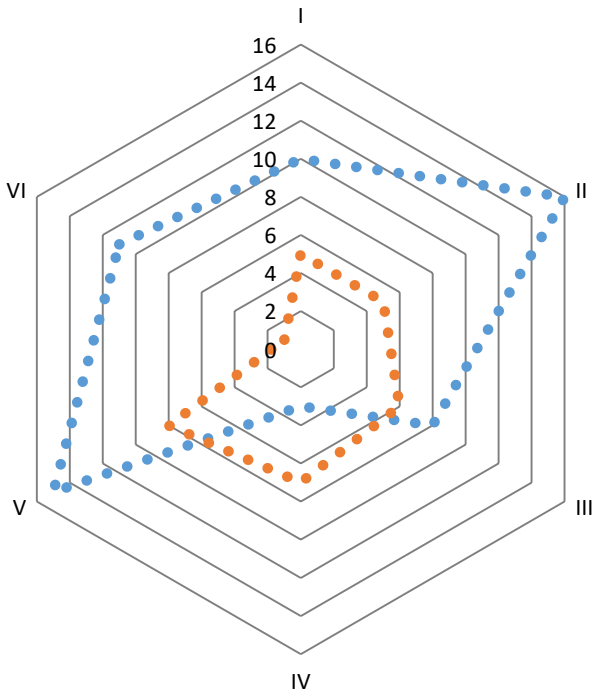
- Druh 3D grafu – osy tvoří dvě proměnné, barva třetí proměnnou
- Lze vytvořit v excelu pomocí podmíněného formátování
- Často ve vícerozměrné analýze pro vizualizaci asociačních matic

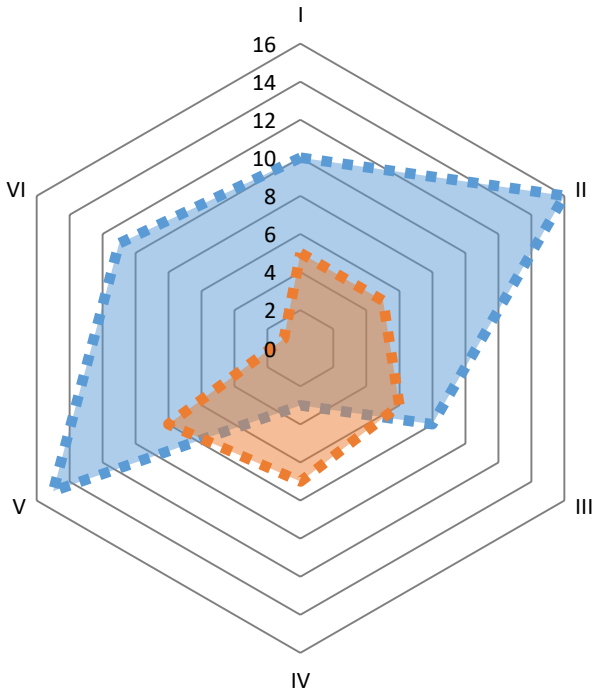
Výskyt indikátorového organismu v závislosti na dvou proměnných

Hloubka v cm vs. Koncentrace polutantu	< 60	60-69	70-74	75-79	80-84	85-89	90-94	95-99	100-109	110-119	120+
<= 30	29.8%	29.2%	27.9%	23.0%	20.5%	19.9%	20.6%	22.1%	22.1%	22.9%	23.3%
31-35	29.4%	28.2%	26.5%	22.0%	20.0%	19.5%	20.4%	21.6%	21.8%	22.6%	23.1%
36-39	18.5%	16.3%	15.8%	13.2%	12.9%	14.1%	15.3%	18.2%	20.4%	23.9%	28.4%
40-44	14.6%	14.3%	12.9%	12.0%	14.3%	20.2%	24.5%	22.2%	21.3%	20.2%	25.0%
45-49	12.6%	11.7%	13.0%	15.0%	17.9%	21.4%	22.5%	19.6%	20.3%	21.1%	30.0%
50+	12.2%	11.4%	13.6%	17.5%	22.0%	25.6%	25.9%	20.4%	19.9%	20.3%	31.3%

Pavoučí / paprskové grafy

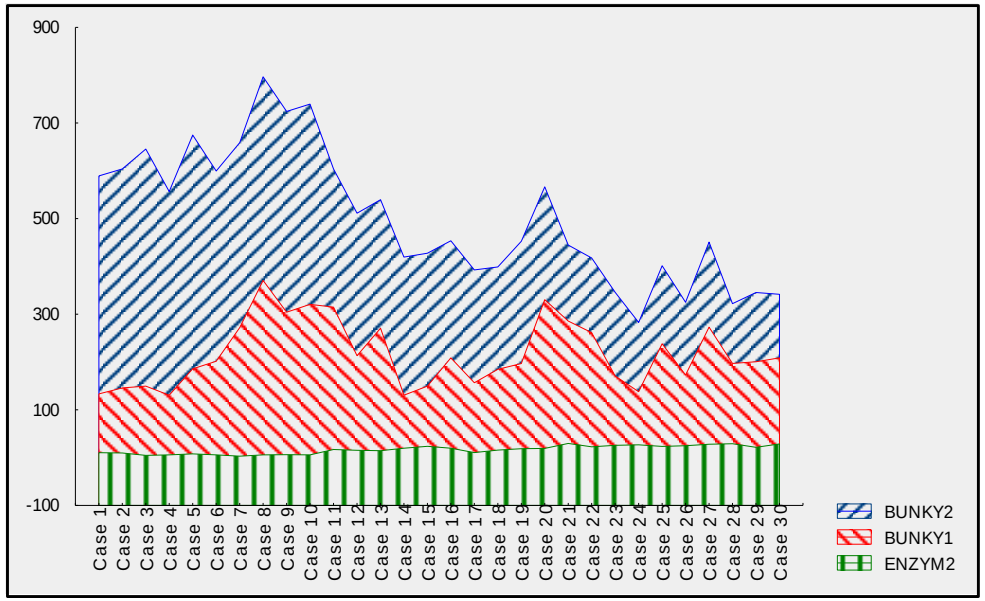
- Vhodné pro srovnání profilů objektů nebo skupin objektů pomocí více proměnných
- Různá grafická forma

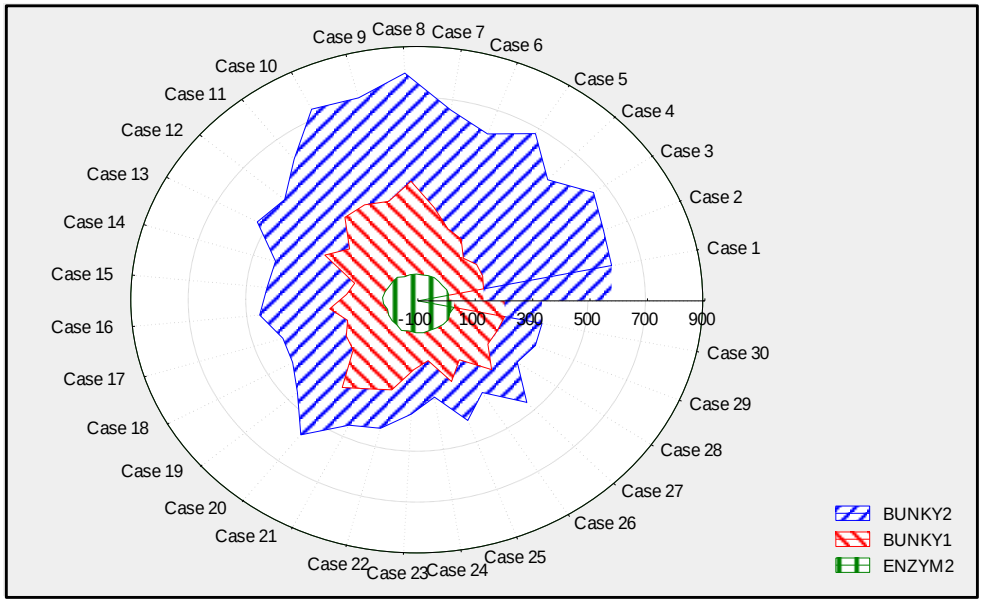




Polární graf

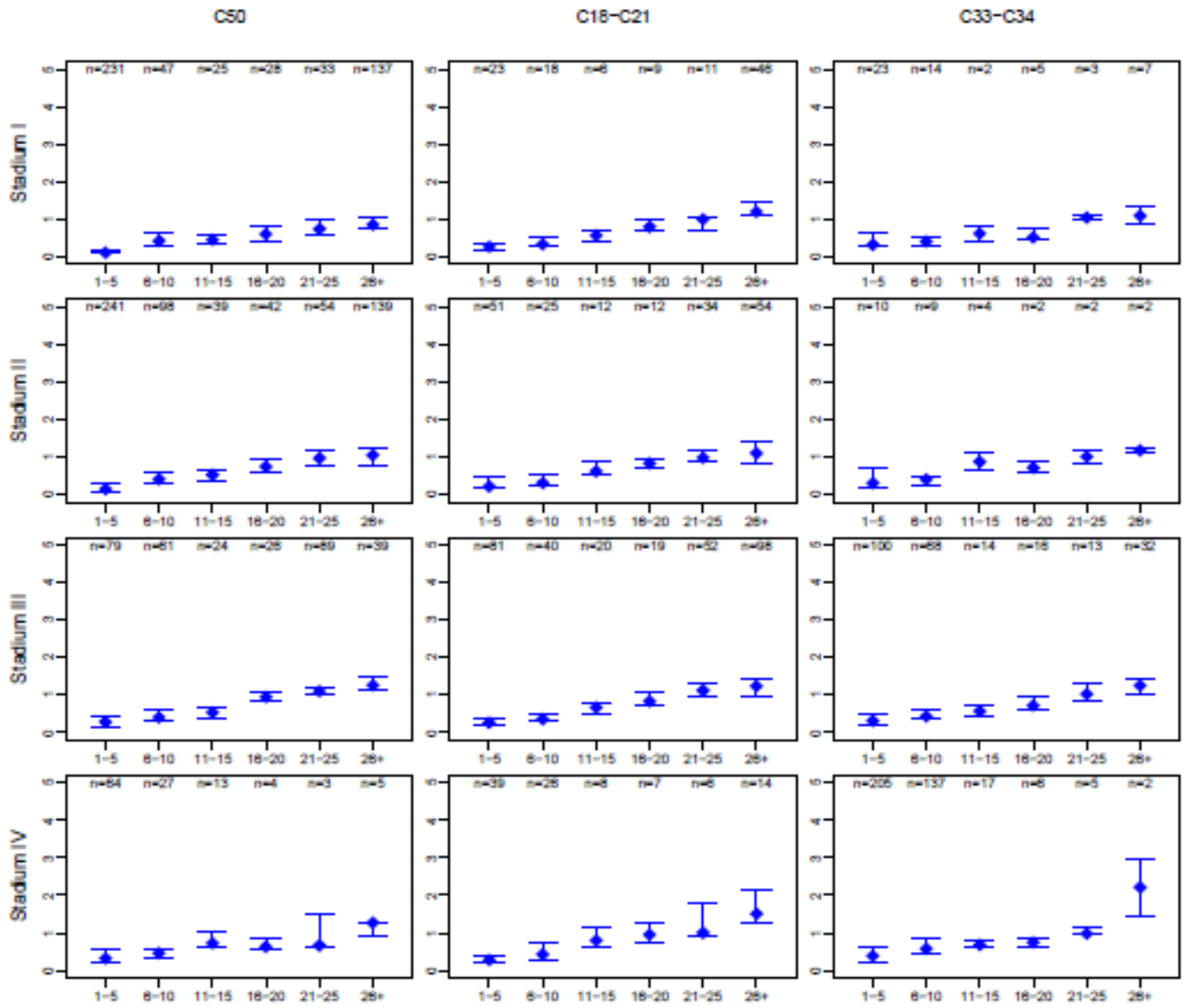
- Obdoba čárového, sloupcového nebo plošného grafu s osou X vynesenou na kružnici
- Vhodný pro cyklická data (cirkadiánní rytmy, sezonalita, směrová statistika pohybu živočichů)





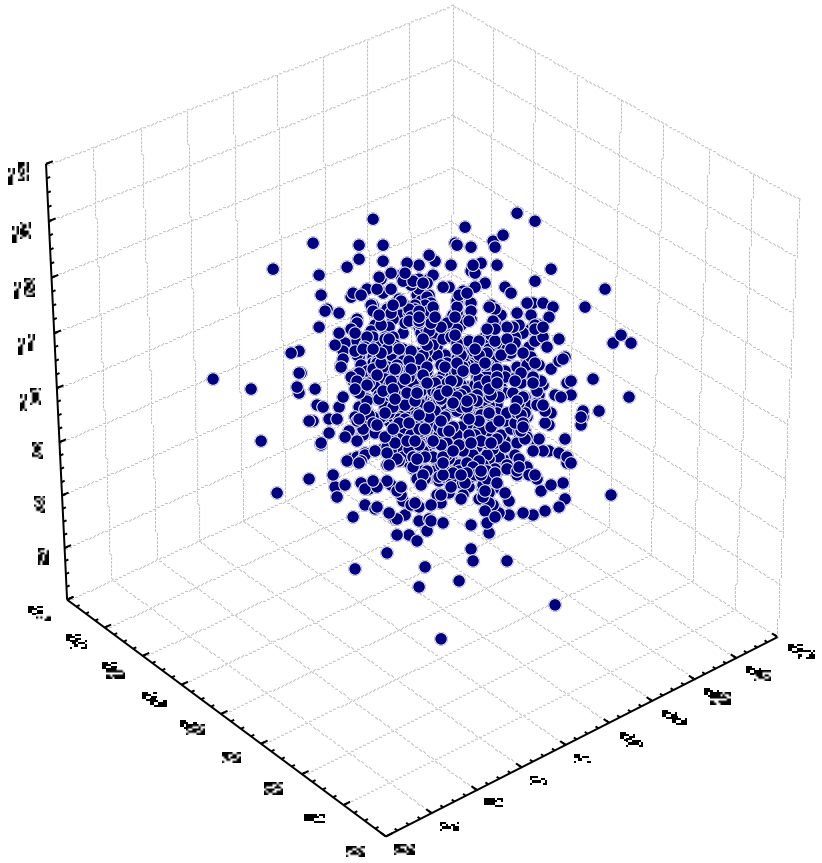
Grafické tabule

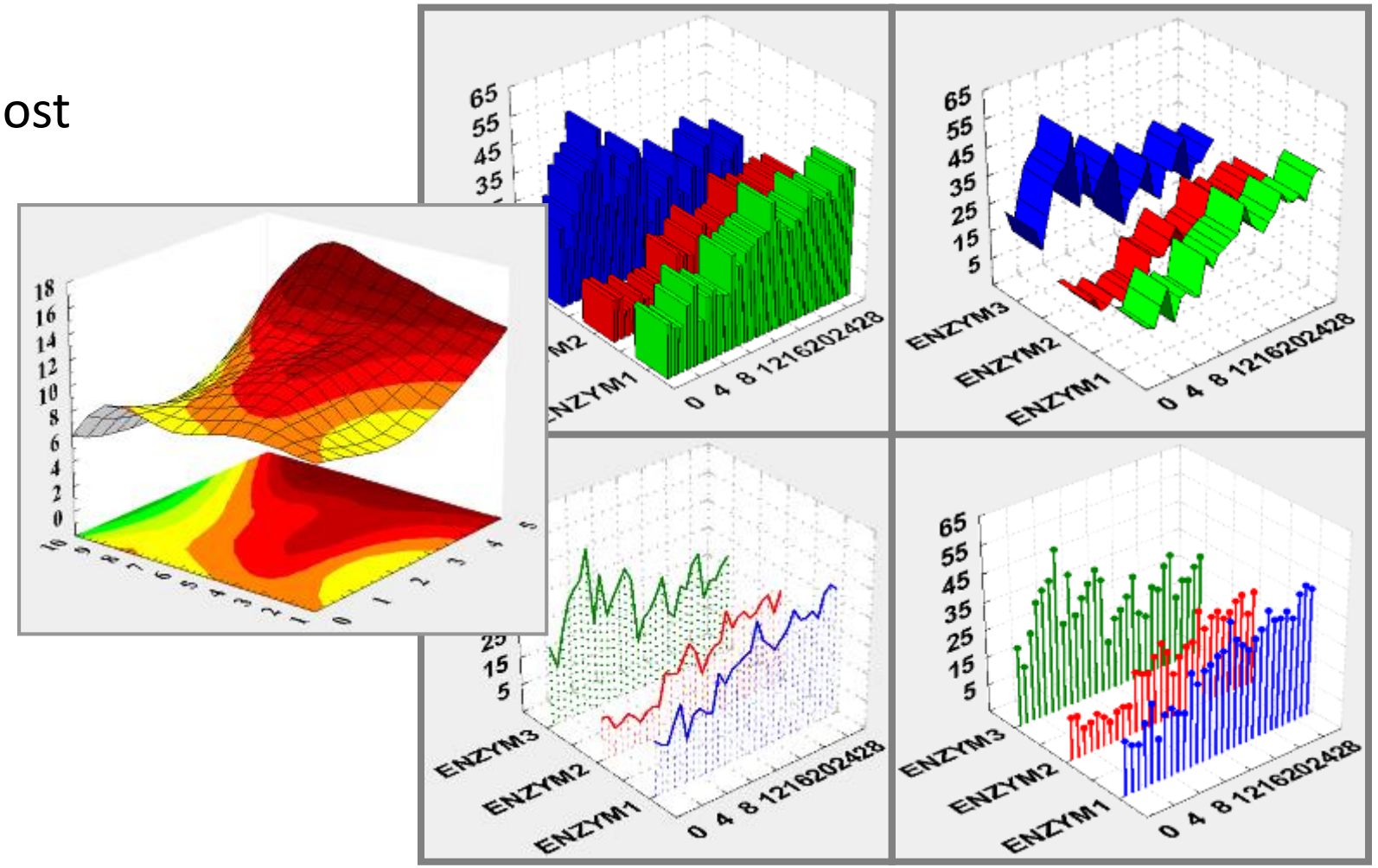
- Více grafů tvořících grafickou tabuli
- Možné skládat z různých grafů jednoho nebo více typů
- Prezentace velkého množství dat na malém prostoru



3D grafy

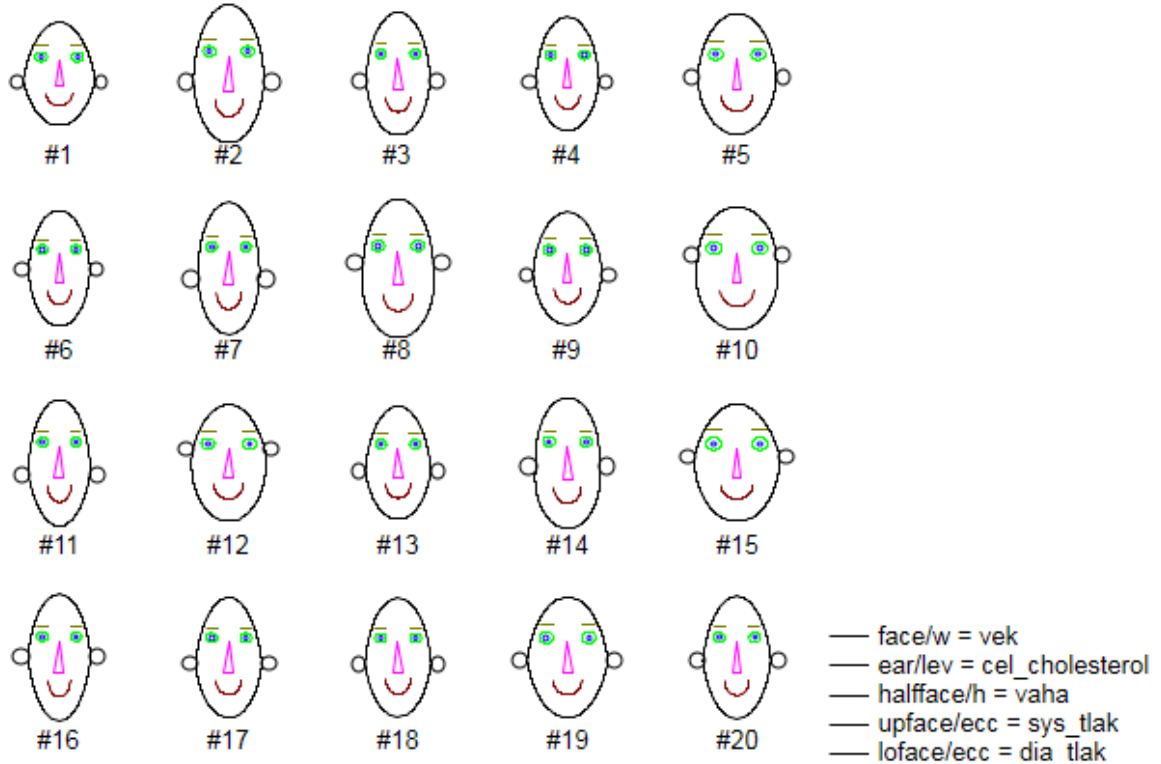
- Mnoho typů
- Velký důraz je třeba klást na interpretovatelnost a smysluplnost

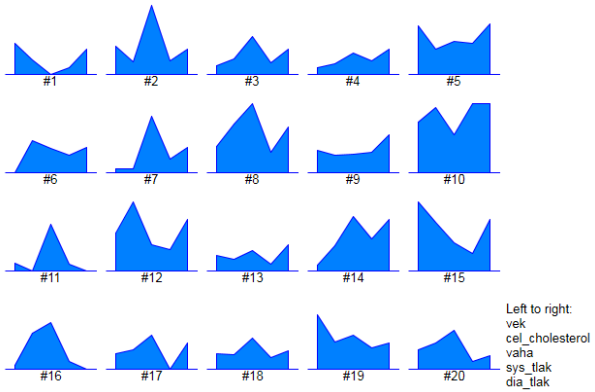


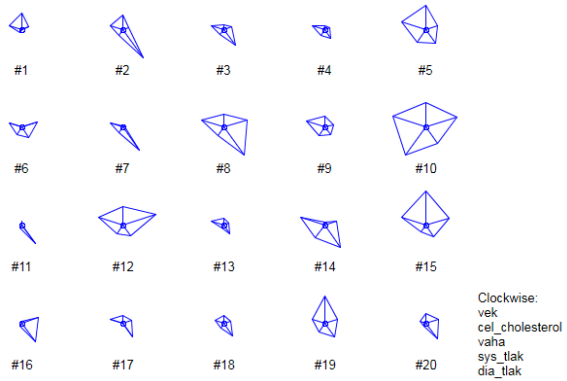


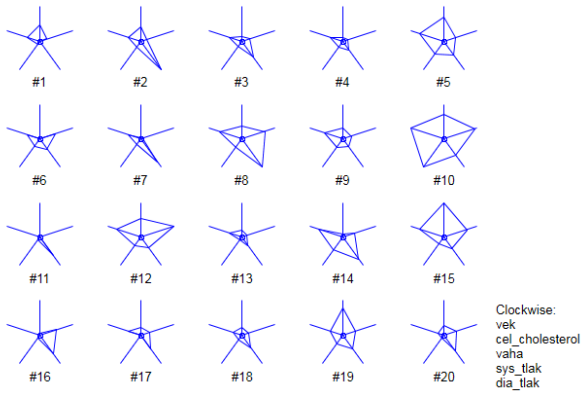
Chernoffovy tváře (ikonové grafy)

- Jednotlivé proměnné jsou zobrazeny jako rysy tváře
- Patří mezi tzv. ikonové grafy
  - hodnoty znaků znázorněny jako geometrické útvary či symboly
  - každému objektu (subjektu) odpovídá jeden obrazec složený z těchto geometrických útvarů či symbolů
  - umožní vizuálně porovnat, které objekty (subjekty) jsou si podobné



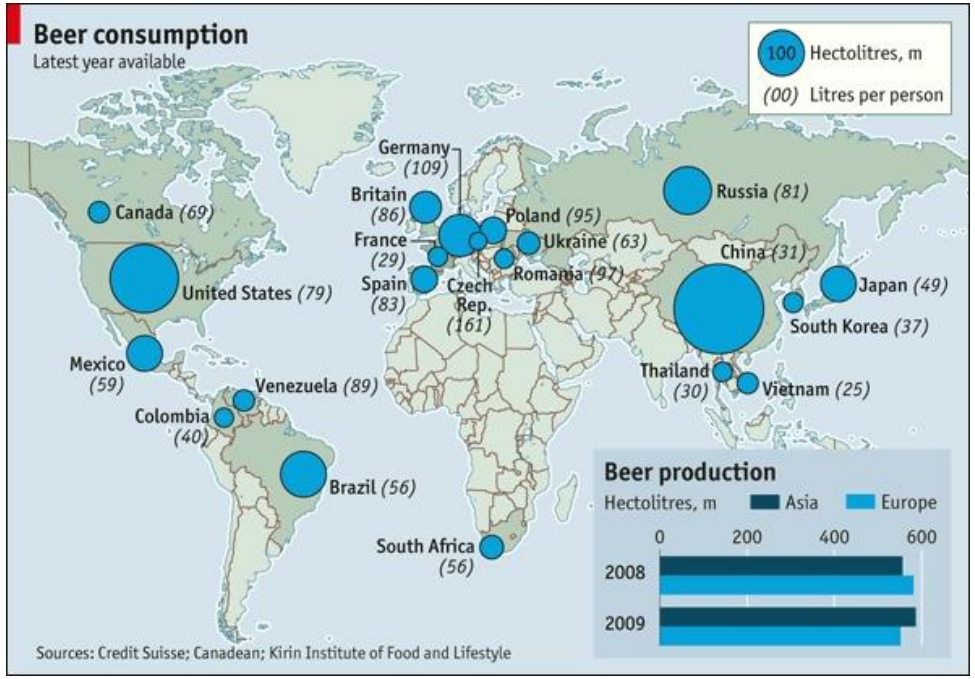


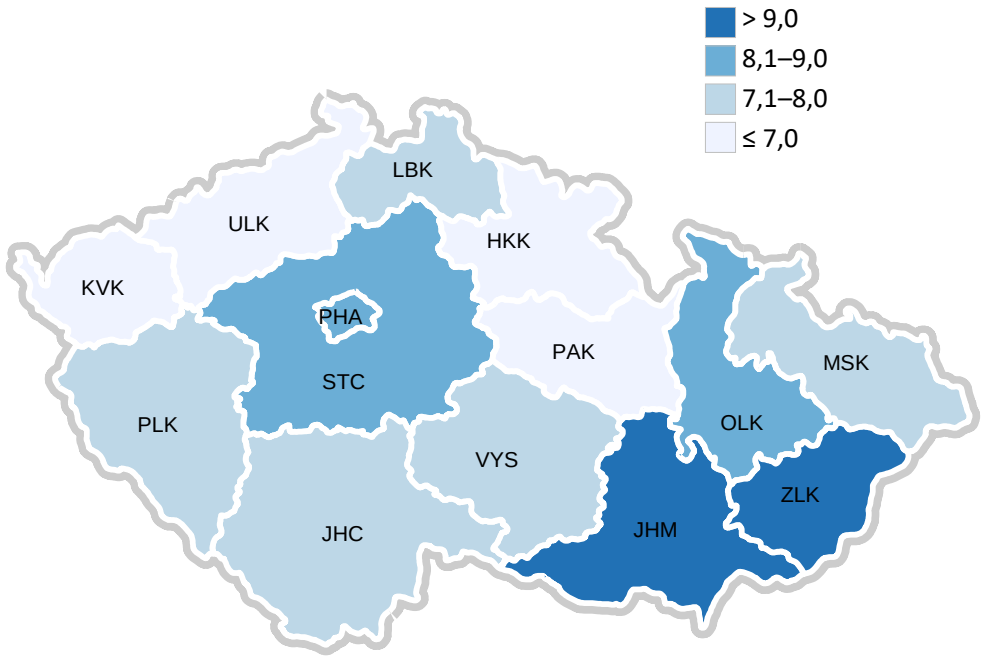




Mapy jsou také grafy

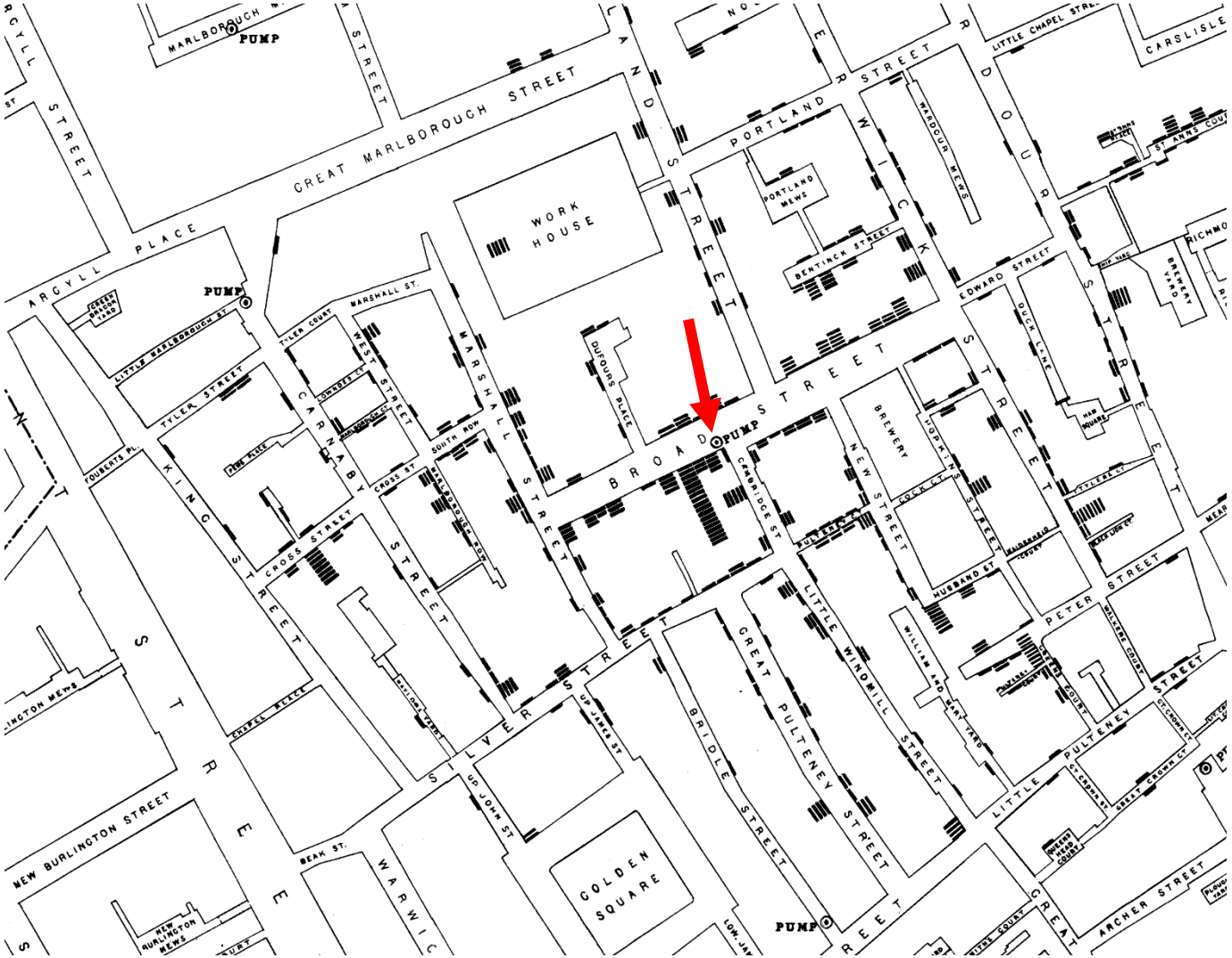
- Samostatná kapitola vizualizace dat
- Obarvení regionů v mapě dle výsledků analýzy nebo přímo vkládání grafů do map (sloupcové, koláčové atd.)
- ArcGIS – další z SW dostupných na inet.muni.cz



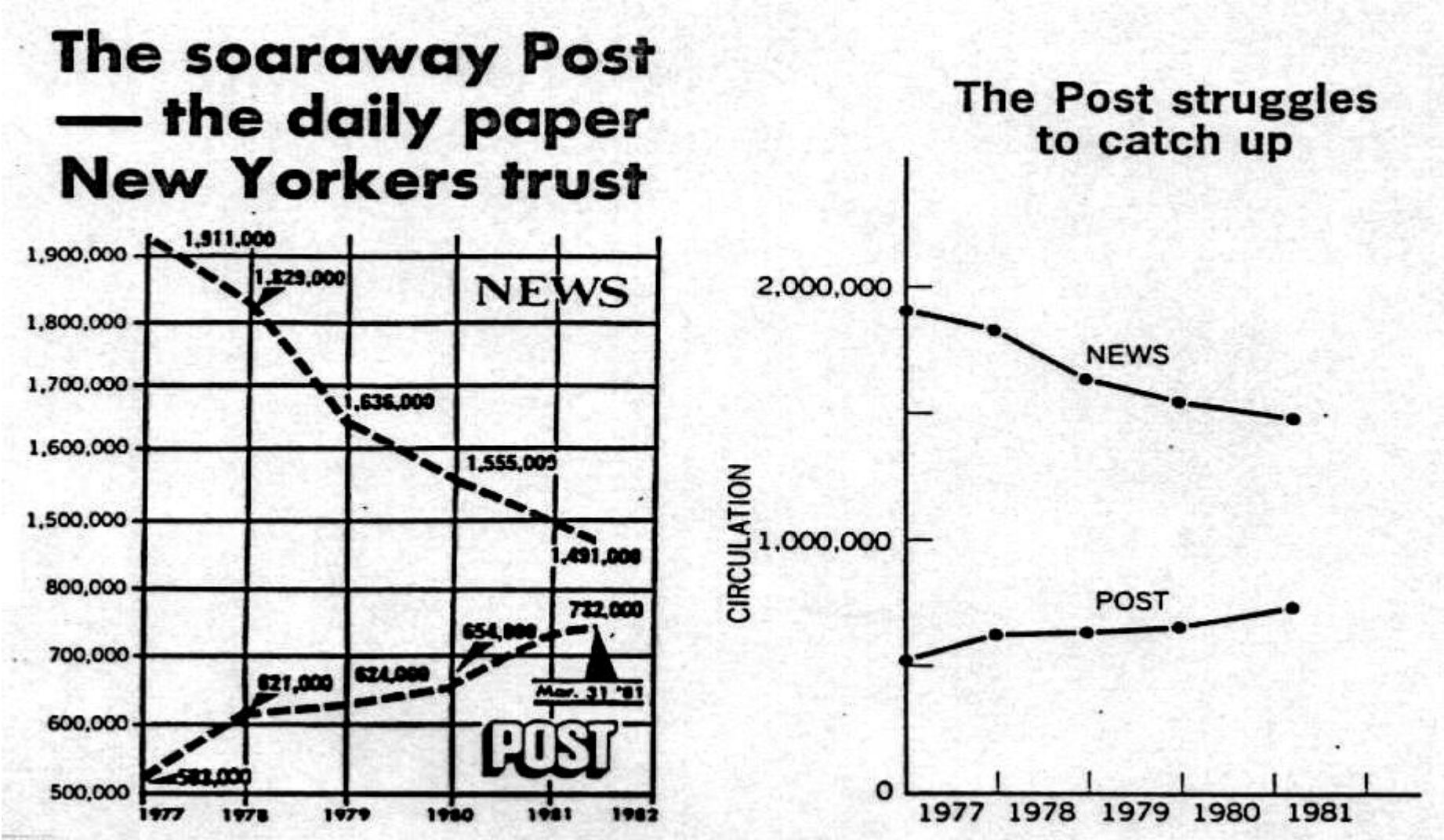


Slavné mapy: John Snow – cholera v Londýně

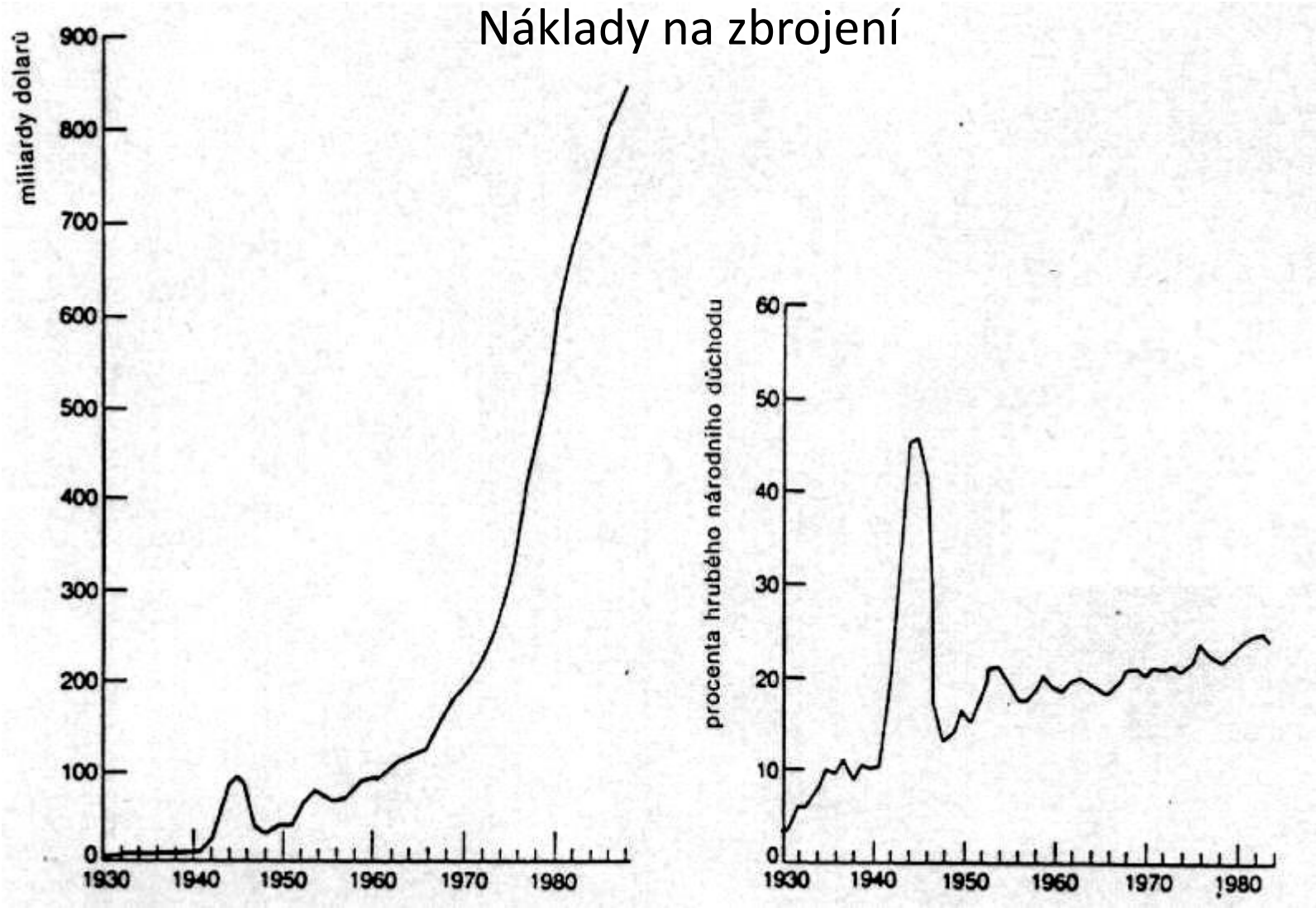
- 1854 Broad Street cholera outbreak
- Počty případů vyneseny jako černé sloupce dle bydliště obětí
- Identifikace zdroje nákazy – kontaminovaná studně
- Jeden z prvních příkladů prostorové analýzy dat a epidemiologického mapování



Nesprávné použití grafů: rozsah os („nevíme jak nakreslit“)



Nesprávné použití grafů: standardizace os („nevíme co kreslíme“)



Přednáška 3

Informace a rozdělení dat

Jak vznikají informace

Rozdělení dat

Anotace

- Základním principem statistiky je pravděpodobnost výskytu nějaké události.
- Prostřednictvím vzorkování se snažíme odhadnout skutečnou pravděpodobnost událostí.
- Klíčovou otázkou je velikost vzorku, čím větší vzorek, tím větší šance na projevení se skutečné pravděpodobnosti výskytu jevu.

Vznik informací: pojmy I

Skutečnost

Pozorovatel

Jev - podmnožina všech možných výsledků pokusu/děje, o které lze říct, zda nastala nebo ne

Jevové pole - třída všech jevů, které jsme se rozhodli nebo jsme schopni sledovat

Skutečnost + Jevové pole = Měřitelný prostor

Vznik informací: pojmy II

- **Experimentální jednotka** - objekt, na kterém se provádí šetření
- **Populace** - soubor experimentálních jednotek (objekt)
- **Znak** - vlastnost sledovaná na objektu
- **Náhodná veličina** - číselná hodnota vyjadřující výsledek náhodného experimentu

- Znak se stává **sledovanou náhodnou veličinou**, pokud se jeho hodnota zjišťuje **vylosováním (vzorkováním)** objektu ze **základního souboru (populace)**

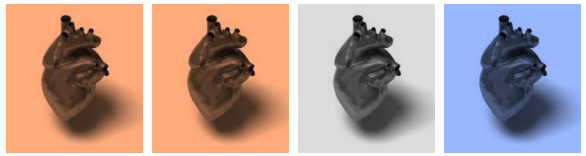
Vznik informací: vzorkování

Statistika hovoří o realitě prostřednictvím výběru z cílové populace

Statistické předpoklady korektního vzorkování je nutné dodržet

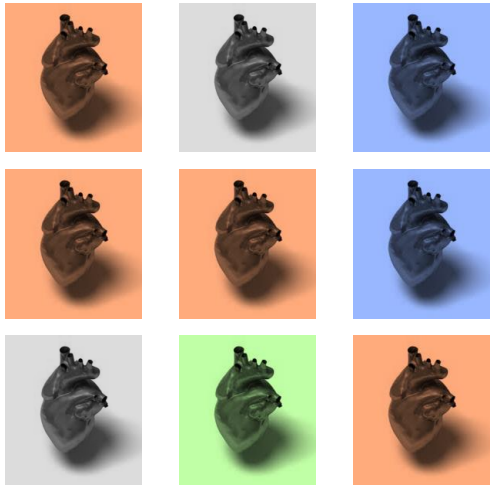
Náhodný výběr z cílové populace

Representativnost: struktura vzorku musí maximálně reflektovat realitu



Nezávislost: několikanásobné vzorkování téhož objektu nepřináší ze statistického hlediska žádnou novou informaci

Cílová populace

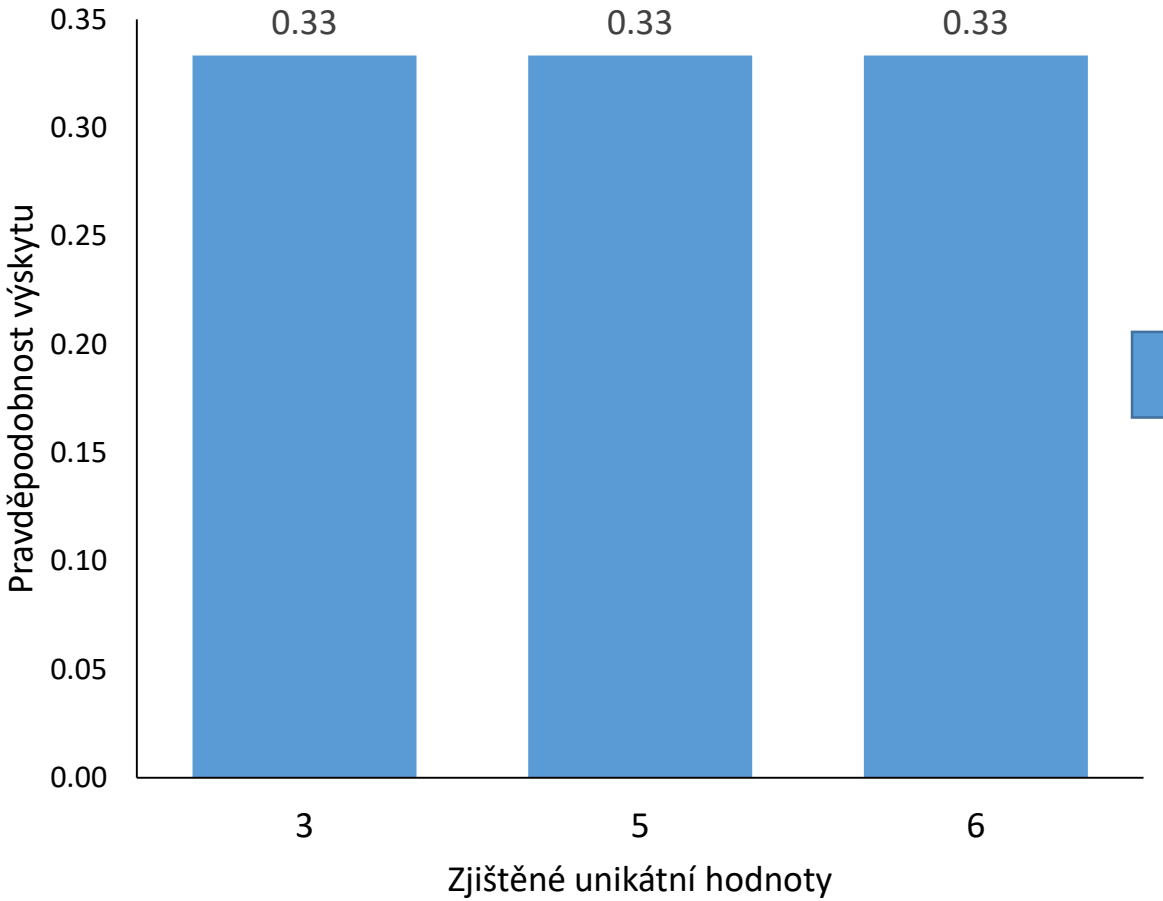


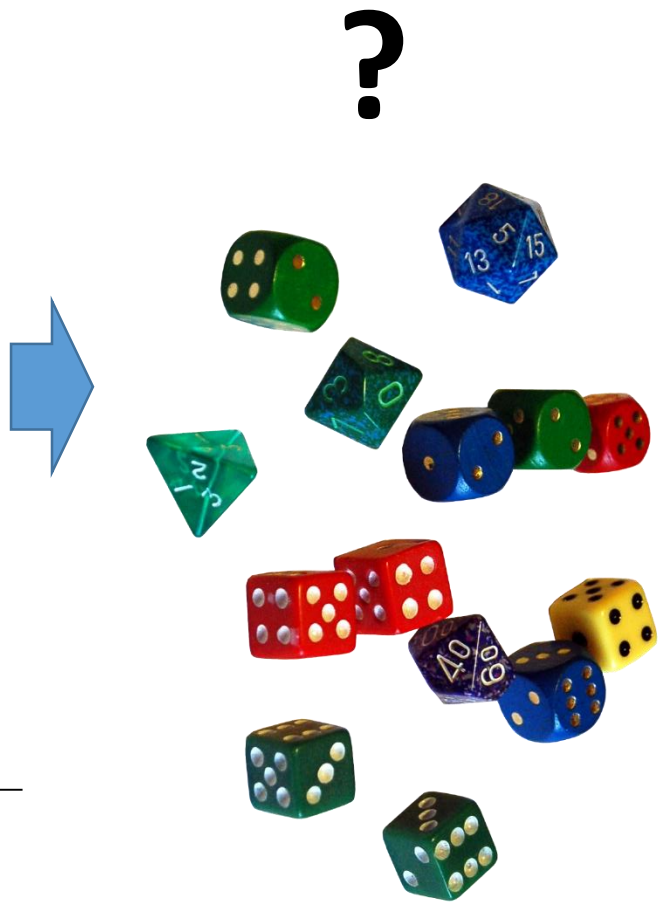
Příklad vzorkování

- Na základě vzorkování chceme zjistit vlastnosti nějakého jevu
- Naší cílovou populací budou hody kostkou s neznámými vlastnostmi

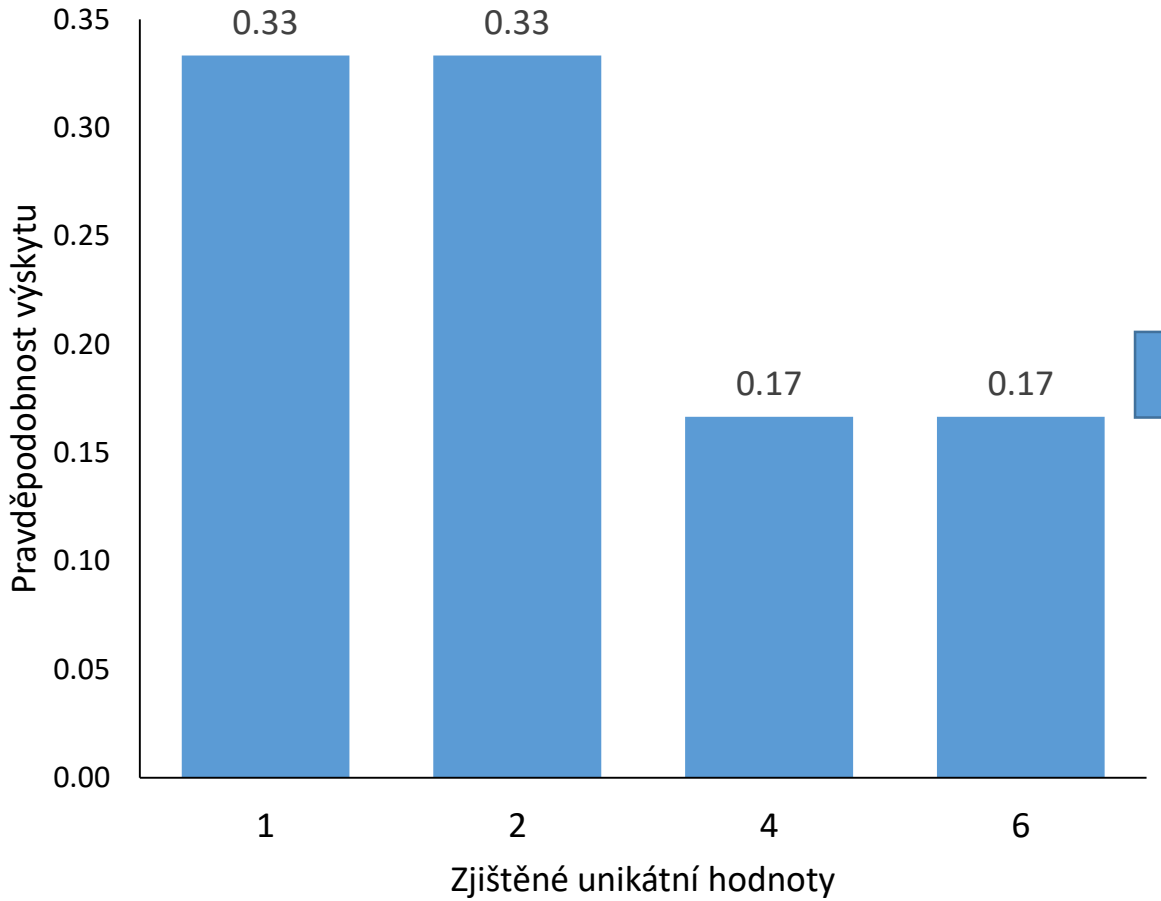
- Chceme zjistit vlastnosti neznámé použité kostky

Příklad vzorkování: N=3

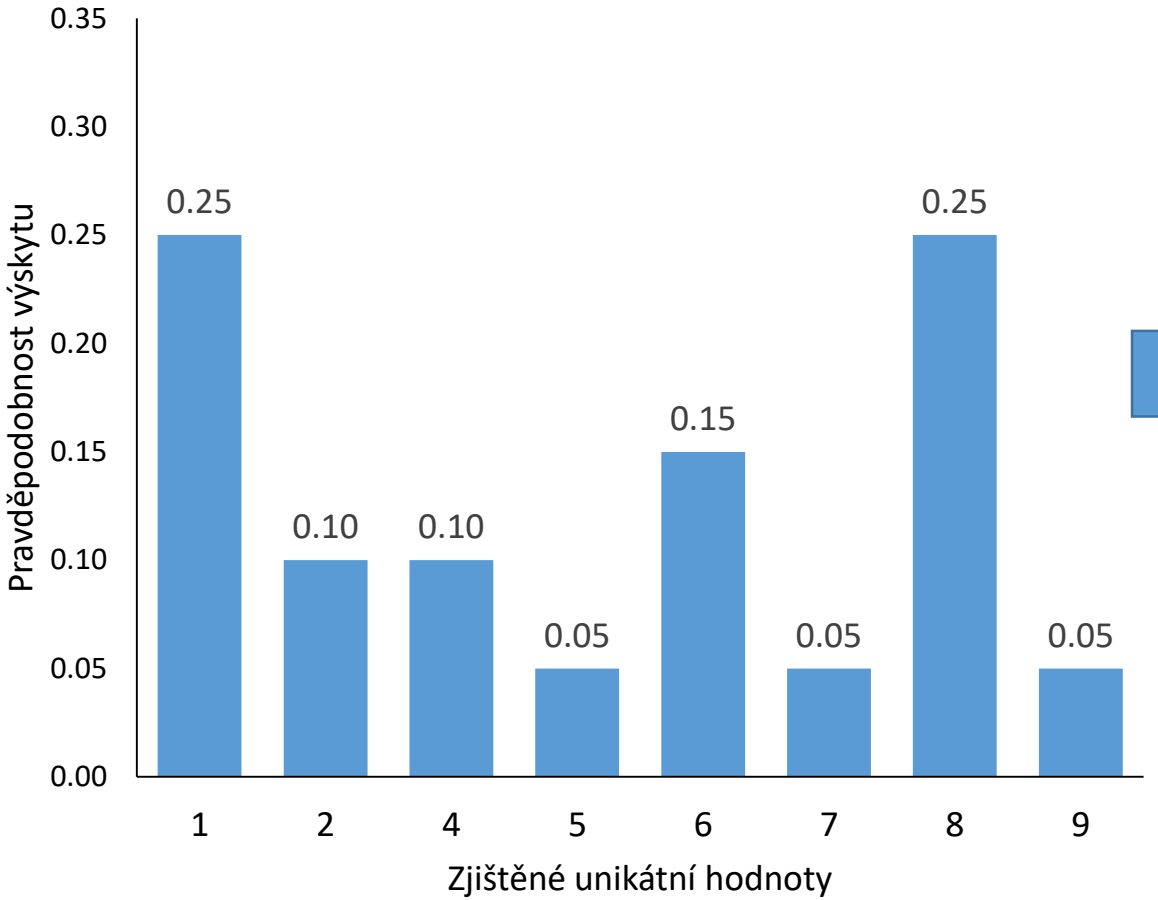




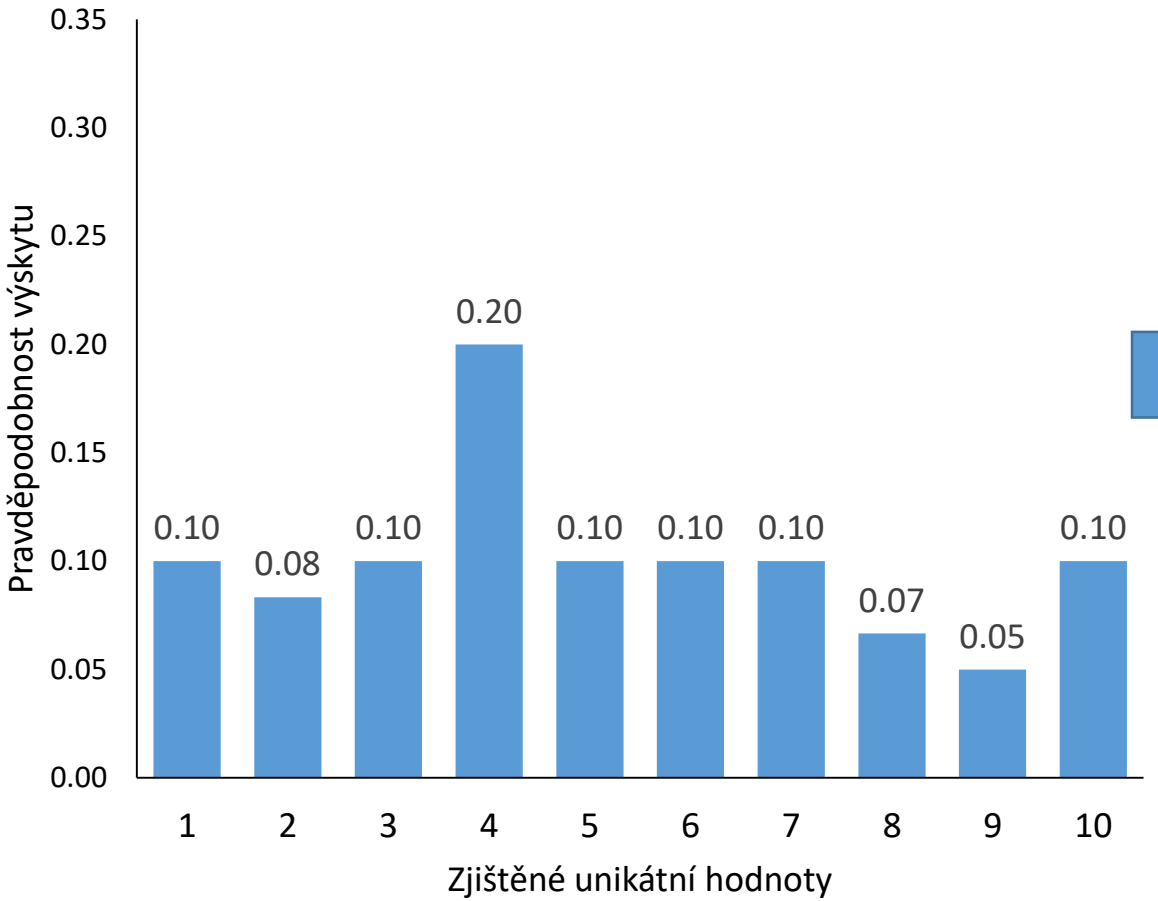
Příklad vzorkování: N=6

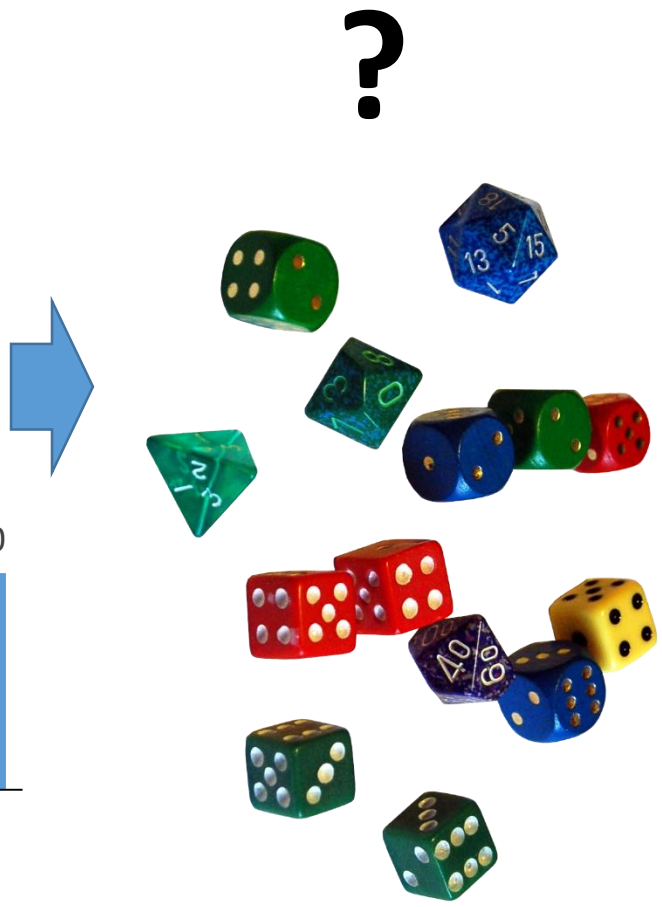


Příklad vzorkování: N=20

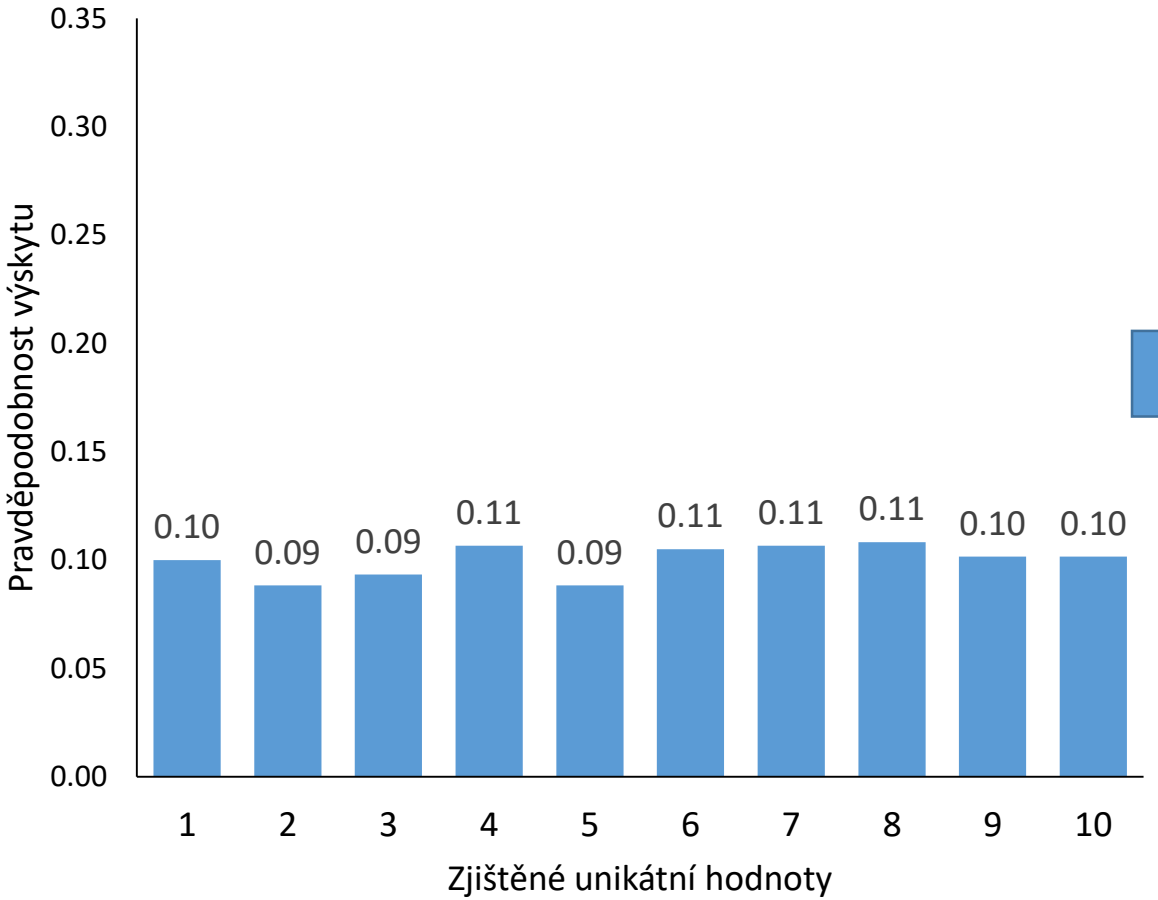


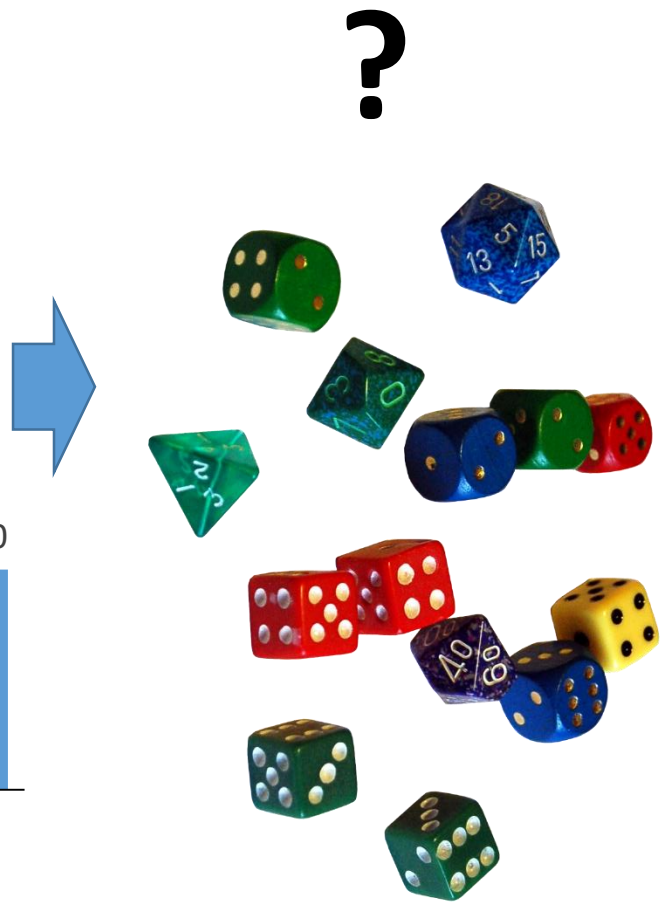
Příklad vzorkování: N=60



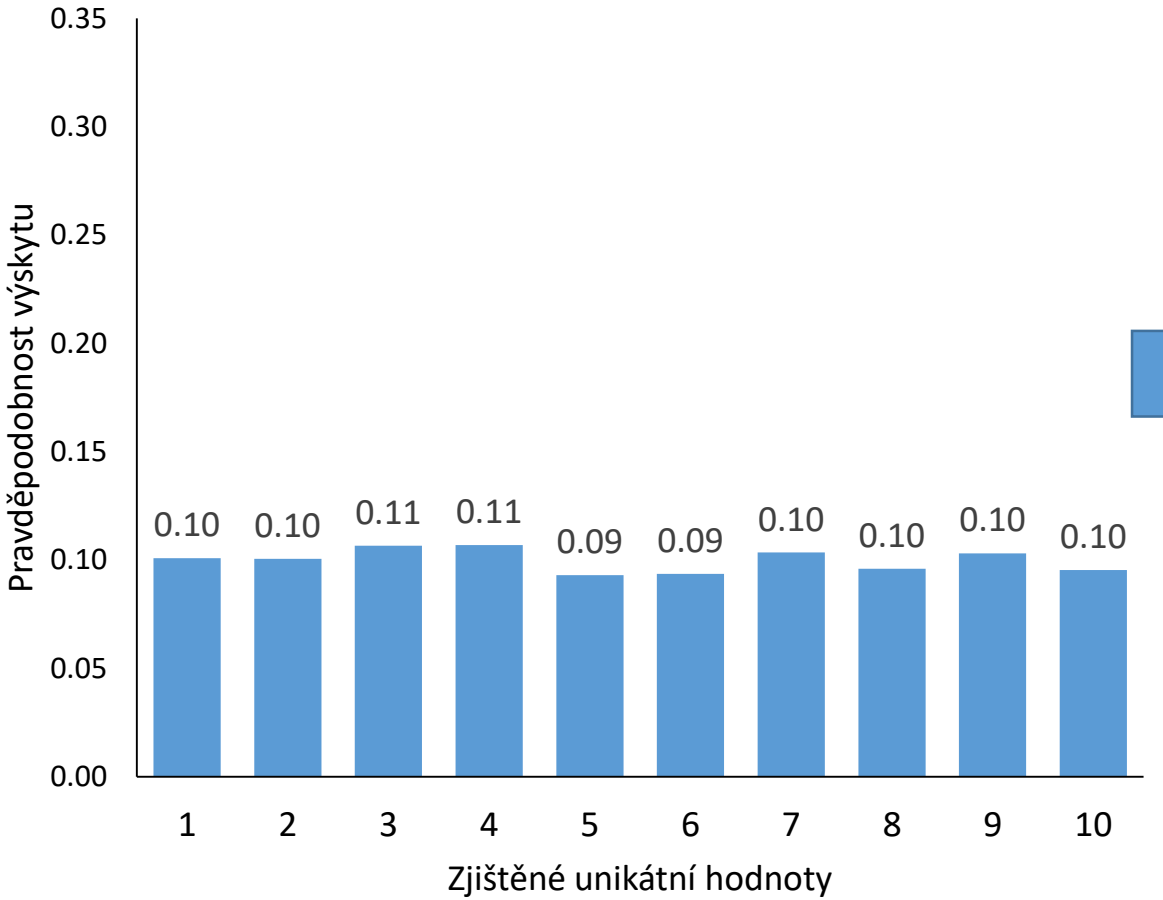


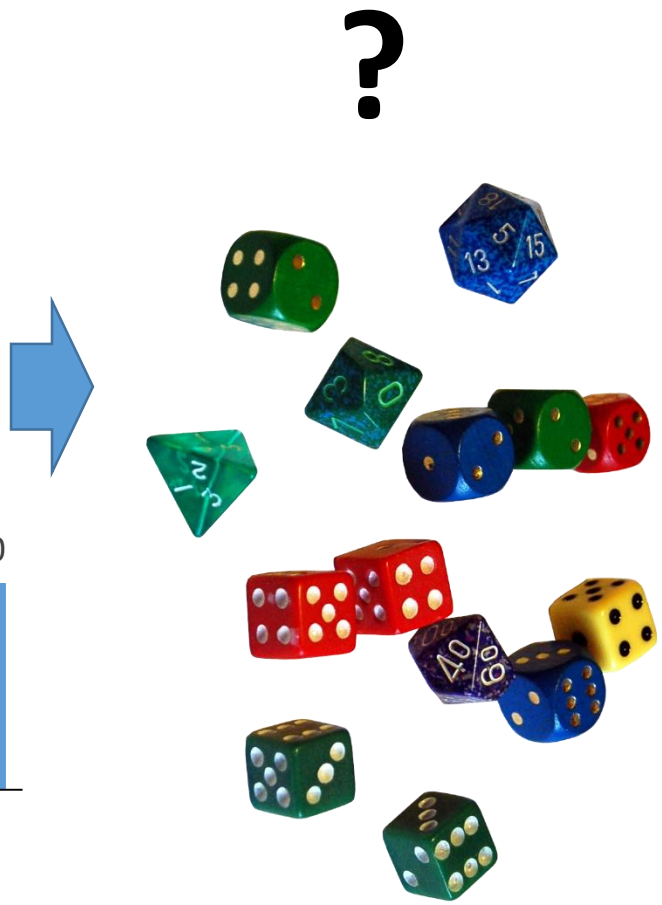
Příklad vzorkování: N=600



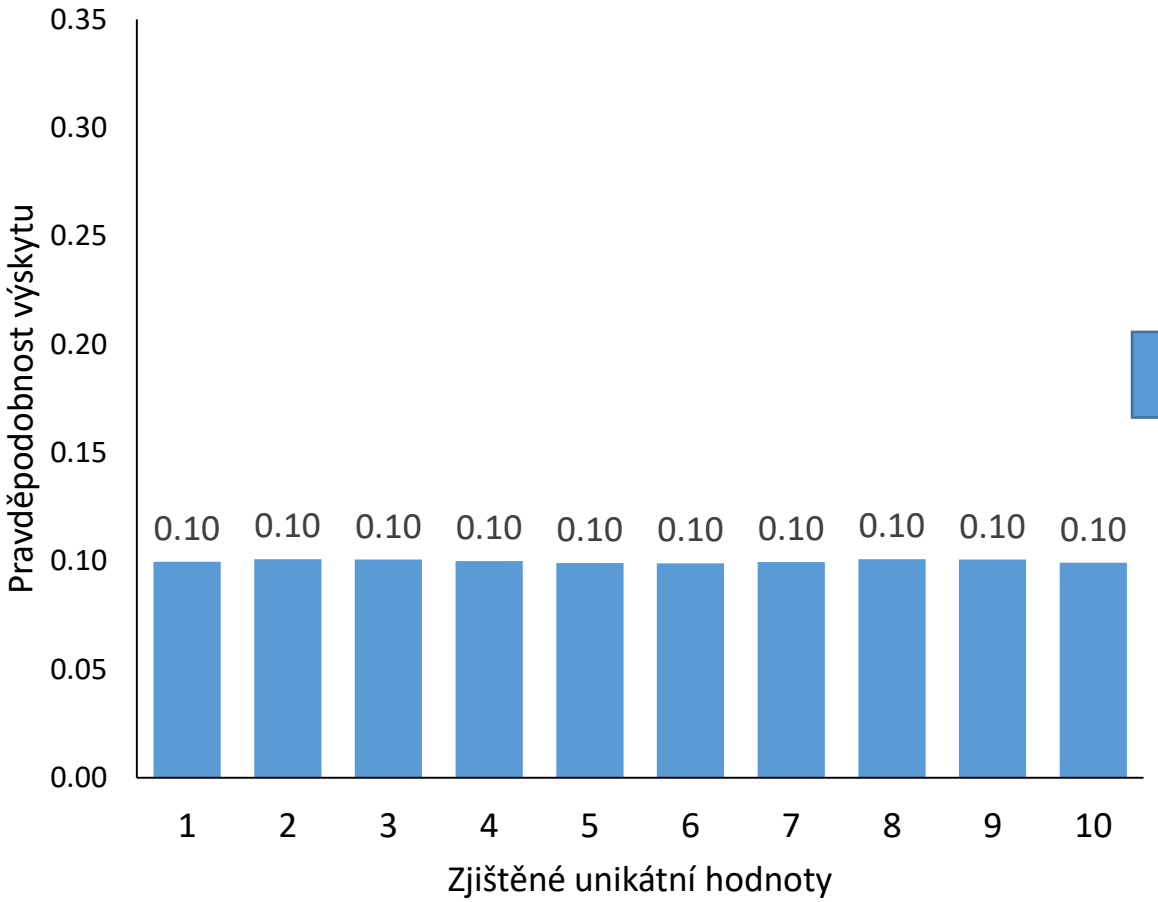


Příklad vzorkování: N=6 000





Příklad vzorkování: N=60 000

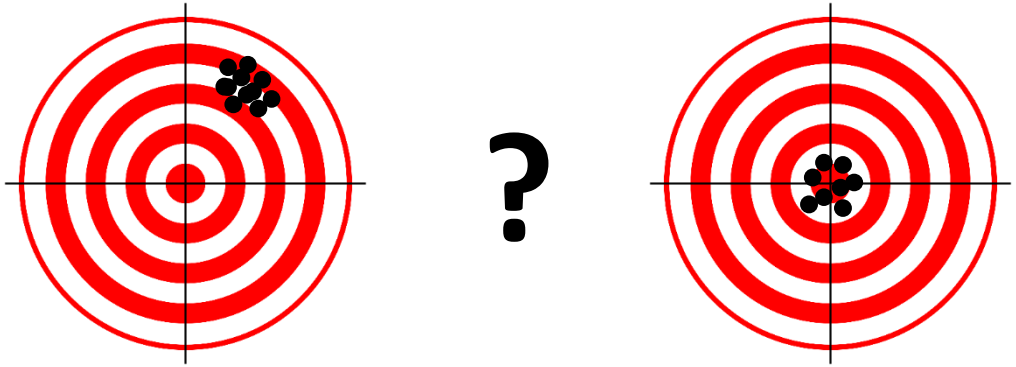


?

Příklad vzorkování: závěr

- Sledovaný jev má pravděpodobně tvar desetistěnné kostky

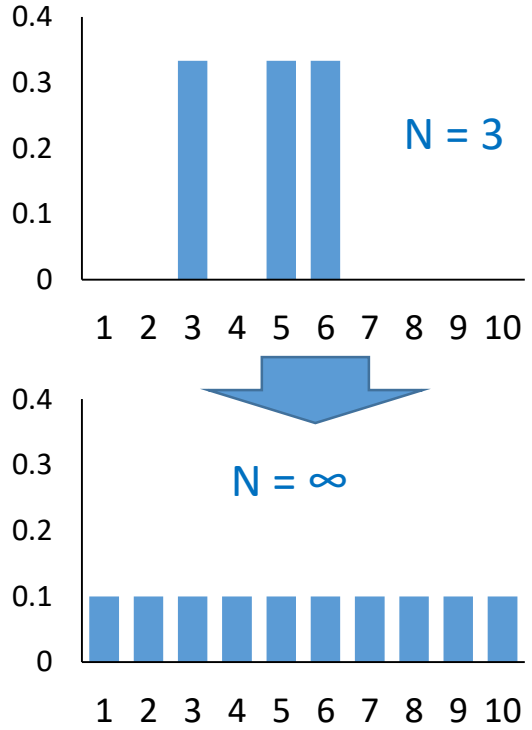
- U složitých stochastických systémů se pravda získá až po odvedení značného množství experimentální práce: musíme dát systému šanci se projevit
- Při realizaci náhodného experimentu roste se zvyšujícím se počtem opakování pravdivá znalost systému (výsledky se stávají stabilnější a spolehlivější)
- Diskutabilní je ovšem míra zobecnění konkrétního experimentu (spolehlivost a stabilita výsledků není totéž co nezkreslený výsledek)



Empirický zákon velkých čísel

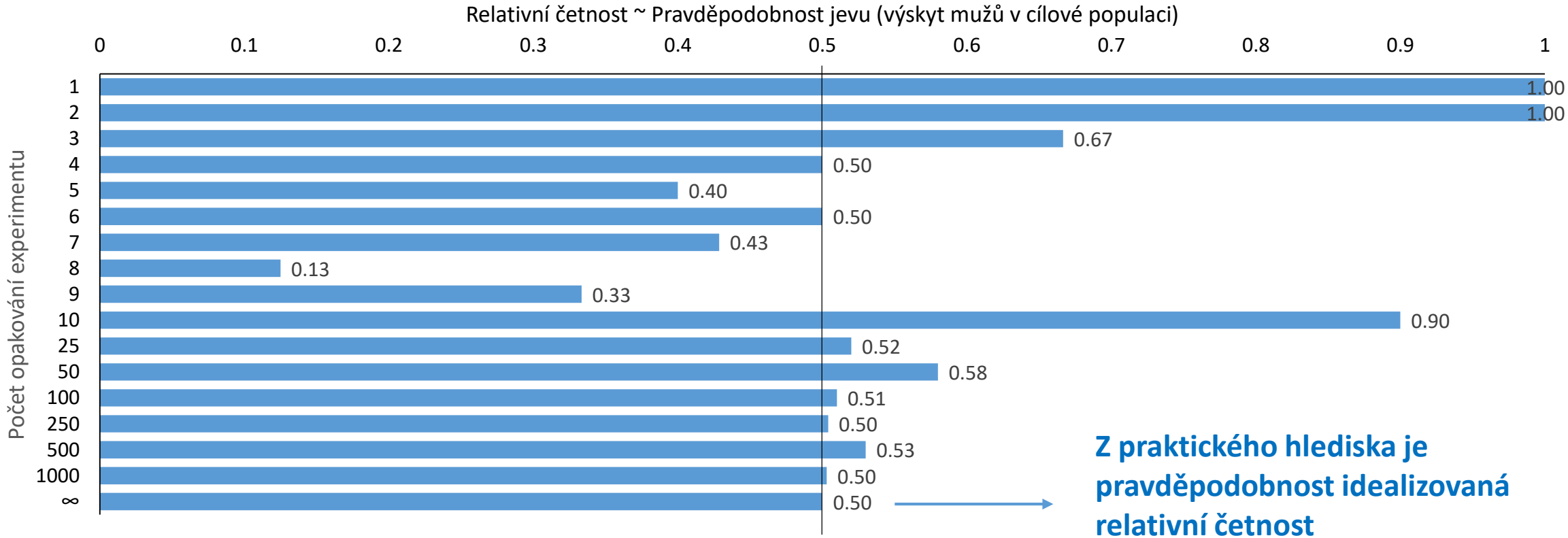
- Při opětovné nezávislé realizaci téhož náhodného experimentu se podíl výskytů sledovaného jevu mezi všemi dosud provedenými realizacemi zpravidla ustaluje kolem konstanty.
- Pravděpodobnost je libovolná reálná funkce definovaná na jevovém poli A (např. hody kostkou), která každému jevu A (např. strany kostky) přiřadí nezáporné reálné číslo P(A) z intervalu 0 - 1.
- **Z praktického hlediska je pravděpodobnost idealizovaná relativní četnost**

- P (A) = 1 …… jev jistý
- P (A) = 0 …… jev nemožný
- P (A ∩ B) = P (A) . P (B)…… nezávislé jevy
- P (A ∩ B) = P (A) . P (B/A) …… závislé jevy
- P (A / B) = P (A ∩ B) / P (B) ……. podmíněná pravděpodobnost



Empirický zákon velkých čísel: příklad

- Hodnotíme výskyt mužů v dané sledované populaci (jev „výskyt muže“)
- Skutečná pravděpodobnost sledovaného jevu je p=0.5 (tu ale ve skutečnosti neznáme)
- Snažíme se na základě opakovaného vzorkování (experimentu) tuto pravděpodobnost zjistit

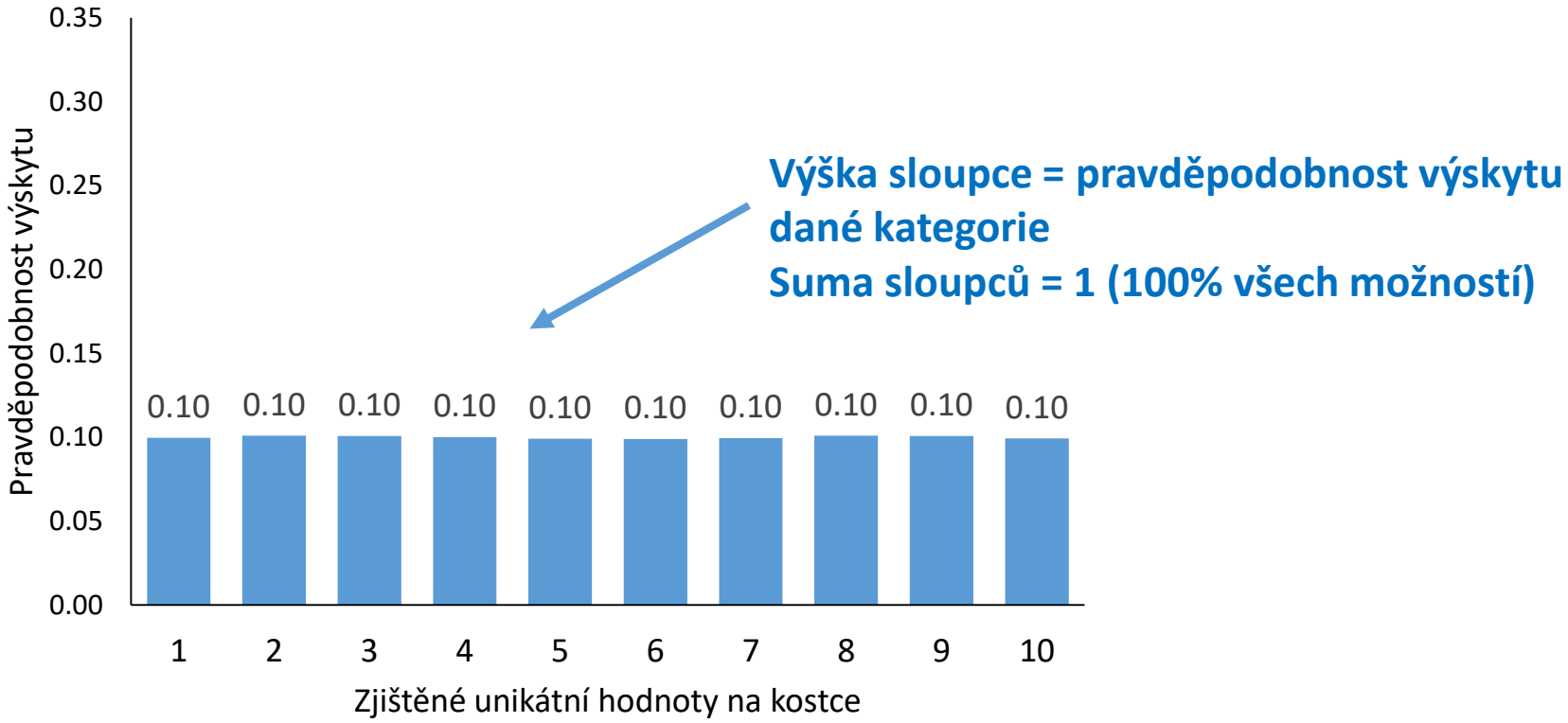


Z praktického hlediska je pravděpodobnost idealizovaná relativní četnost

P=0.5

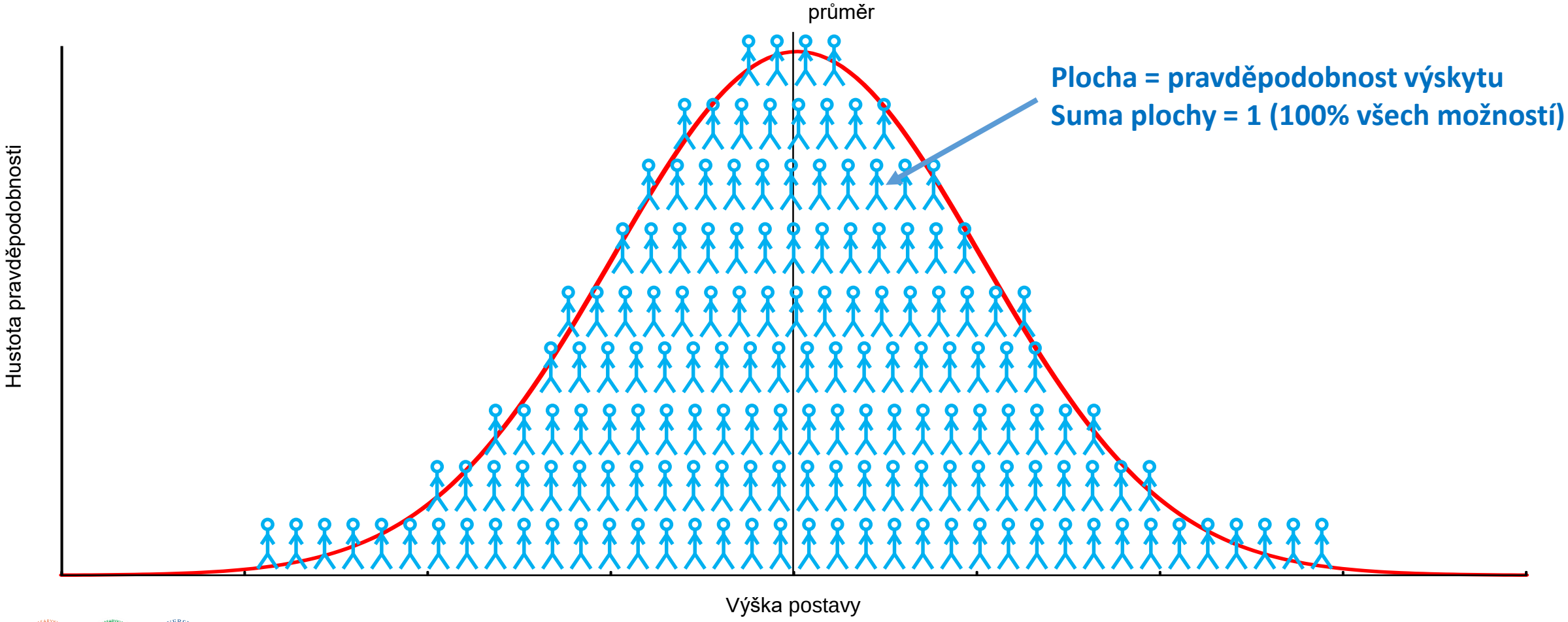
Pravděpodobnost výskytu jevu – rozložení kategoriálních dat

- existuje pravděpodobnost výskytu jevů (nedeterministické závěry)
- „vše je možné“: pouze jev s pravděpodobností 0 nikdy nenastane



Pravděpodobnost výskytu jevu – rozložení spojitých dat

- existuje pravděpodobnost výskytu jevů (nedeterministické závěry)
- „vše je možné“: pouze jev s pravděpodobností 0 nikdy nenastane



Základní typy dat

Spojitá a kategoriální data

Základní popisné statistiky

Grafický popis dat

Anotace

- Realitu můžeme popisovat různými typy dat, každý z nich se specifickými vlastnostmi, výhodami, nevýhodami a vlastní sadou využitelných statistických metod
- Od binárních přes kategoriální, ordinální až po spojitá data roste míra informace v nich obsažené.
- Základním přístupem k popisné analýze dat je tvorba frekvenčních tabulek a jejich grafických reprezentací – histogramů.

Jak vznikají data?

- Záznamem skutečnosti…

Jak vznikají data?

- Záznamem skutečnosti…

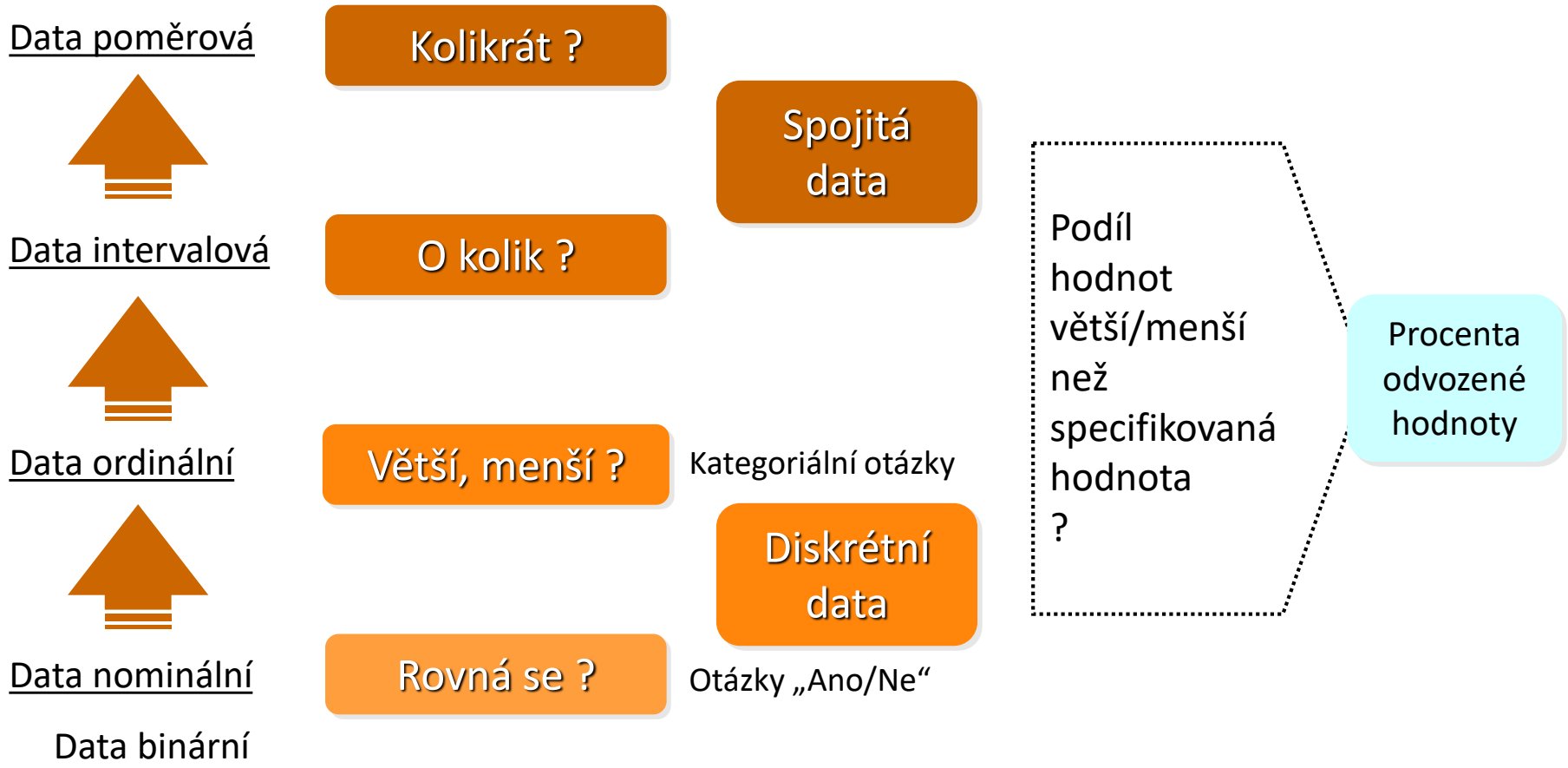
… **kterou chceme dále studovat** → smysluplnost?

(koncentrace polutantu x nadmořská výška, krevní tlak, glykémie × počet srdcí, počet domů)

… více či méně dokonalým → kvalita?

(variabilita = informace + chyba)

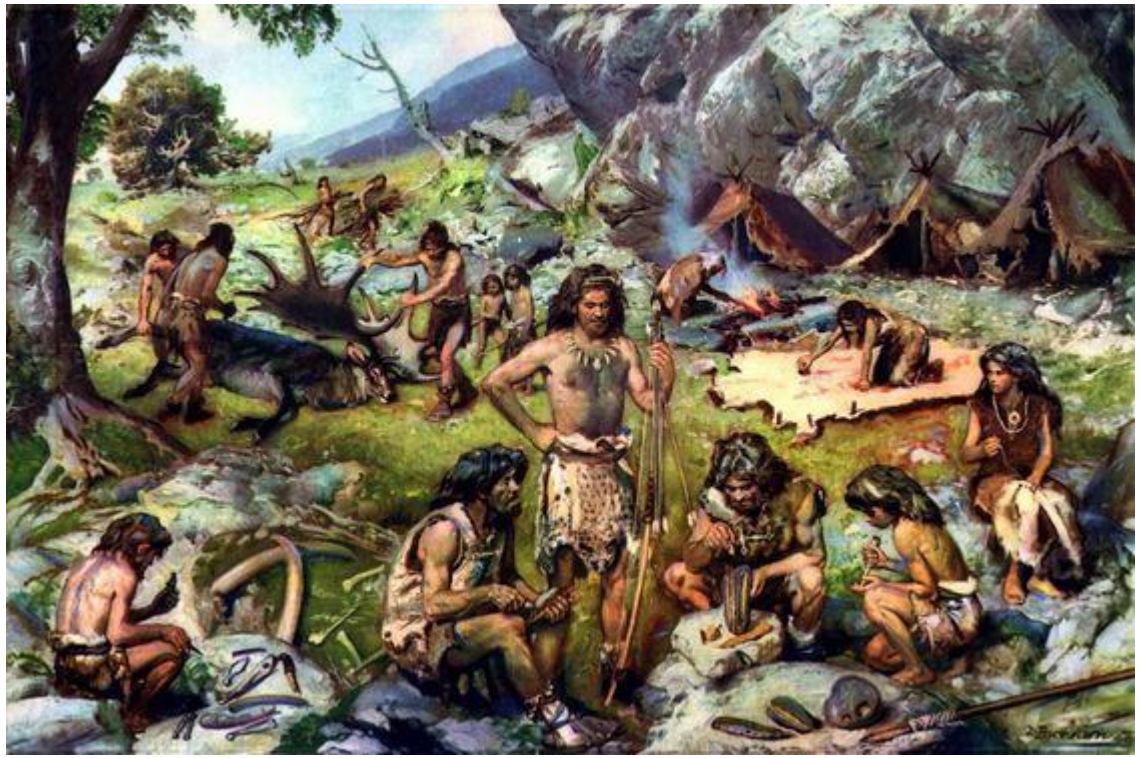
Jak vznikají informace - různé typy dat znamenají různou informaci



Samotná znalost typu dat ale na dosažení informace nestačí …….

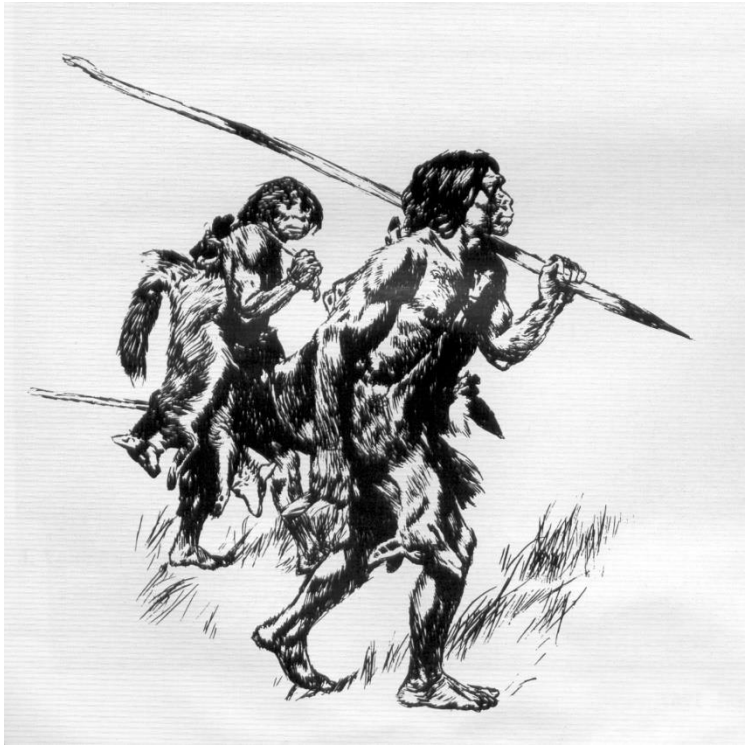
Typy dat a jejich informační hodnota

- Statistika je užitečná v každé době ☺
- I v době ledové ……
- Šaman sedí před jeskyní a přemýšlí:
  - Zima se blíží a je třeba udělat zásoby na zimu
  - Ale musím vymyslet jak **správně** popsat co jsme vlastně ulovili za zásoby
  - Nebo pomřeme hlady ……



Cílová populace

- Vzorkujeme 3 kategorie sledované proměnné kořist



Kořist

Veverka *Jelen* *Mamut*

Binární data – chytili jsme něco?

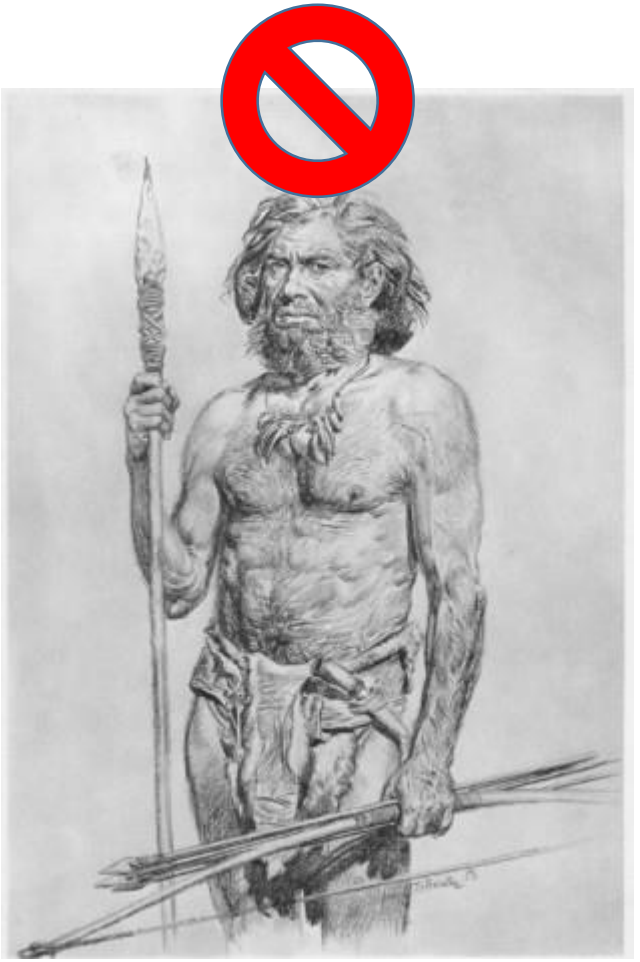
- Informačně nejméně obsáhlá jsou data binární

Hodnotíme dva možné stavy:

Přinesl x nepřinesl kořist

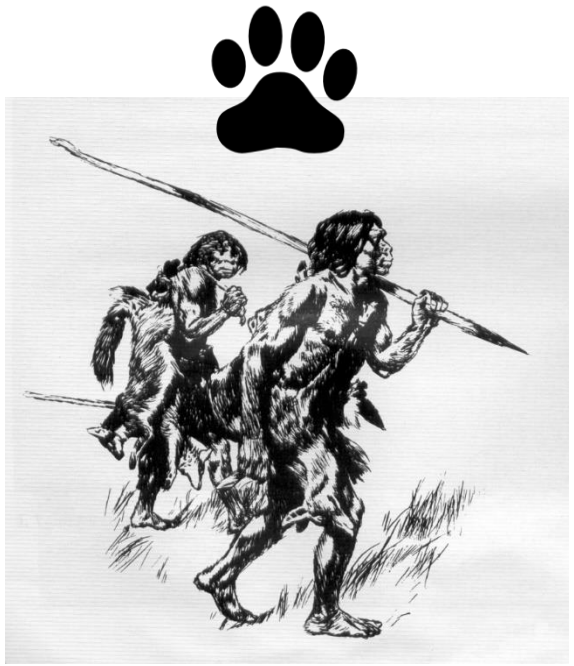
Jak můžeme popsat:

?



Binární data – chytili jsme něco?

- Informačně nejméně obsáhlá jsou data binární

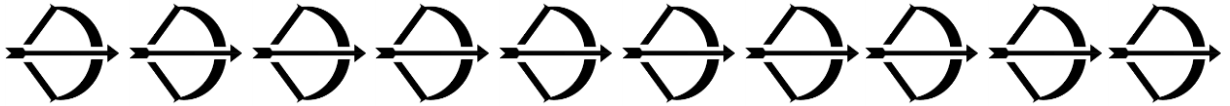


Hodnotíme dva možné stavy:

Přinesl x nepřinesl kořist

Jak můžeme popsat:

Celkový počet lovů (báze hodnocení)



N=10

Počet úlovků (absolutní četnost)

N=7

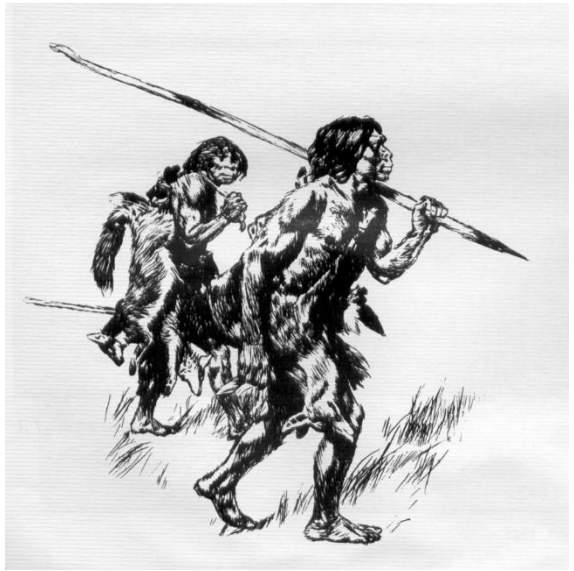
Podíl úspěšných lovů (relativní četnost) nebo nejčetnější kategorie (modus)

N = 7 (70%)

Jsou binární data dostatečná za všech okolností?

Kategoriální data – co jsme chytili?

- Více informací získáme z dat kategoriálních



Hodnotíme několik možných stavů:

Jak můžeme popsat:

Celkový počet lovů (báze hodnocení)

N = 4 (40%)

Počet různých kategorií úlovků (absolutní četnost)

N = 1 (10%)

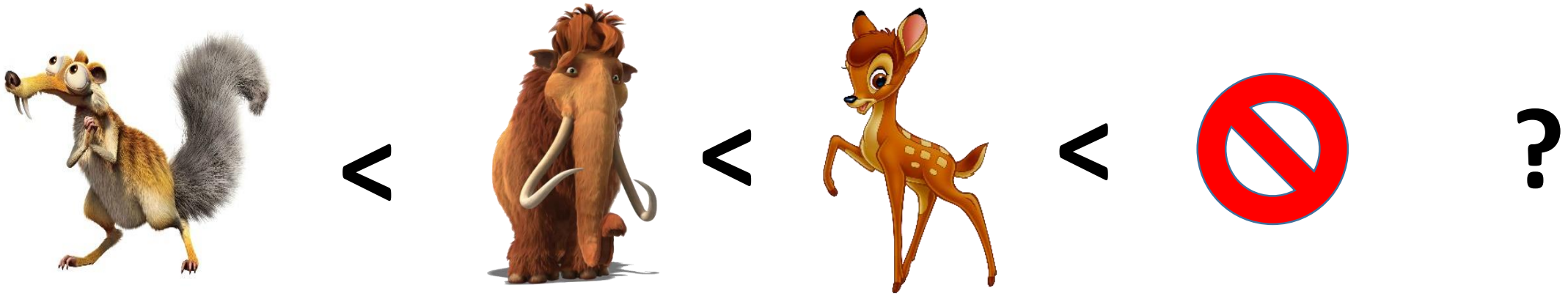
Podíl úspěšných lovů různých kategorií úlovků (relativní četnost) nebo nejčetnější kategorie (modus)

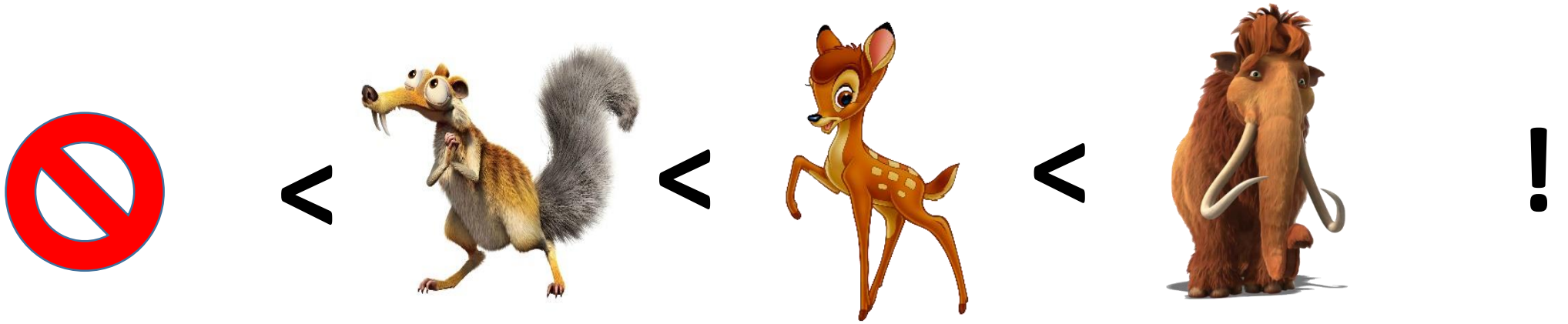
N = 2 (20%)

N = 3 (30%)

Jsou kategoriální data dostatečná za všech okolností?

Jsou kategorie seřaditelné?





- Seřaditelné kategorie = ordinální data
- Ordinální data je možné popsat stejně jako data kategoriální + u seřiditelných dat je možné počítat i medián

Jsou kategoriální data dostatečná za všech okolností?

Pozor na medián u ordinálních dat

- Je medián vždy vhodným ukazatelem středu ordinálních dat?

Medián?

Vs.

Medián?

Pozor na medián u ordinálních dat

Vs.

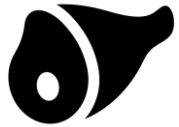
Medián je shodný !

- Medián je shodný, nicméně interpretace dat je odlišná
- Možnost a formální správnost výpočtu statistiky neznamená, že jde o vhodnou metodu.

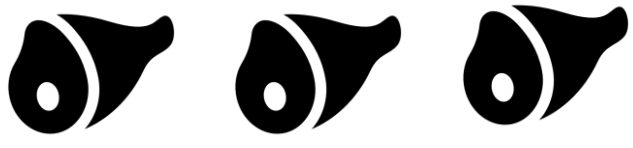
Kvantitativní data – jaký je objem kořisti ?

- Informačně nejhodnotnější jsou data kvantitativní
- Pro popis je nezbytné posoudit jejich rozložení
  - Průměr
  - Medián
  - Směrodatná odchylka
  - Minimum, maximum
  - Percentily
  - Atd.

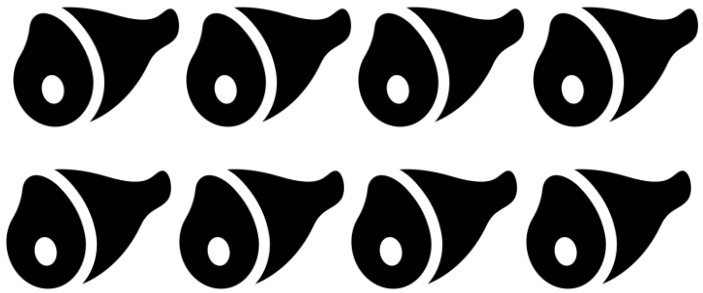
=



=



=



Typy dat: shrnutí

- Kvalitativní proměnná (kategoriální) – lze ji řadit do kategorií, ale nelze ji kvantifikovat, resp. nemá smysl přiřadit jednotlivým kategoriím číselné vyjádření.
- Příklady: pohlaví, HIV status, užívání drog, barva vlasů

- Kvantitativní proměnná (numerická) – můžeme jí přiřadit číselnou hodnotu.
  Rozlišujeme dva typy kvantitativních proměnných:
  - Spojité: může nabývat jakýchkoliv hodnot v určitém rozmezí.
    - Příklady: výška, váha, vzdálenost, čas, teplota.
  - Diskrétní: může nabývat pouze spočetně mnoha hodnot.
    - Příklady: počet krevních buněk, počet hospitalizací, počet krvácivých epizod za rok, počet dětí v rodině.

Kvalitativní data lze dělit dále

- Binární data – pouze dvě kategorie typu ano / ne.

- Nominální data – více kategorií, které nelze vzájemně seřadit.
  - Nemá smysl ptát se na relaci větší/menší.

- Ordinální data – více kategorií, které lze vzájemně seřadit.
  - Má smysl ptát se na relaci větší/menší.

Kvalitativní data – příklady

- Binární data
  - diabetes (ano/ne)
  - pohlaví (muž/žena)

- Nominální data
  - krevní skupiny (A/B/AB/0)
  - stát EU (Belgie/…/Česká republika/…/Velká Británie)

- Ordinální data
  - stupeň bolesti (mírná/střední/velká/nesnesitelná)
  - spotřeba cigaret (nekuřák/ex-kuřák/občasný kuřák/pravidelný kuřák)
  - stadium maligního onemocnění (I/II/III/IV)

Jak vznikají informace – popis různých typů dat

Statistika středu

Data poměrová

Data intervalová

Data ordinální

Data nominální

Data binární

PRŮMĚR

MEDIÁN

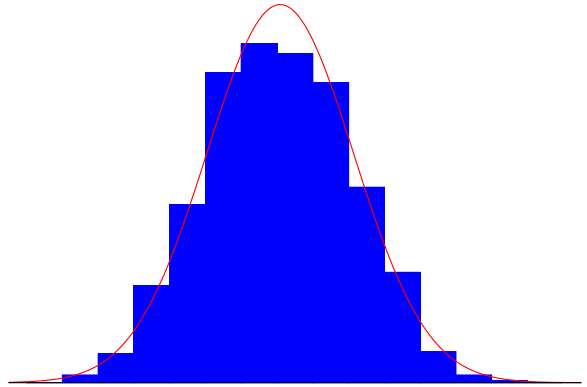
MODUS

Absolutní a relativní četnosti

Spojitá data

Diskrétní data

- Kvantitativní data - četnost hodnot rozložení v jednotlivých intervalech.

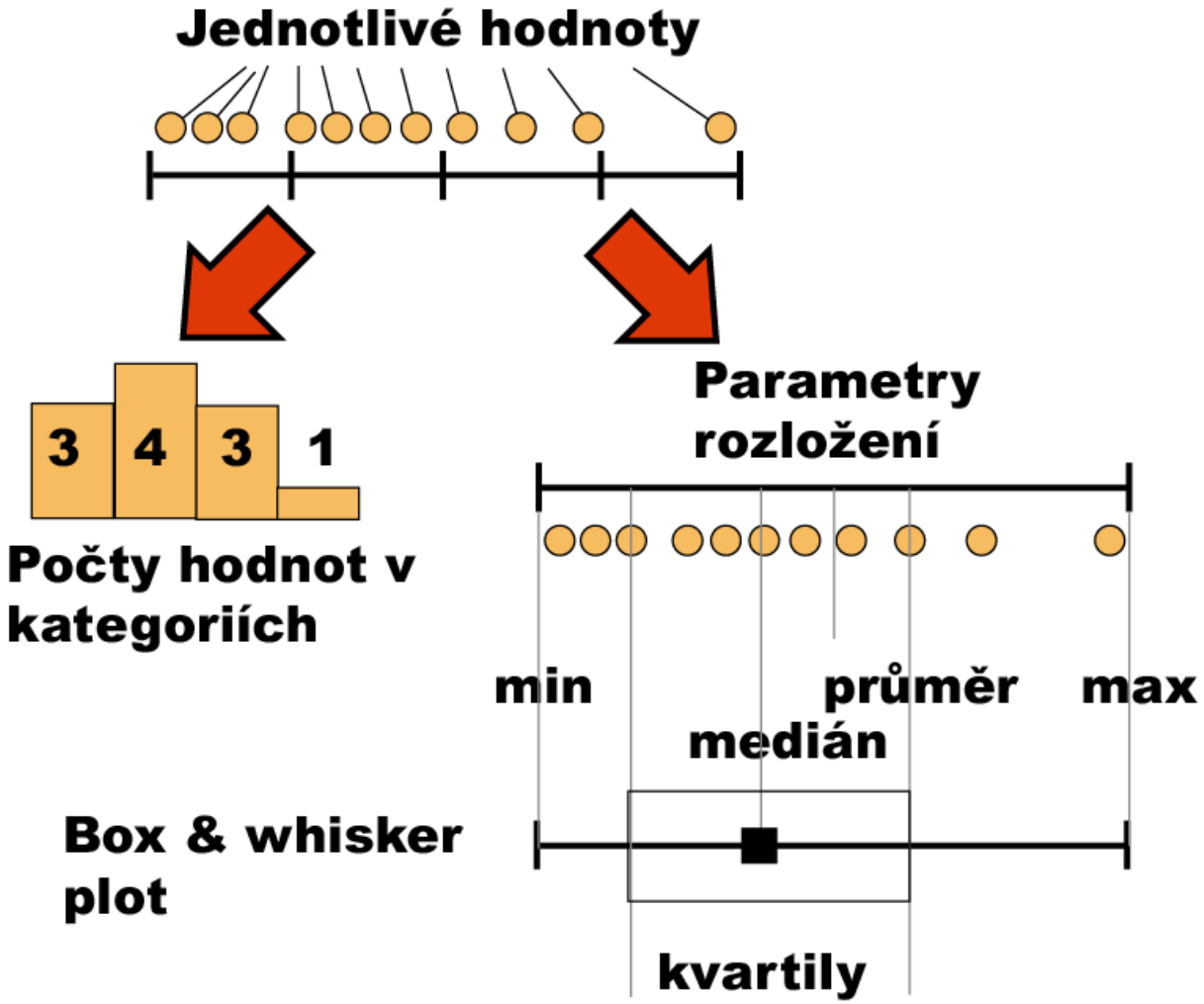


- Kvalitativní data - tabulka s četností jednotlivých kategorií.

Kategorie	Četnost
B	5
C	8
D	1

Řada dat a její vlastnosti

- V analýze je často možné zvolit několik možných cest popisu dat
- Kritériem výběru není pouze formální matematická správnost, ale také smysluplnost a informační hodnota použité popisné statistiky v dané situaci



Odvozená data: pozor na odvozené indexy

- X: Průměrný počet výrobků v prodejně
- Y: Odhad prostoru průměrně nabízeného k vystavení výrobku

- Popsáno průměrem a rozsahem min-max
  - X: 1,2 : (1,15 - 1,24) → + / - 3,8 %
  - Y: 1,8 : (1,75 - 1,84) → + / - 2,5 %

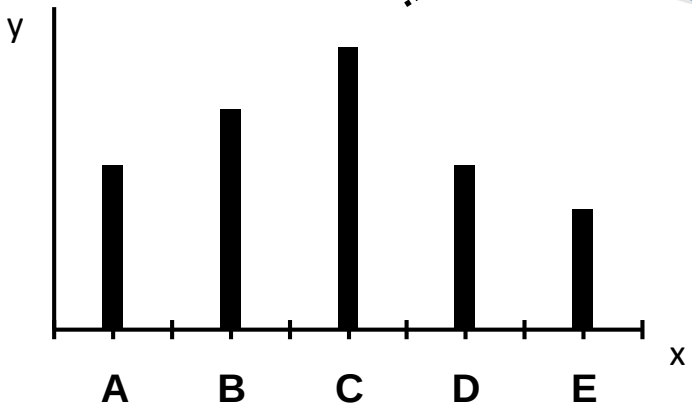
  - $\frac{X}{Y} = 0{,}667 : \left(\frac{1{,}15}{1{,}84} - \frac{1{,}24}{1{,}75}\right)$ → + / - 6,2 %

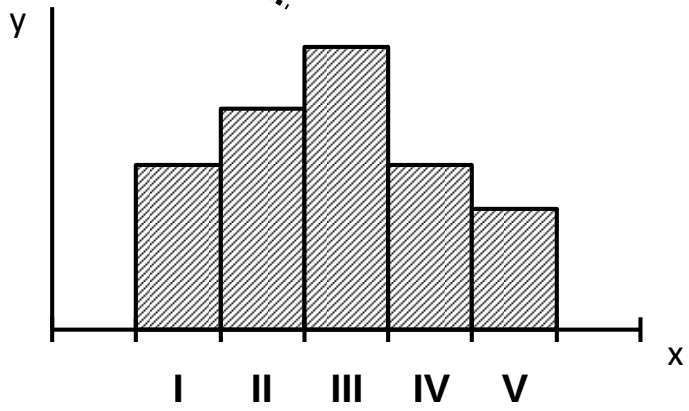
- Nová veličina má jinou šířku rozpětí než ty, ze kterých je odvozená

Vznik informací: opakovaná měření informují rozložením hodnot

Y: frekvence -
absolutní / relativní

KOLIK se naměřilo





CO se naměřilo

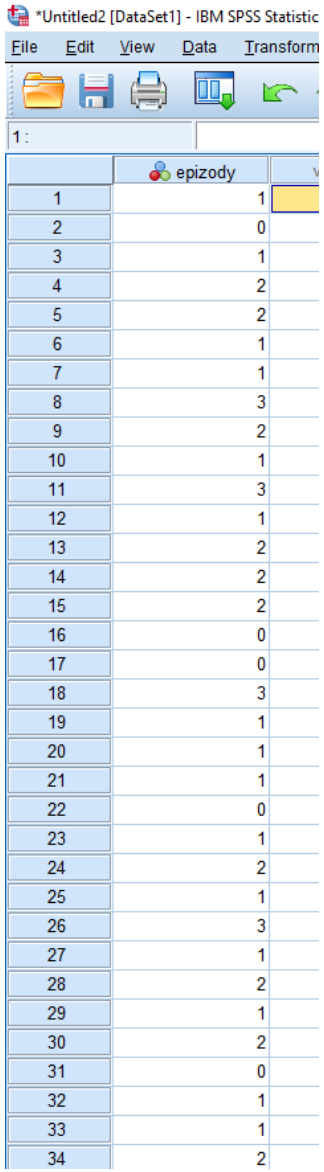
X: měřený znak

Diskrétní data

Spojitá data

Frekvenční sumarizace - základní nástroj popisu dat: kvalitativní data

- Cílem sumarizace je zjednodušení dat do přehledné formy
- N = 100 pacientů s hemofilií
- Hodnocenou proměnnou je počet krvácivých epizod za měsíc
- Nejjednodušší sumarizací je frekvenční tabulka



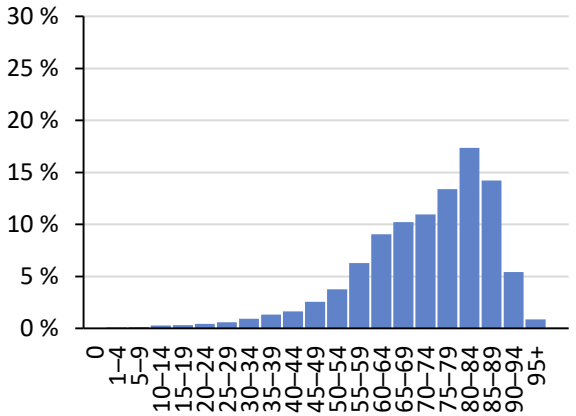
epizody

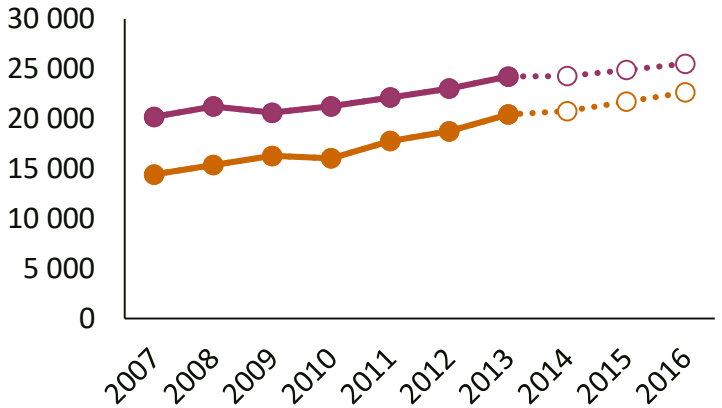
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	22	22,0	22,0	22,0
	1	27	27,0	27,0	49,0
	2	29	29,0	29,0	78,0
	3	22	22,0	22,0	100,0
	Total	100	100,0	100,0	

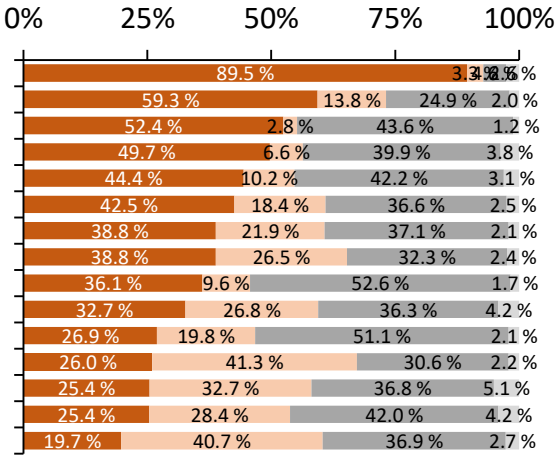
- Tabulka ukazuje unikátní hodnoty v datech
- Frequency = počet hodnot v kategorii (absolutní četnost)
- Percent = procentuální zastoupení kategorie (relativní četnost)
- Valid percent = procentuální zastoupení kategorie (bez započtení chybějících hodnot)
- Cumulative percent = kumulativní procentuální zastoupení kategorií až po danou kategorii (kumulativní relativní četnost; má smysl pouze pro ordinální data, obdobně existuje i kumulativní absolutní četnost)

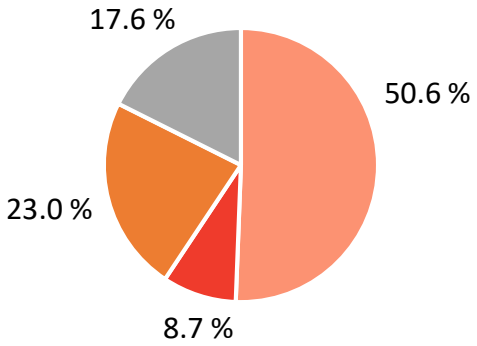
Vizualizace frekvenční tabulky kvalitativních dat

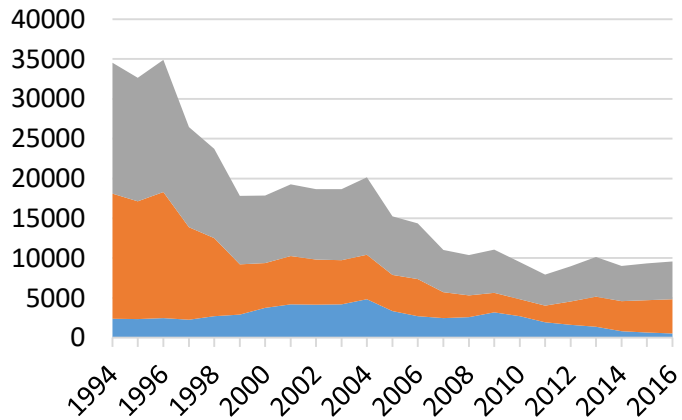
- Libovolné grafy umožňující vizualizaci počtů a procent (koláčový, páskový, sloupcový, čárový)

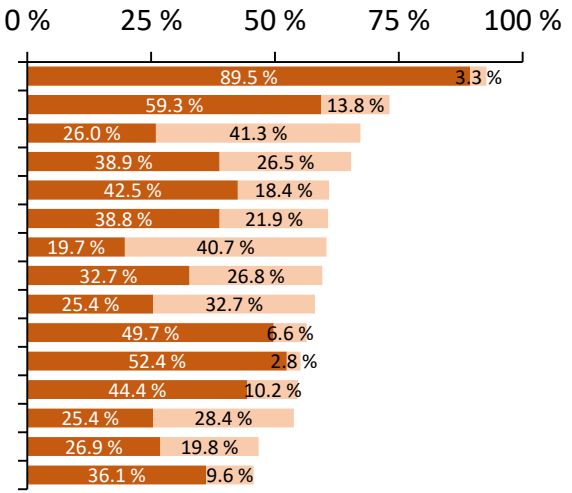






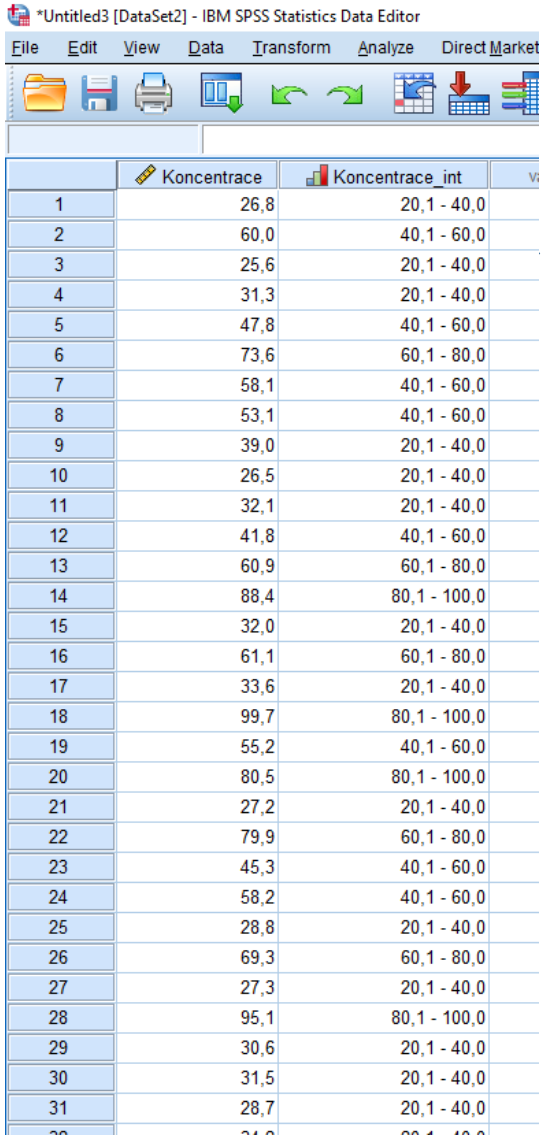






Frekvenční sumarizace - základní nástroj popisu dat: kvantitativní data

- Cílem sumarizace je zjednodušení dat do přehledné formy
- N = 100 pacientů s
- Hodnocenou proměnnou je koncentrace látky v krvi
- Nejjednodušší sumarizací je opět frekvenční tabulka
- Další možností je výpočet zástupných sumárních statistik (průměr, medián aj.)



	Koncentrace	Koncentrace_int
1	26,8	20,1 - 40,0
2	60,0	40,1 - 60,0
3	25,6	20,1 - 40,0
4	31,3	20,1 - 40,0
5	47,8	40,1 - 60,0
6	73,6	60,1 - 80,0
7	58,1	40,1 - 60,0
8	53,1	40,1 - 60,0
9	39,0	20,1 - 40,0
10	26,5	20,1 - 40,0
11	32,1	20,1 - 40,0
12	41,8	40,1 - 60,0
13	60,9	60,1 - 80,0
14	88,4	80,1 - 100,0
15	32,0	20,1 - 40,0
16	61,1	60,1 - 80,0
17	33,6	20,1 - 40,0
18	99,7	80,1 - 100,0
19	55,2	40,1 - 60,0
20	80,5	80,1 - 100,0
21	27,2	20,1 - 40,0
22	79,9	60,1 - 80,0
23	45,3	40,1 - 60,0
24	58,2	40,1 - 60,0
25	28,8	20,1 - 40,0
26	69,3	60,1 - 80,0
27	27,3	20,1 - 40,0
28	95,1	80,1 - 100,0
29	30,6	20,1 - 40,0
30	31,5	20,1 - 40,0
31	28,7	20,1 - 40,0

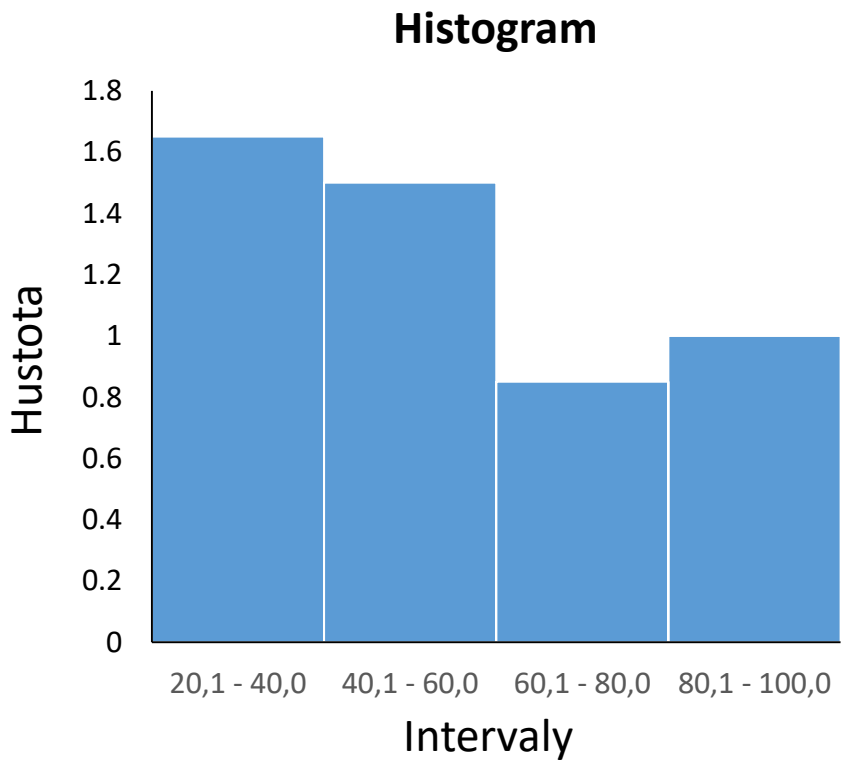
Koncentrace intervaly

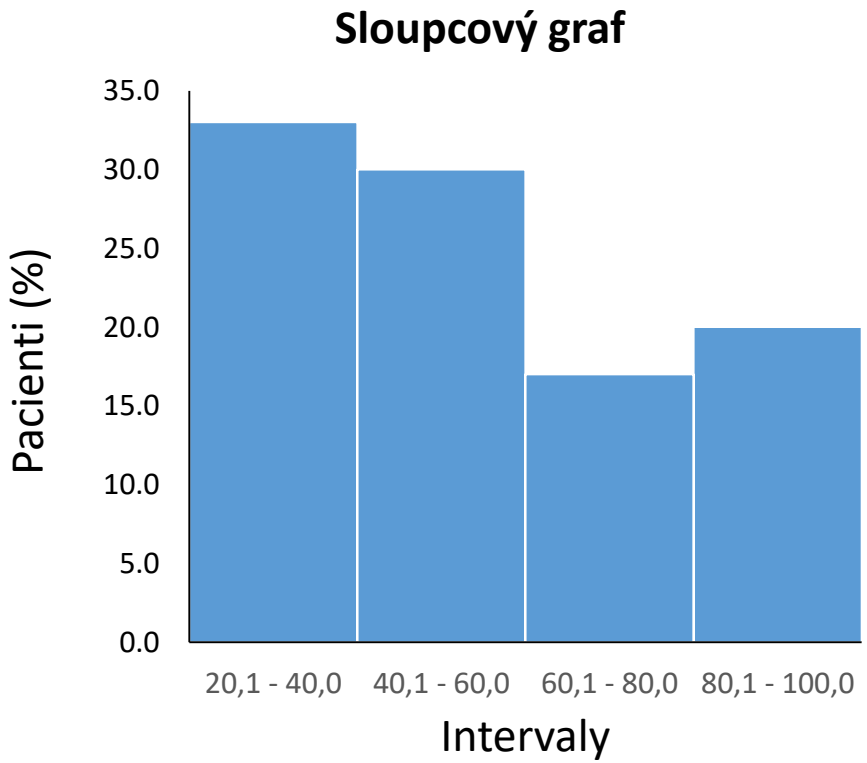
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	20,1 - 40,0	33	33,0	33,0	33,0
	40,1 - 60,0	30	30,0	30,0	63,0
	60,1 - 80,0	17	17,0	17,0	80,0
	80,1 - 100,0	20	20,0	20,0	100,0
	Total	100	100,0	100,0	

- Tabulka ukazuje unikátní hodnoty v datech
- Na rozdíl od kvalitativních dat je nezbytné pro smysluplnost výstupu stanovit v datech intervaly (o stejné nebo různé šířce)
- Frequency = počet hodnot v kategorii (absolutní četnost)
- Percent = procentuální zastoupení kategorie (relativní četnost)
- Valid percent = procentuální zastoupení kategorie (bez započtení chybějících hodnot)
- Cumulative percent = kumulativní procentuální zastoupení kategorií až po danou kategorii (kumulativní relativní četnost; obdobně existuje i kumulativní absolutní četnost)

Vizualizace frekvenční tabulky kvantitativních dat

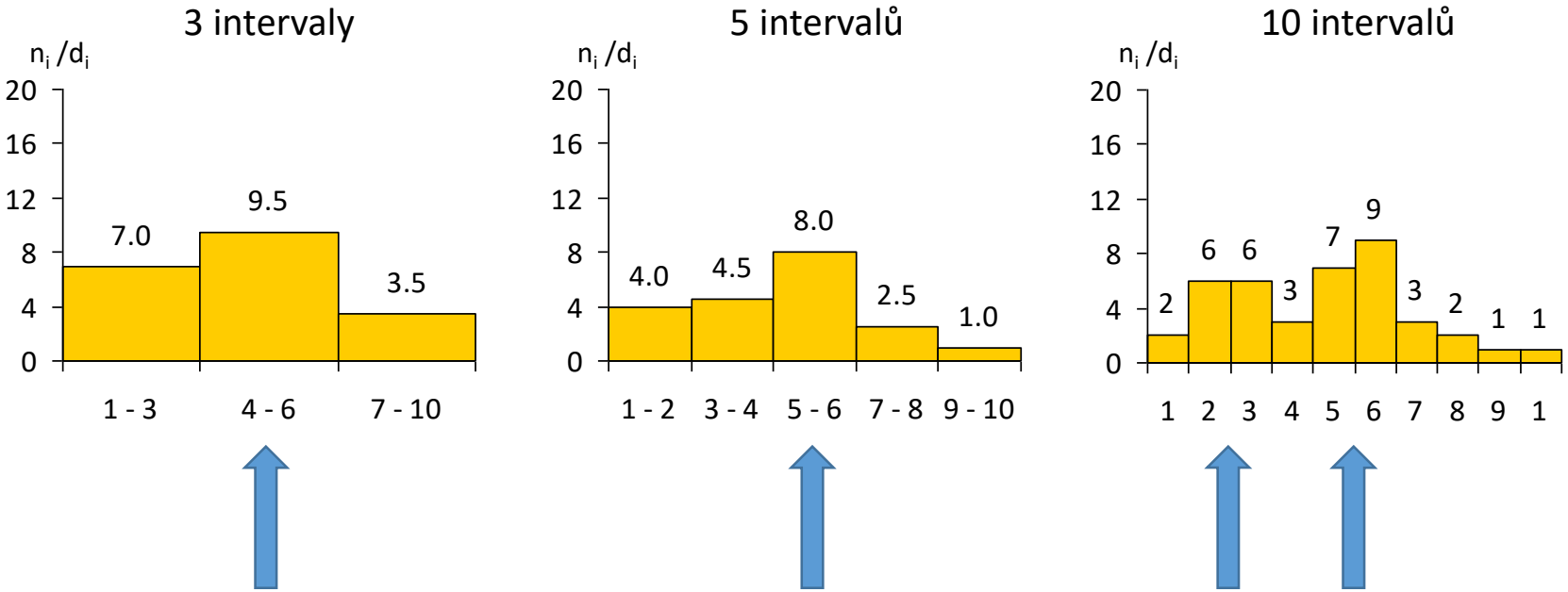
- Základním nástrojem vizualizace spojitých dat založeným na frekvenční tabulce je histogram
- Na rozdíl od sloupcového grafu představuje vizualizovanou hodnotu plocha sloupce, nikoliv jeho výška





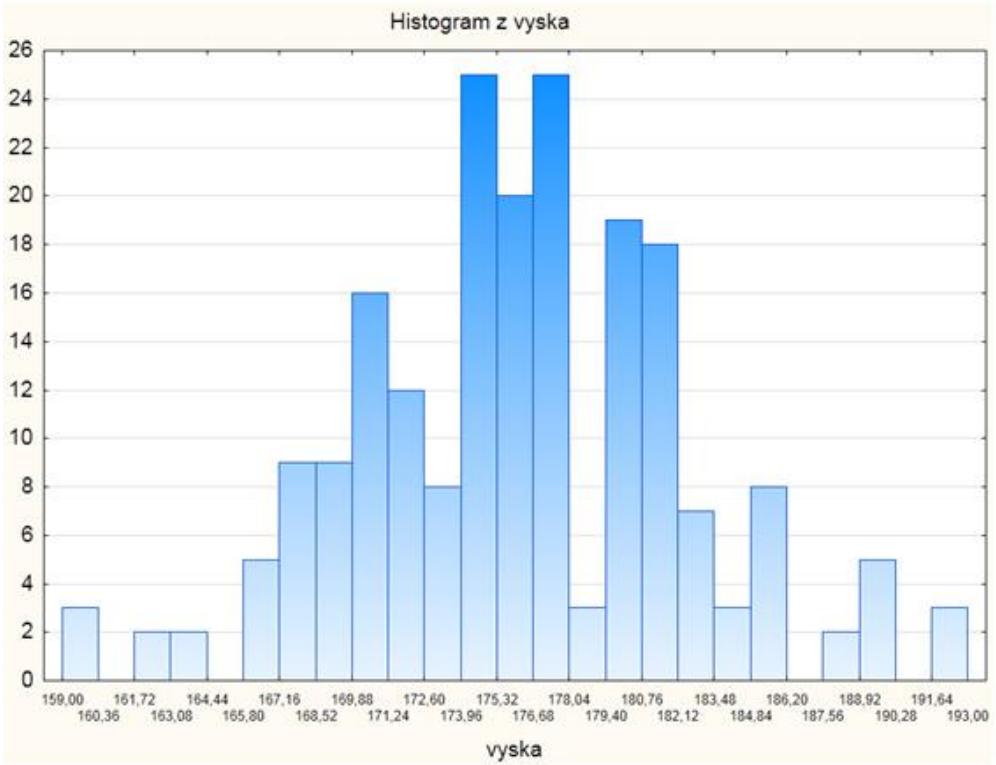
Histogram: vliv kategorizace dat

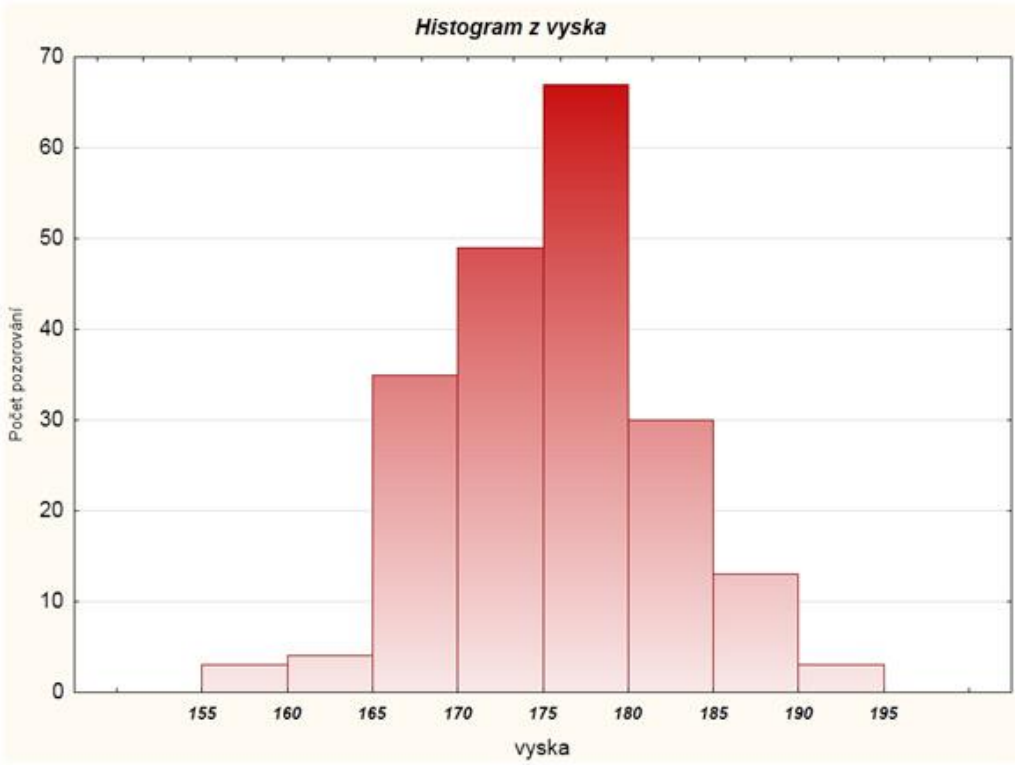
- Počtem zvolených intervalů v histogramu rozhodujeme o tom, jak bude vypadat. Při malém počtu můžeme přehlédnout důležité prvky v datech, při velkém zase může být informace roztříštěná.



Histogram: vliv kategorizace dat

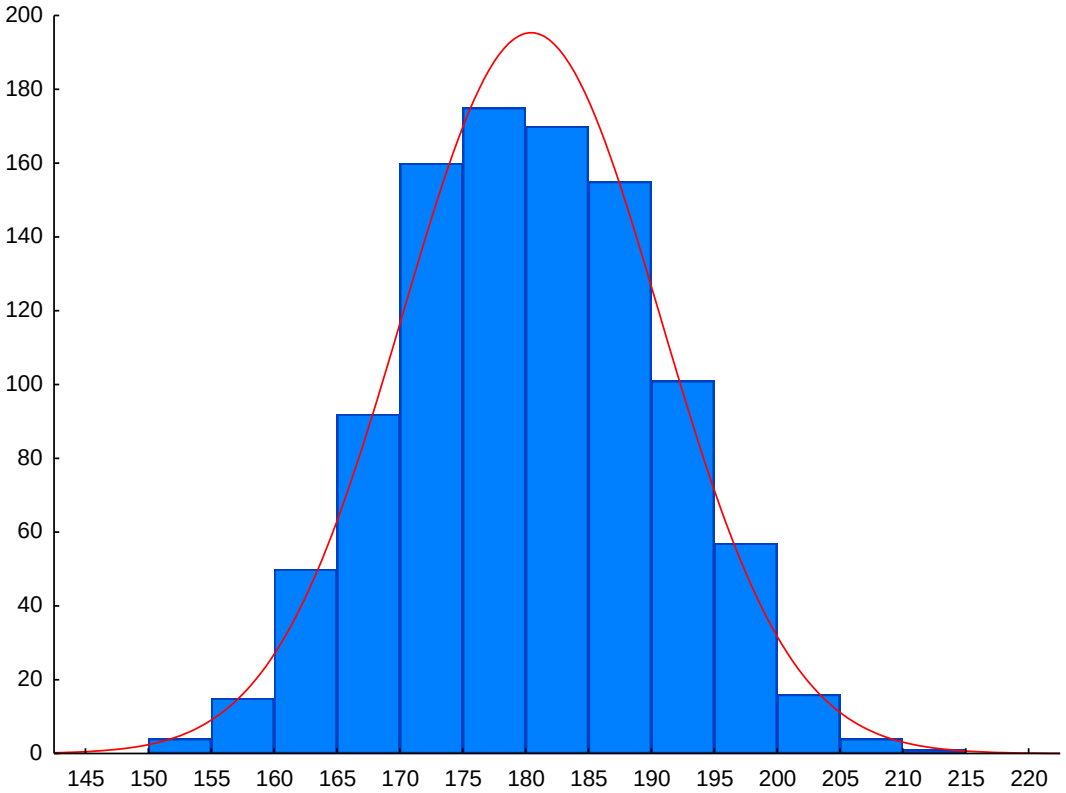
- Výběr počtu kategorií – důležitý pro interpretaci
- Ruční nebo automatický výběr – různé algoritmy (závisí na velikosti vzorku a variabilitě dat)



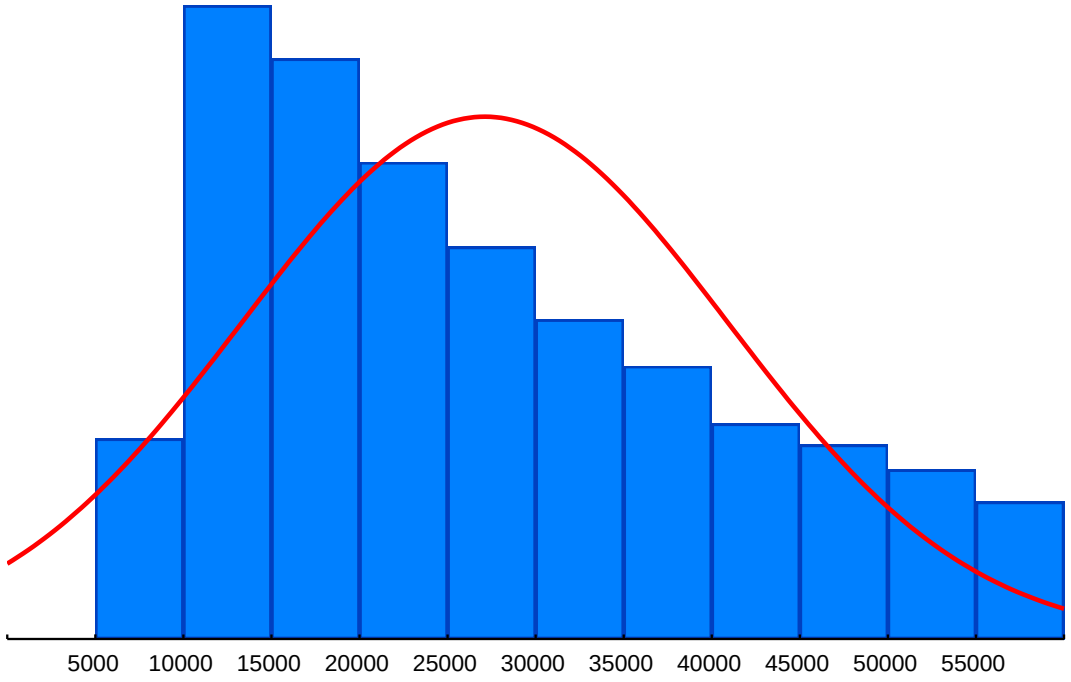


Histogram: nástroj posouzení rozložení dat

- Histogram reálných dat má vazbu na modelové rozdělení



?

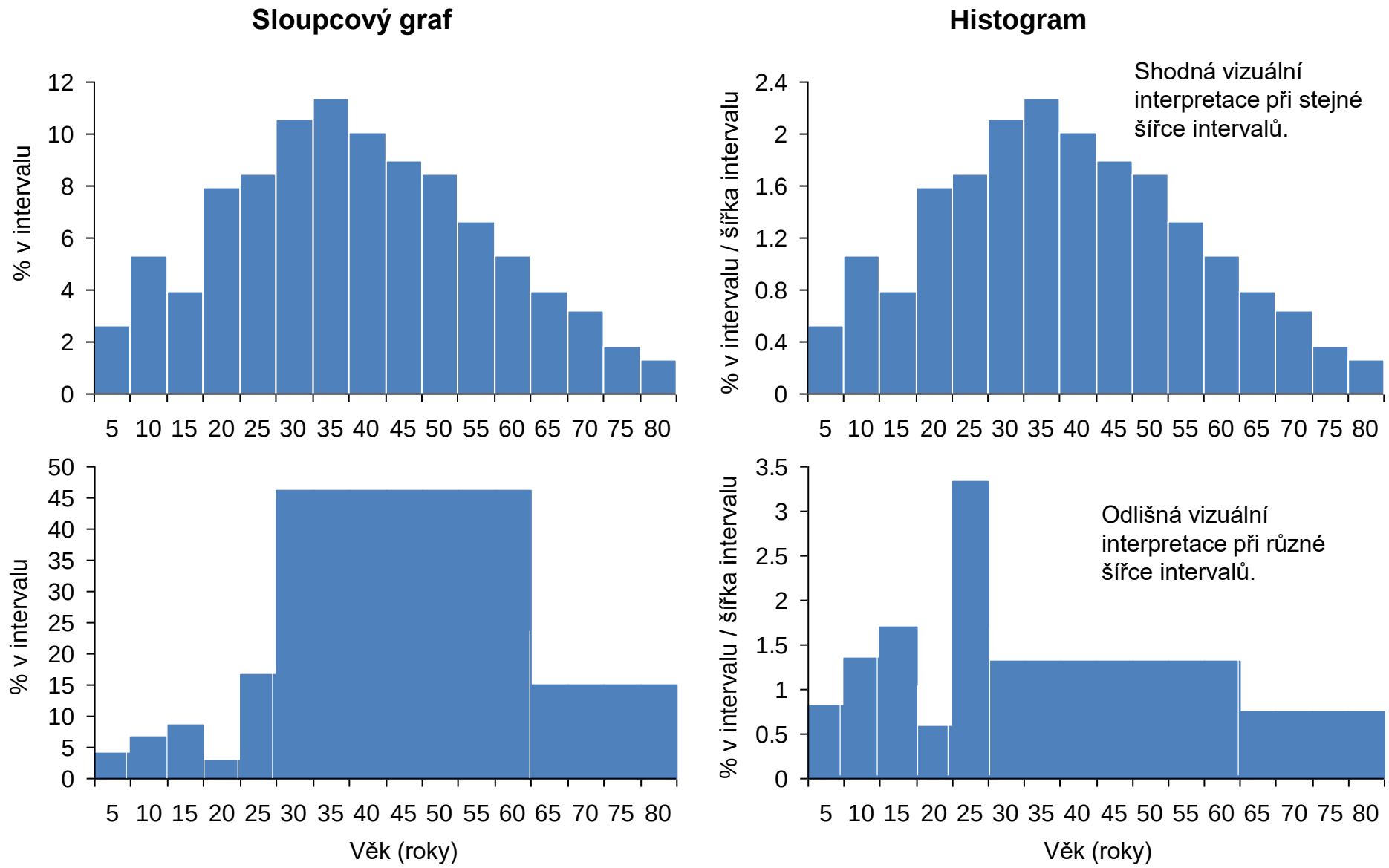


Proč je důležité vědět co je to skutečný histogram I

- Většina lidí uvažuje vizuálně – vizualizace dat je tak nesmírně důležitá pro první vjem a interpretaci dat
- Díky odlišné vizuální interpretaci histogramu a sloupcového grafu v případě použití různě širokých intervalů může být za některé situace použití sloupcového grafu zavádějící

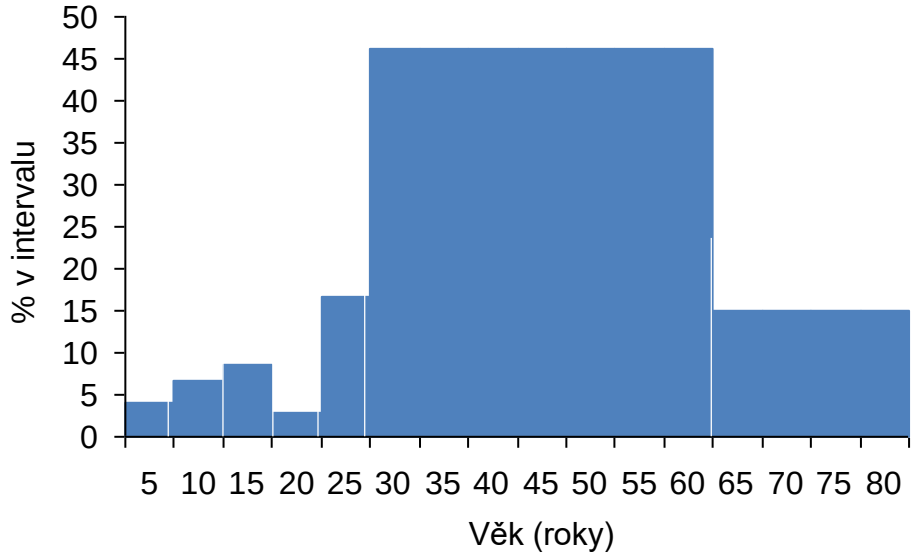
- V praxi se nicméně často používá namísto „pravého“ histogramu sloupcový graf (i výrobci statistických SW)
- V případě stejné šířky intervalů interpretační problém nevzniká (při různé šířce intervalu vypínají SW některé volby = nastavení pro pokročilé uživatele)

Histogram a sloupcový graf

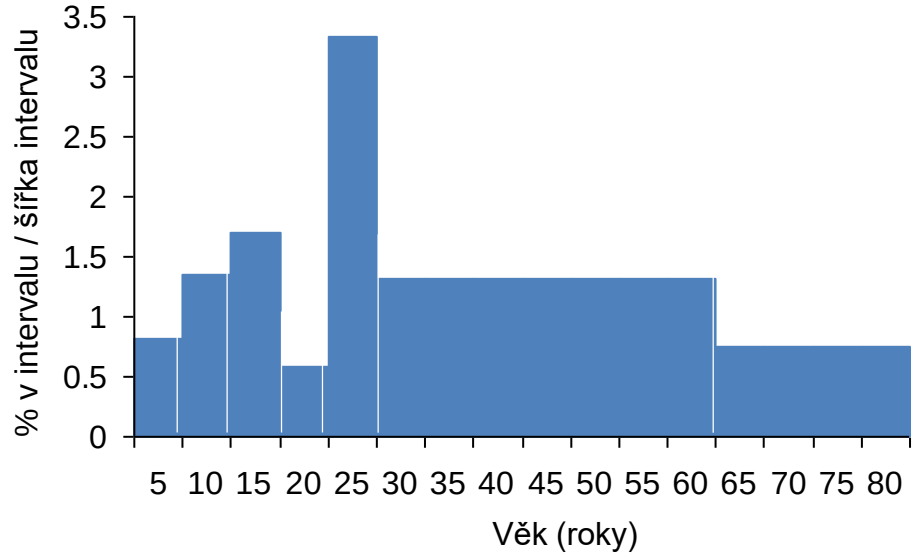


Příklad: věk účastníků vážných dopravních nehod

- Analyzován byl věk účastníků vážných dopravních nehod v jedné londýnské čtvrti
- Liší se interpretace dat vizualizovaných pomocí sloupcového grafu a histogramu?
- Která interpretace Vám přijde smysluplnější a proč?



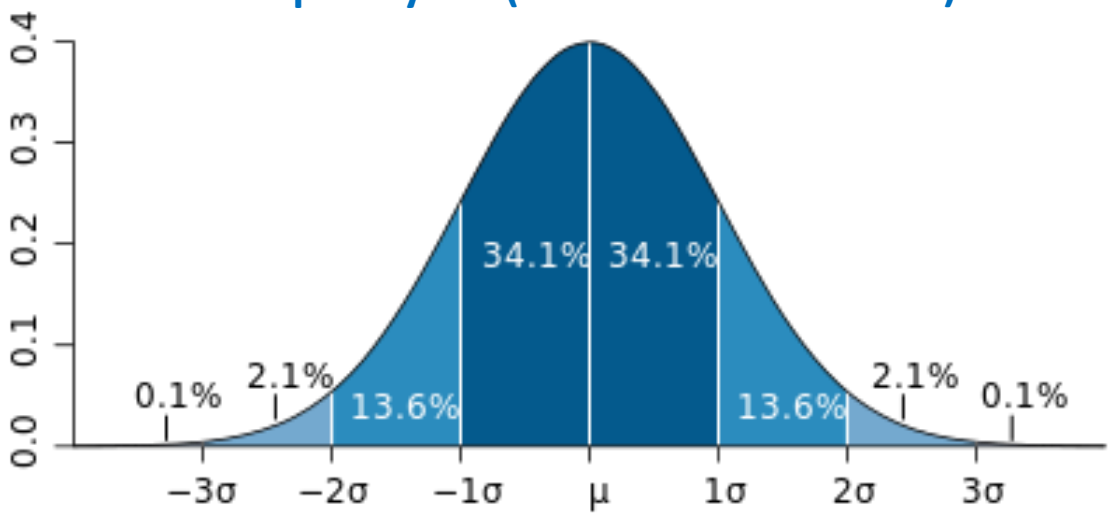
Věk	N	%
0 - 4	28	4,1%
5 - 9	46	6,7%
10-15	58	8,5%
16 - 19	20	2,9%
20 - 24	114	16,6%
25 - 59	316	46,1%
> 60	103	15,0%

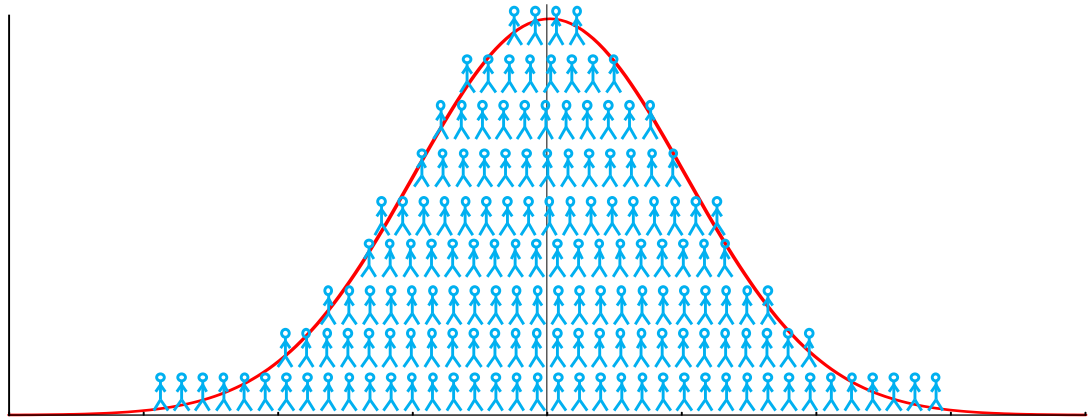


Proč je důležité vědět co je to skutečný histogram II

- Statistické analýzy jsou postaveny na modelových rozděleních, které používáme ve výpočtech jako zástup naměřených dat (pokud reálná data odpovídají svým rozložením modelu, můžeme model využít ve výpočtech místo něj)
- Modely popisují rozdělení hustoty pravděpodobnosti výskytu dané hodnoty = pravděpodobnost výskytu hodnot je dána plochou grafu
- **Rozložení** = reálná data
- **Rozdělení** = model

Plocha = pravděpodobnost výskytu
Suma plochy = 1 (100% všech možností)





Příklad: optimalizace skladových zásob oblečení

- Představte si, že vlastníte obchod s oblečením a chcete optimalizovat skladové zásoby různých velikostí oblečení = potřebujete zjistit kolik % lidí v populaci potřebuje jaké oblečení
- Jaké je rozdělení lidí v populaci co do velikosti?
- Rovnoměrné, normální, lognormální ???

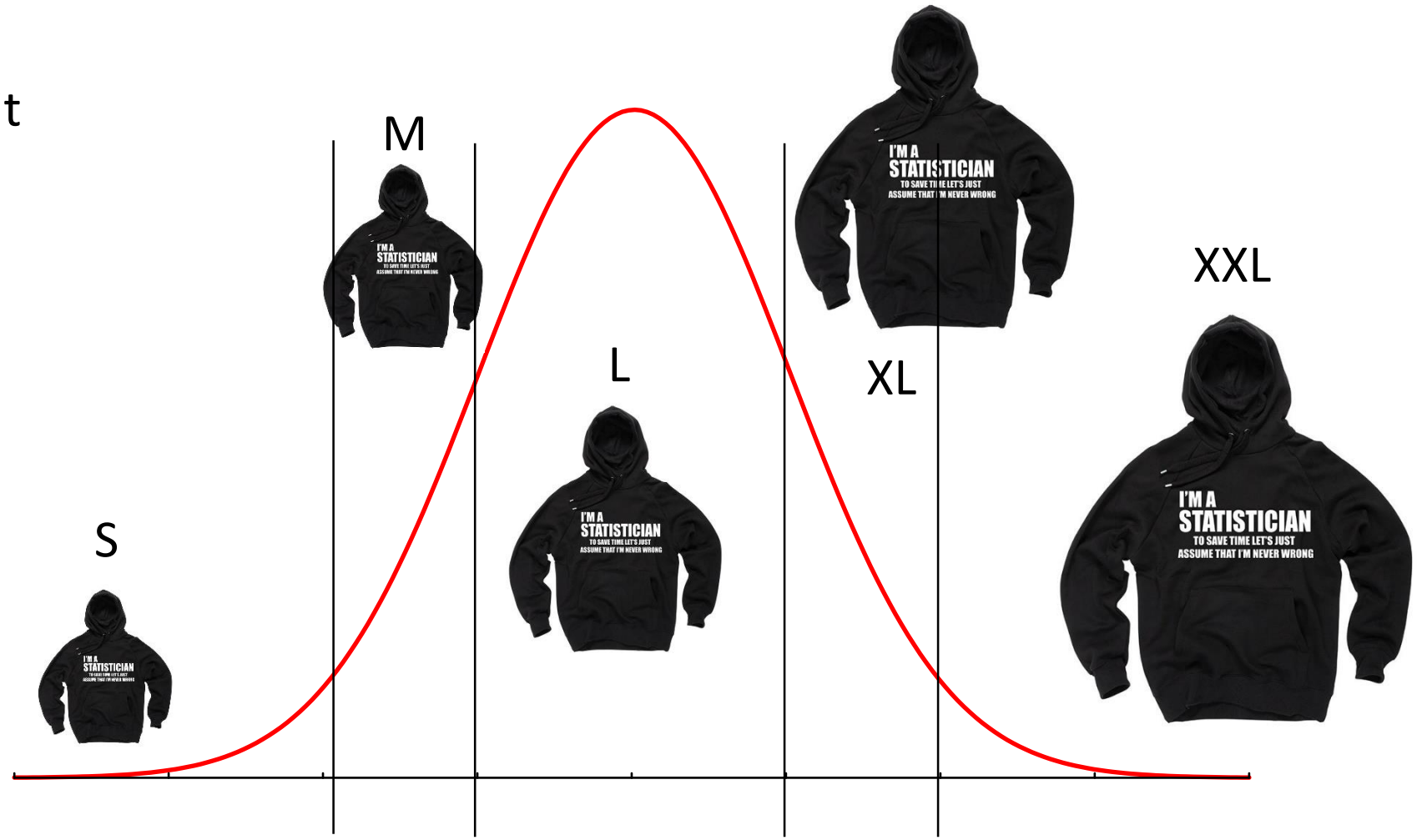
S M L XL XXL

Příklad: optimalizace skladových zásob oblečení

- Dá se předpokládat, že velikost lidí je rozložena normálně
- Pokud jsme schopni stanovit rozsahy hodnot pro různé velikosti oblečení, můžeme podíly skladových zásob odečíst z křivky normálního rozdělení

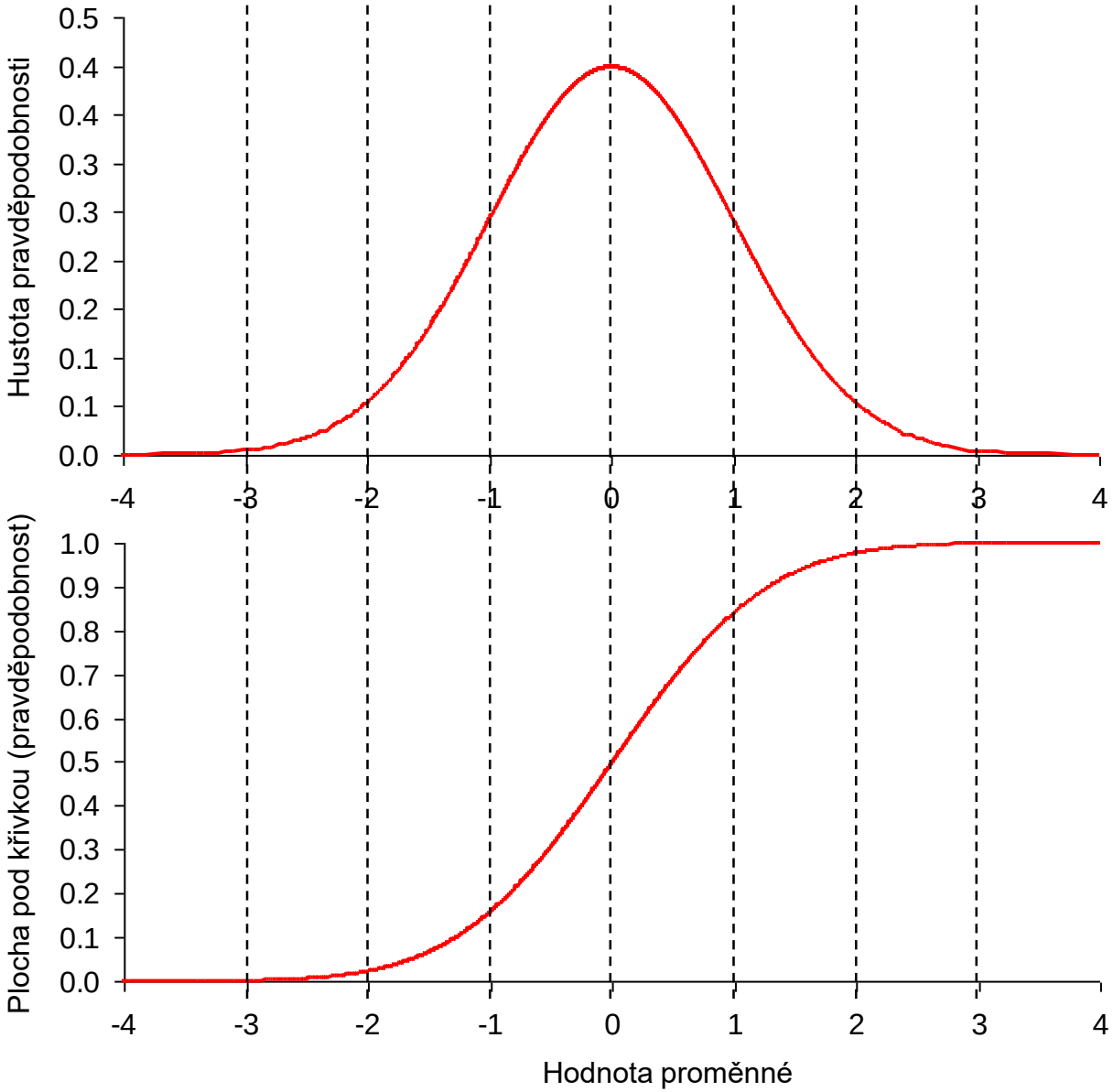
- Integrovat?

- Lze jednodušeji?



Normální rozdělení a jeho distribuční funkce

- K modelovým rozdělením existují jejich distribuční funkce
- Pro danou hodnotu rozdělení uvádějí plochu (=pravděpodobnost) pod křivkou do dané hodnoty
- Základní nástroj v řadě statistických výpočtů
- **Kvantil modelového rozdělení:** hodnota jíž odpovídá daná plocha pod křivkou rozdělení (např. 95% kvantil je hodnota proměnné pod níž leží 95% všech hodnot)

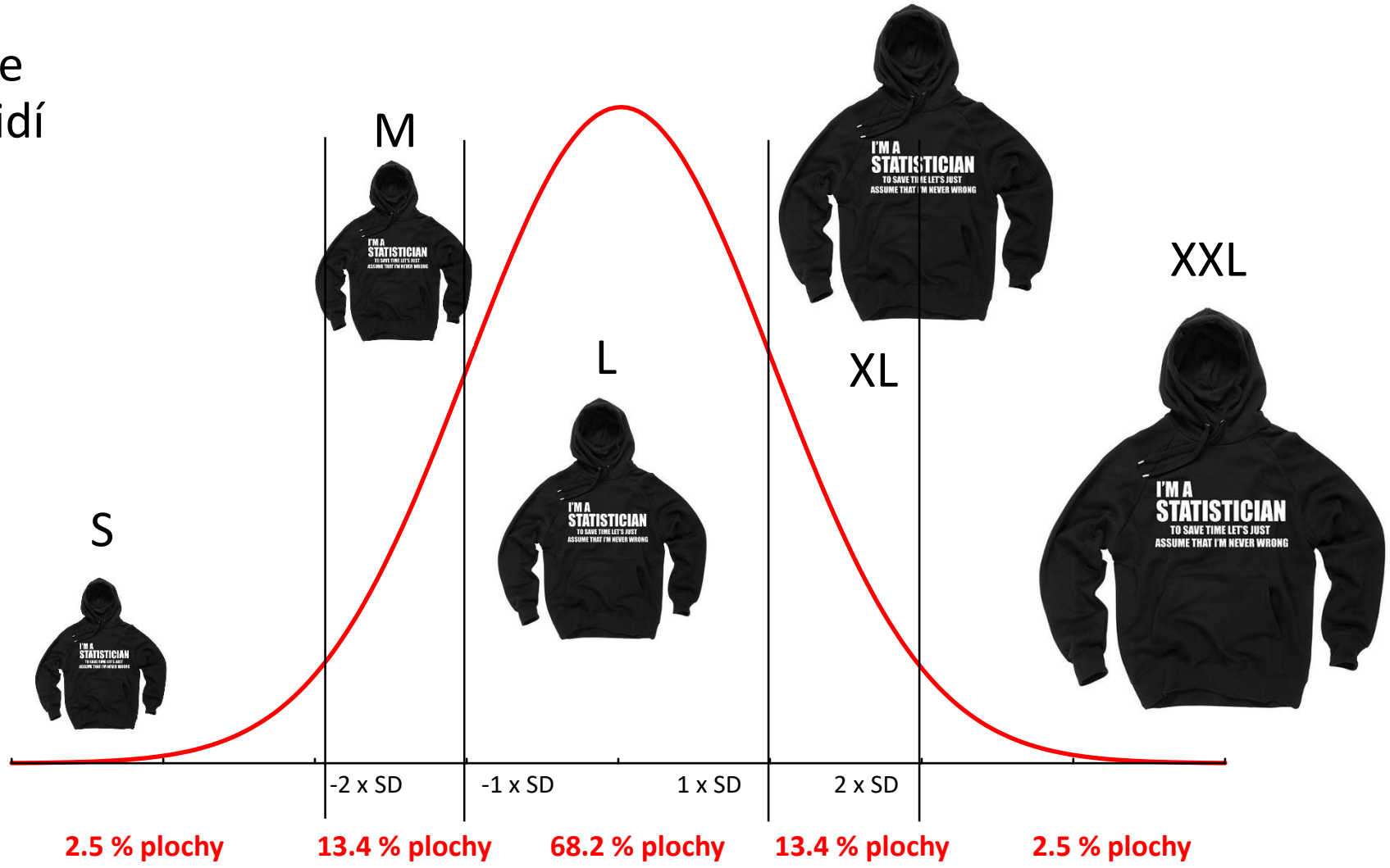


Normální rozdělení

Distribuční funkce normálního rozdělení

Příklad: optimalizace skladových zásob oblečení

- Řešení příkladu odvodíme ze znalosti rozdělení velikosti lidí v cílové populaci a jeho distribuční funkce
- Přibližné podíly různých velikostí oblečení:
  - S: 2.5%
  - M: 13.4%
  - L: 68.2%
  - XL: 13.4%
  - XXL: 2.5%



Velikost člověka relevantní k velikosti oblečení