

Přednáška 1

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

Statistika ve vědecké praxi

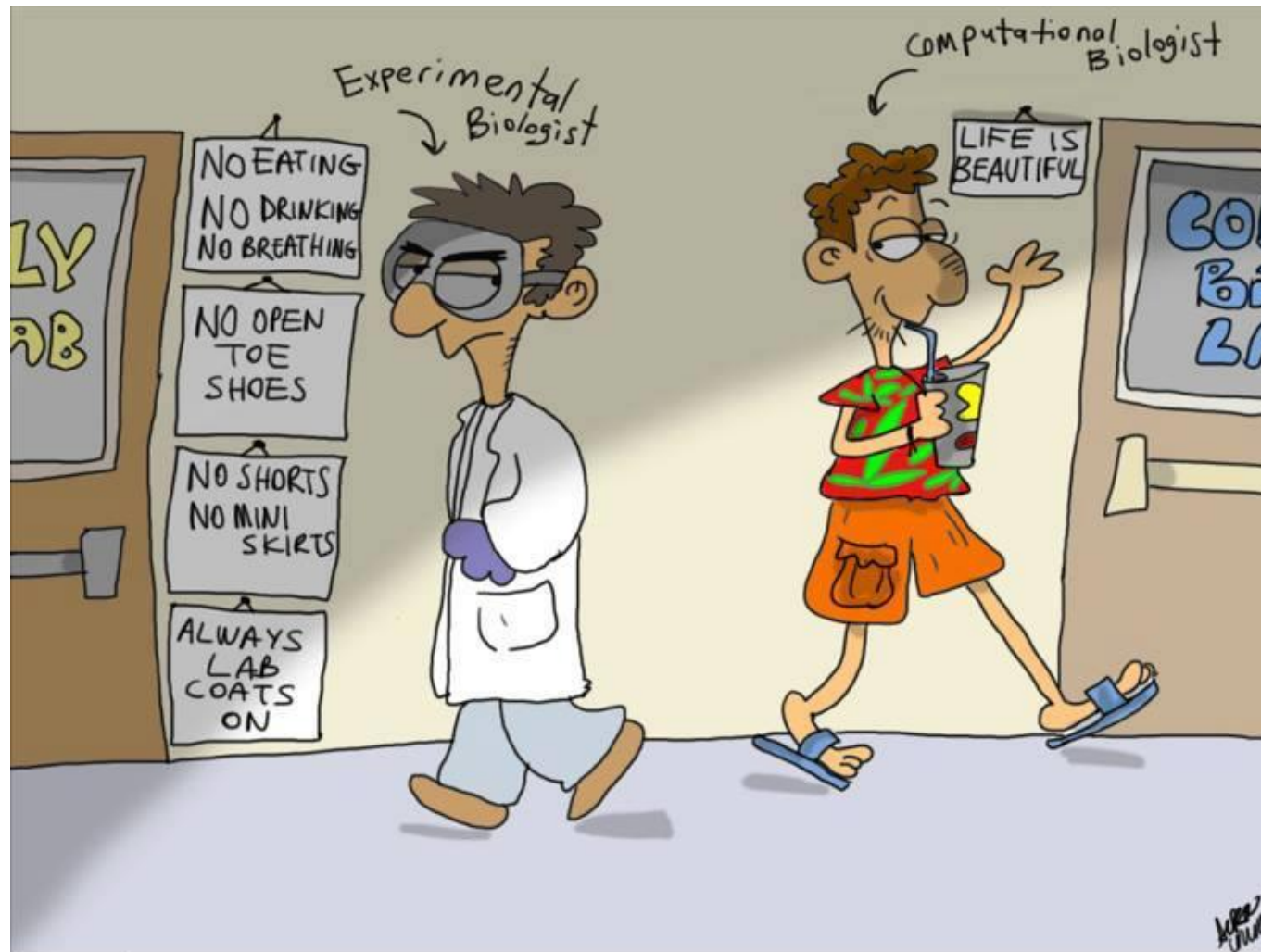
Pozice statistické analýzy ve vědě a klinické praxi

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 rozbořem 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?

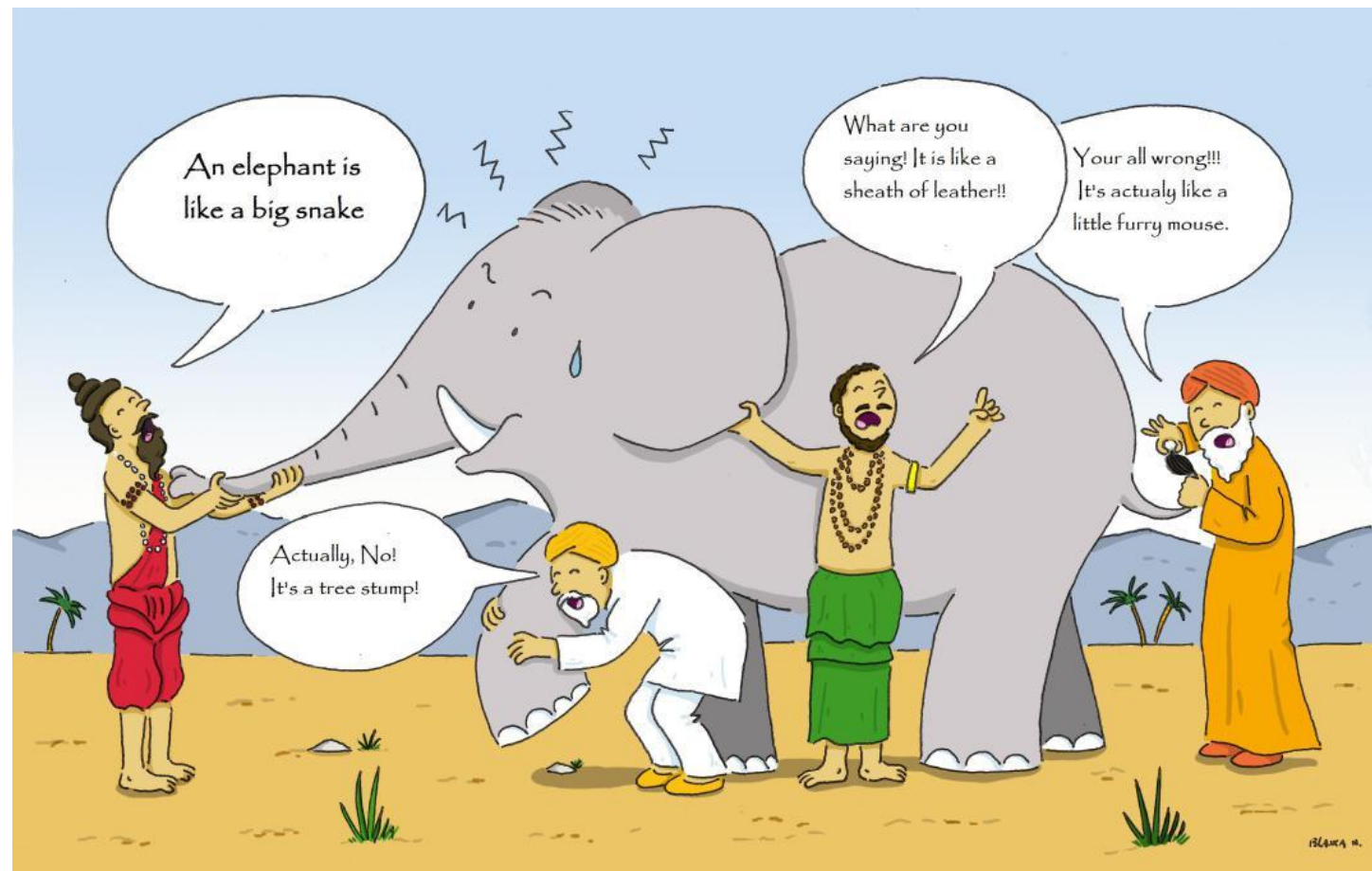


- **Statistika** je jedním z nástrojů umožňujícím popis a komunikaci výsledků výzkumu.
- Ale je to pouze nástroj, co je skutečně důležité jsou **data**.



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

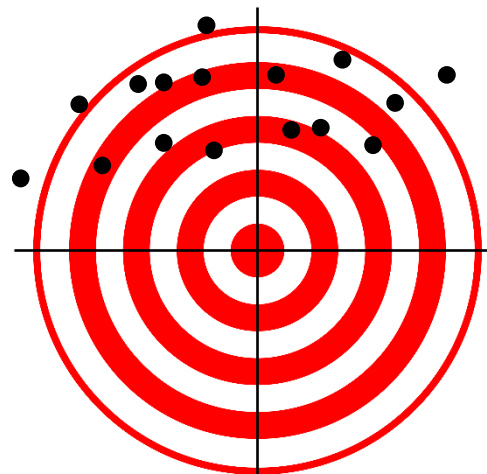


?

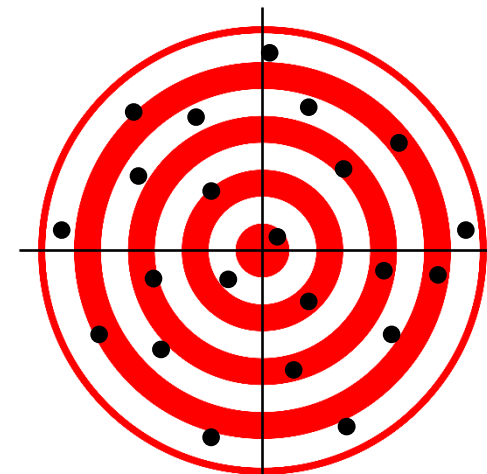


Spolehlivost a přesnost měření

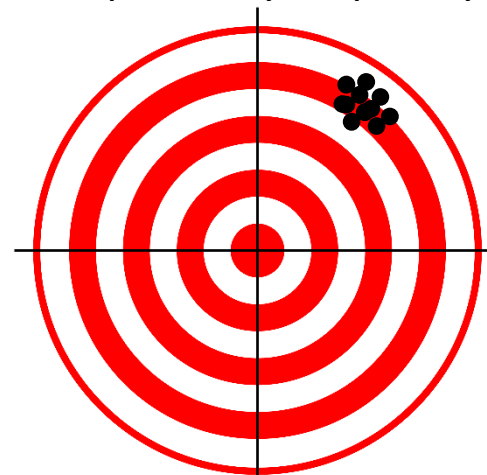
- Kvalita dat je klíčová pro jakékoliv statistické hodnocení
- Bez spolehlivých a přesných dat není možné získat spolehlivé a přesné výsledky statistického hodnocení
- Ve statistické analýze dat musíme zohlednit jak střed měření, tak variabilitu a zamyslet se nad přesností popisu reality



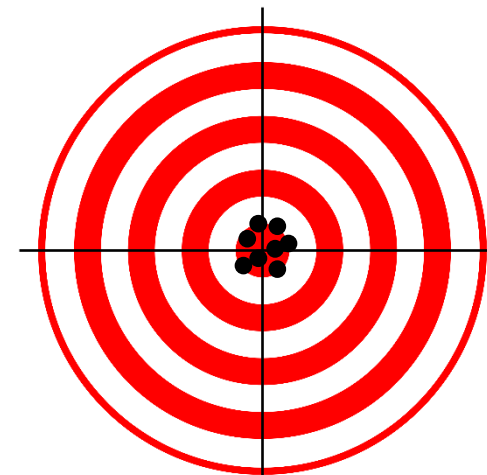
Nespolehlivý, nepřesný



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 !!!

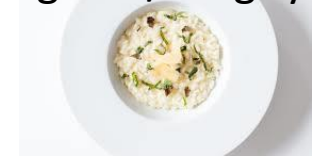
Zamícháte



5g soli / 1 kg rýže



5g soli / 1 kg rýže

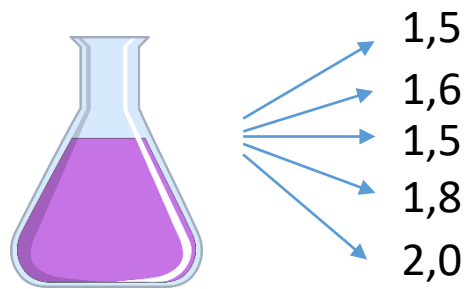


Průměr: 5g soli / 1 kg rýže
Vše OK !!!

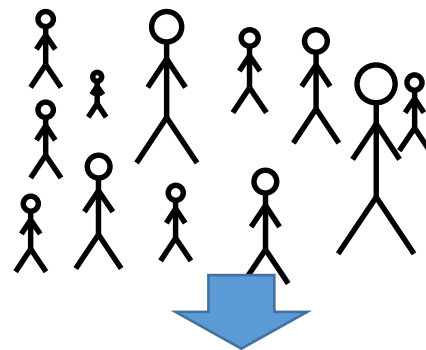
**Průměr není vše, je
nezbytné zohlednit
variabilitu**

Různé úrovně variability

Variabilita opakovaných měření

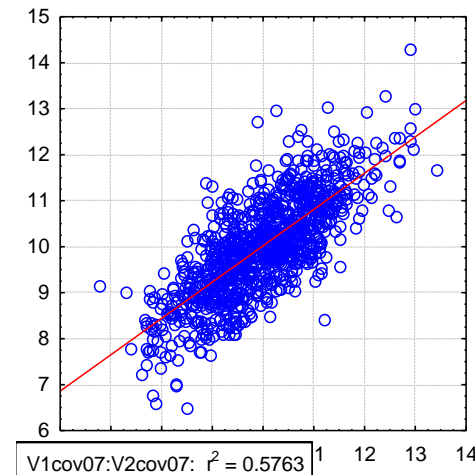


Variabilita dat v populaci

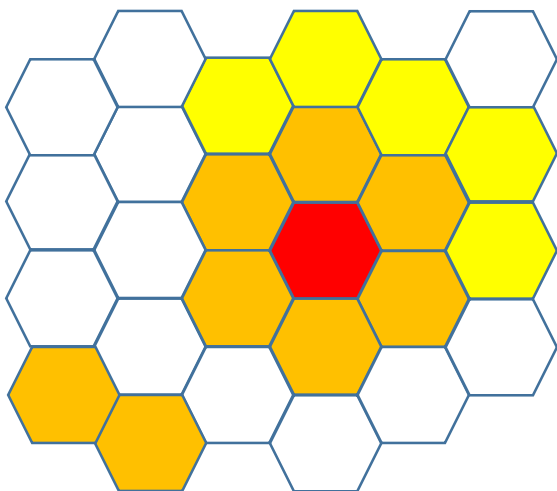


Hlavní téma kurzu

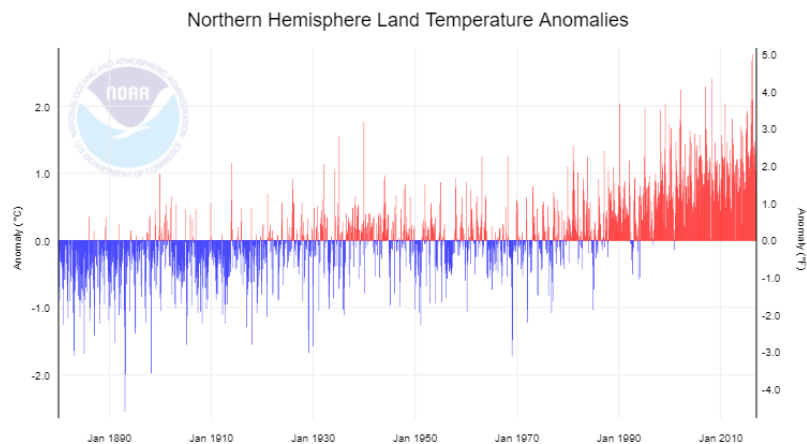
Variabilita v modelech



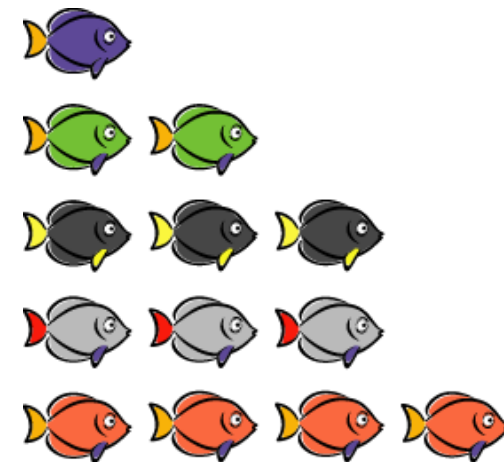
Geografická variabilita



Variabilita časových řad

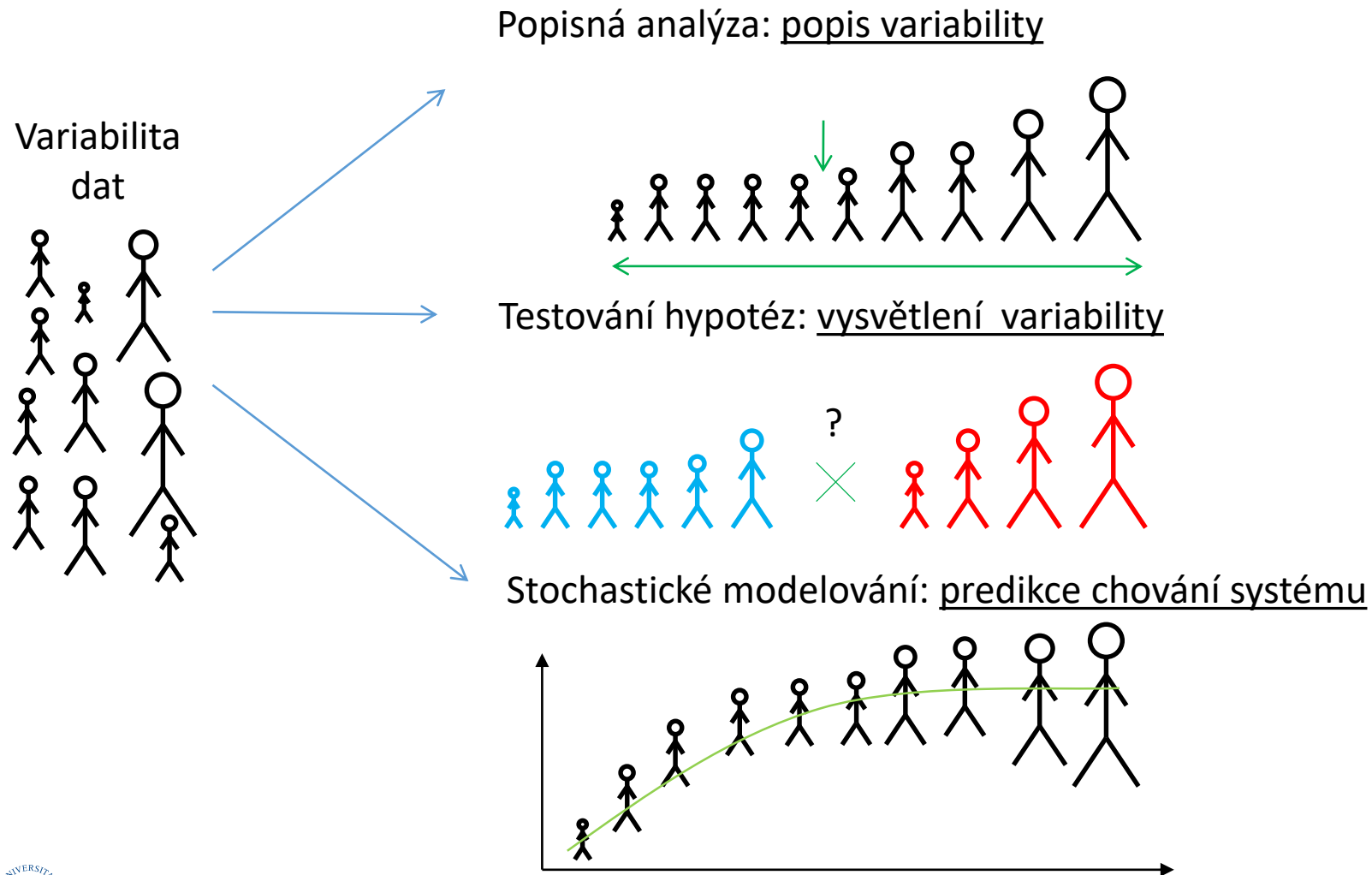


Biodiverzita



Práce s variabilitou v analýze dat

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



Statistika – definice

WWW.WIKIPEDIA.ORG:

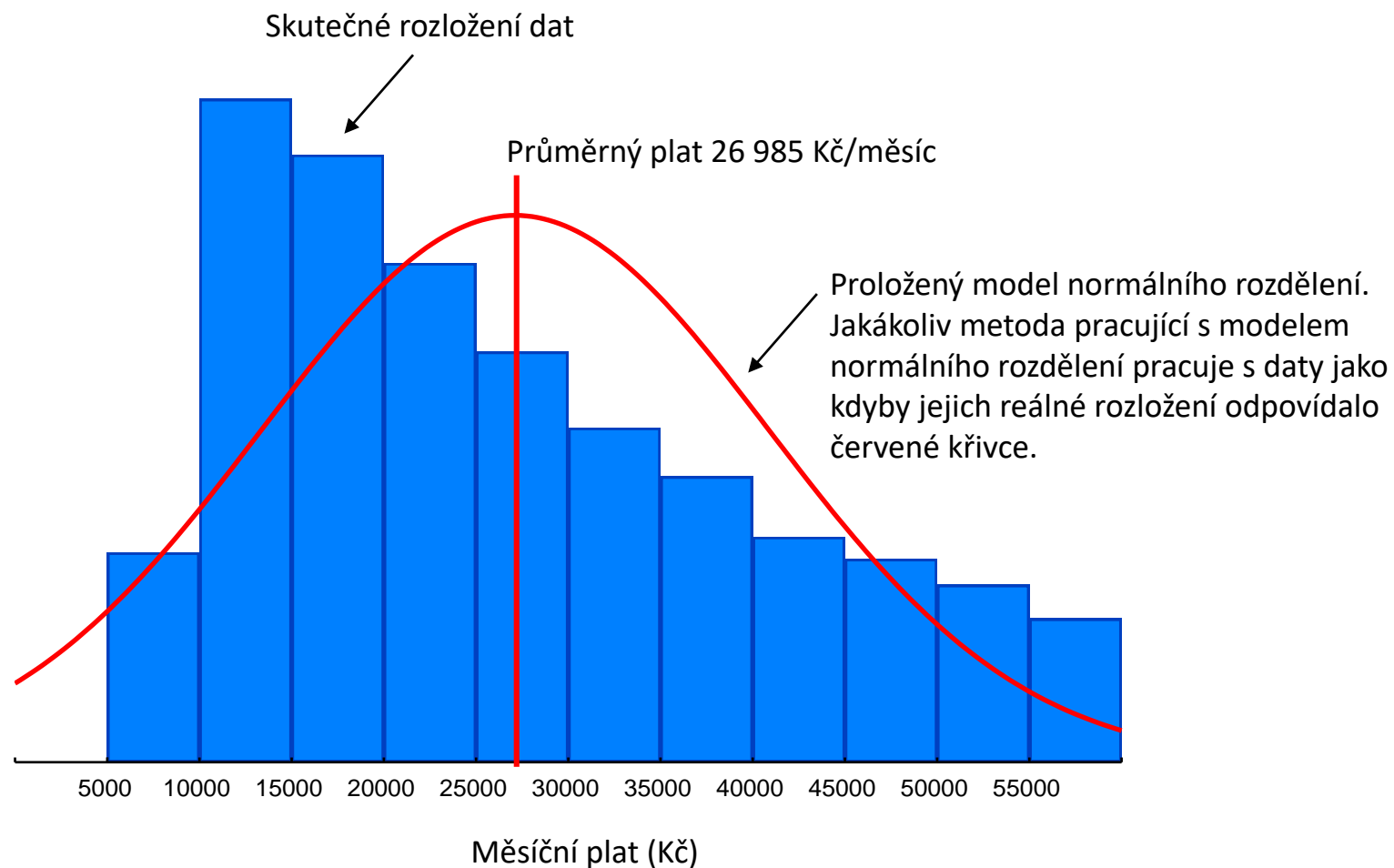
Statistika je matematickou vědou zabývající se shromážděním, analýzou, interpretací, vysvětlením a prezentací dat. Může být aplikována v širokém spektru vědeckých disciplín od přírodních až po sociální vědy. Statistika je využívána i jako podklad pro rozhodování, kdy nicméně může být záměrně i nevědomky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



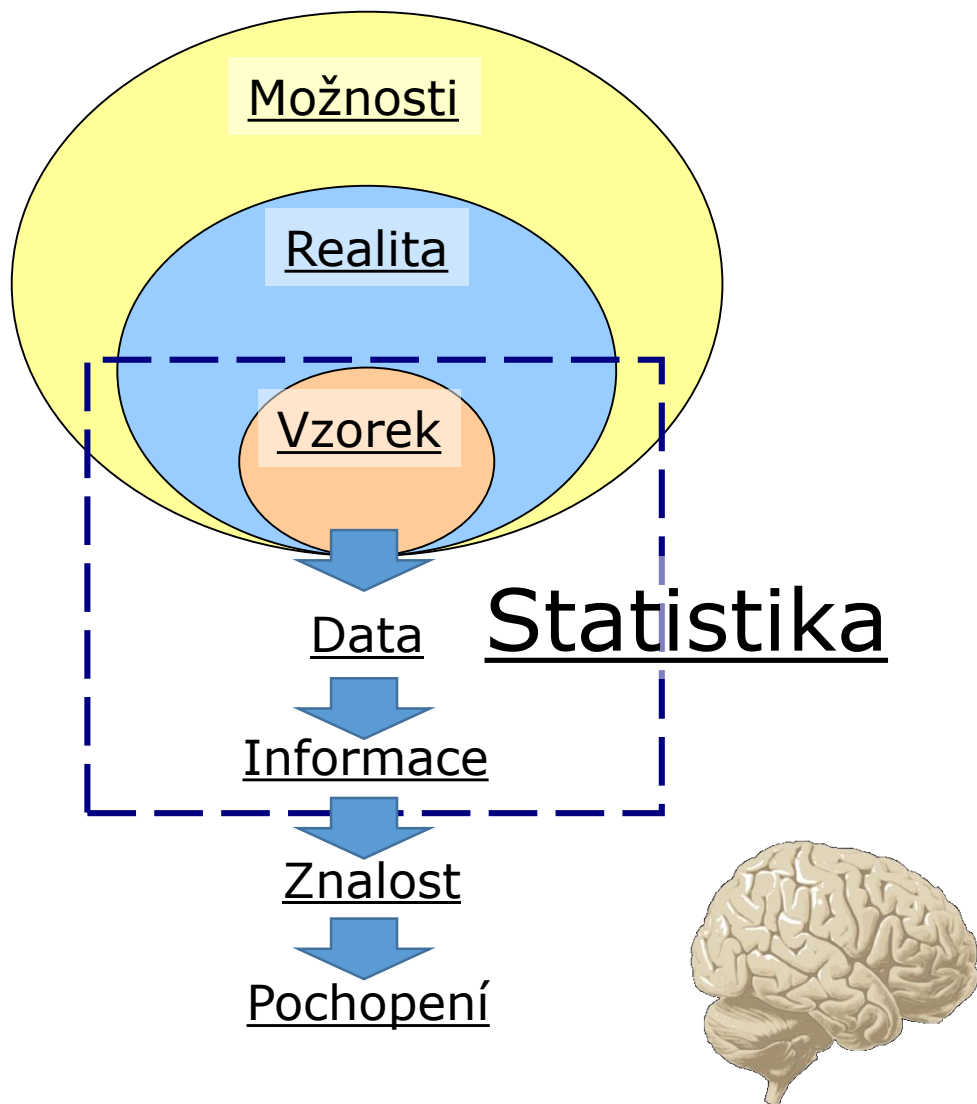
Statistika využívá matematické modely reality k zobecnění výsledků experimentů a vzorkování. Statistika funguje korektně pouze pokud jsou splněny předpoklady jejích metod a modelů.

Nesprávná aplikace modelu -> zkreslené závěry

- Různé popisné statistiky a testy jsou spjaté 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.

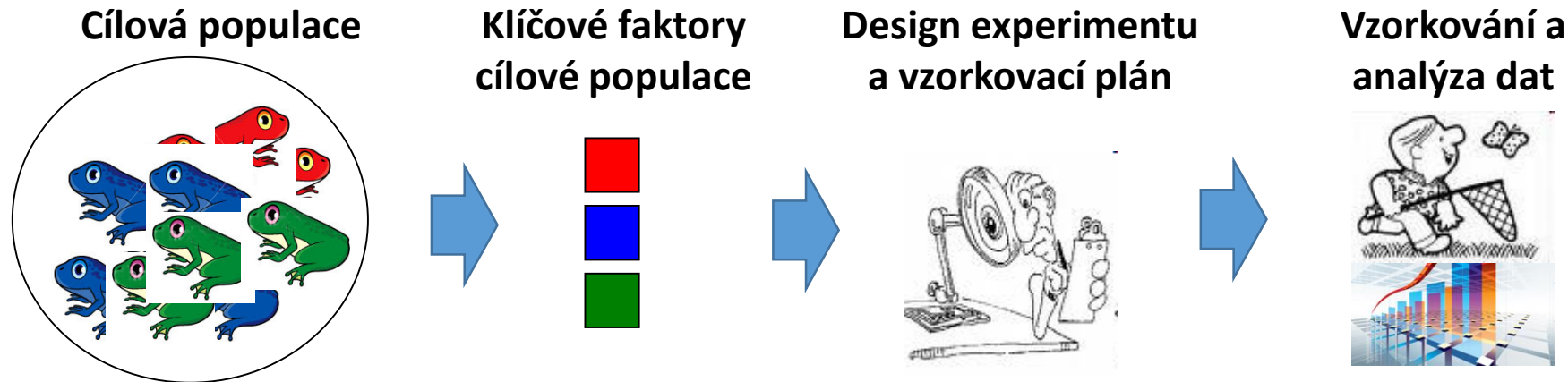
Statistika není náhradou naší inteligence !!!

Co musíme vědět před zahájením studie nebo experimentu?

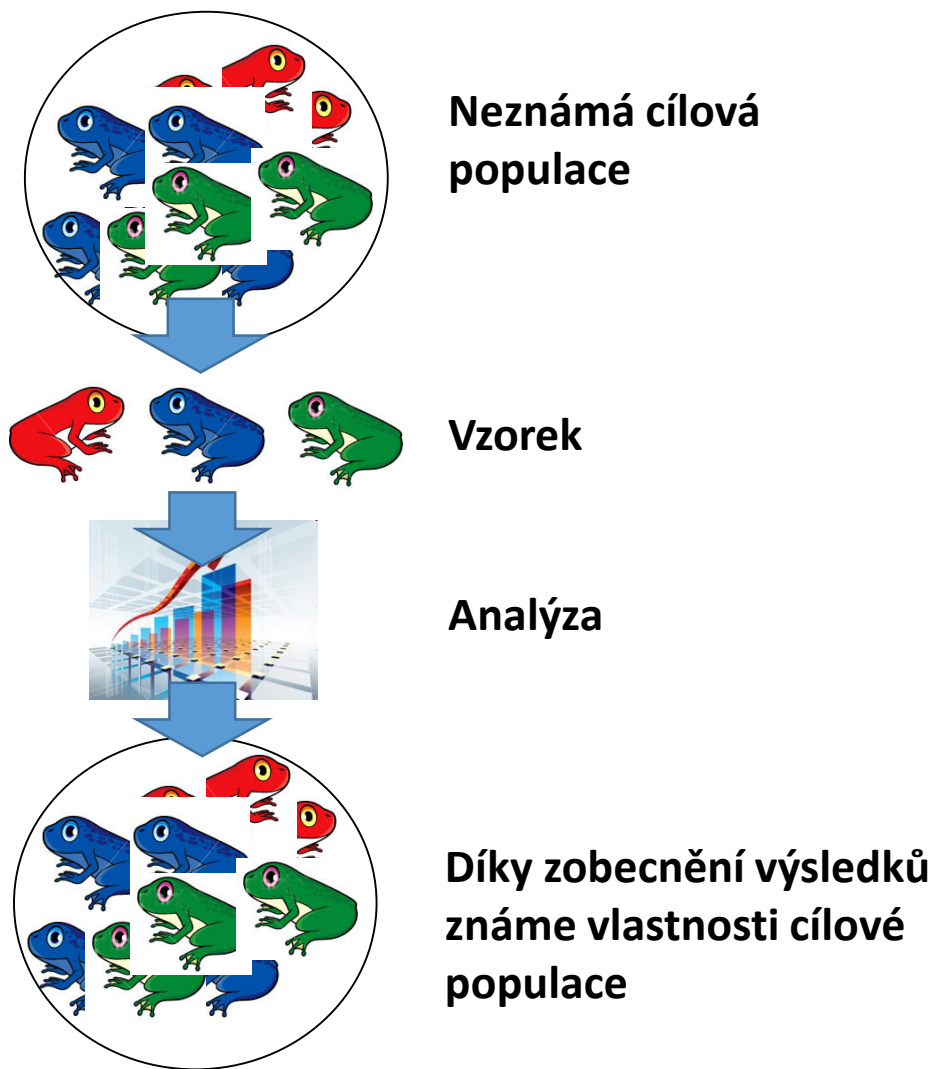
- Cílová populace
 - Skupina objektů (pacientů, lokalit atd.) na něž je studie zaměřena
- Primární hypotézy
 - Hlavní otázka položená ve studii – odhad velikosti vzorku a design studie je vypracován vzhledem k primární hypotéze (v řadě případů nelze v reálném výzkumu formální power analýzu vypracovat, 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



Statistika a zobecnění výsledků

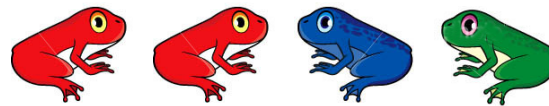


- Cílem analýzy není pouhý popis a analýza vzorku, ale zobecnění výsledků ze vzorku na jeho cílovou populaci
- Pokud vzorek nereprezentuje cílovou populaci, vede zobecnění k chybným závěrům

Vzorkování a jeho význam ve statistice

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

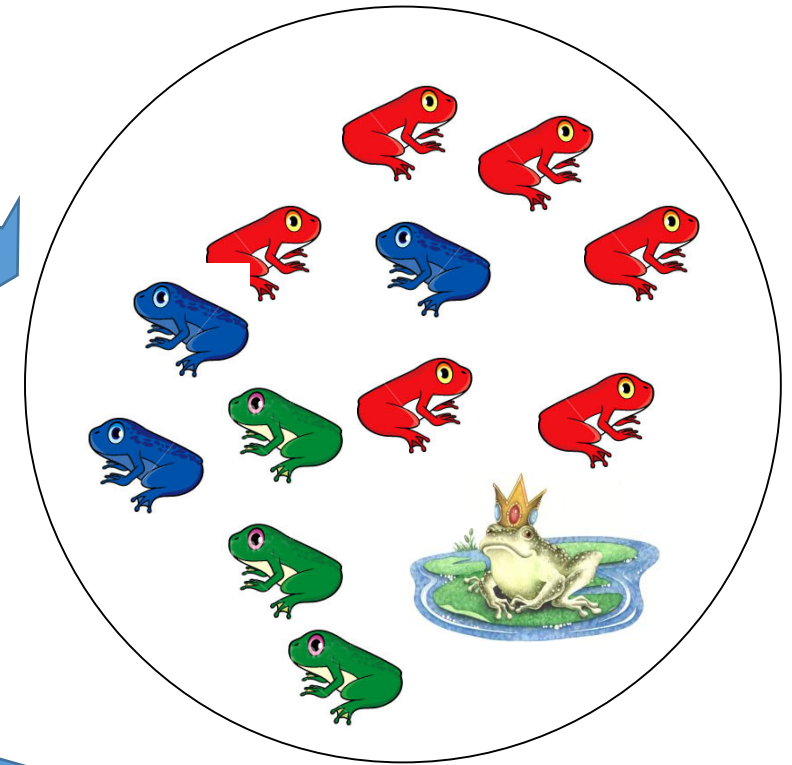
- **Representativnost:** struktura vzorku musí maximálně reflektovat realitu



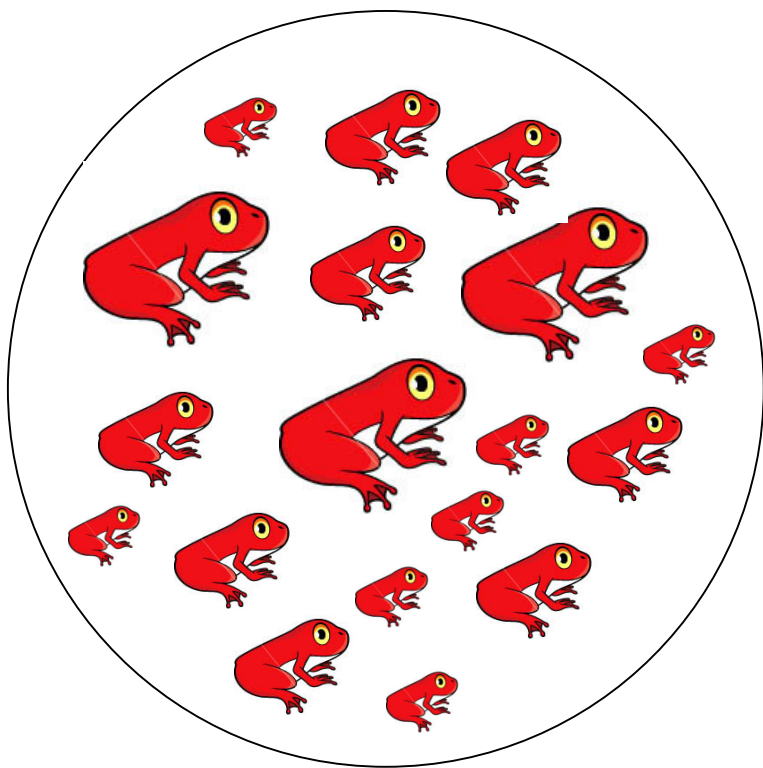
- **Nezávislost:** několikanásobné vzorkování téhož objektu nepřináší ze statistického hlediska žádnou novou informaci



- **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



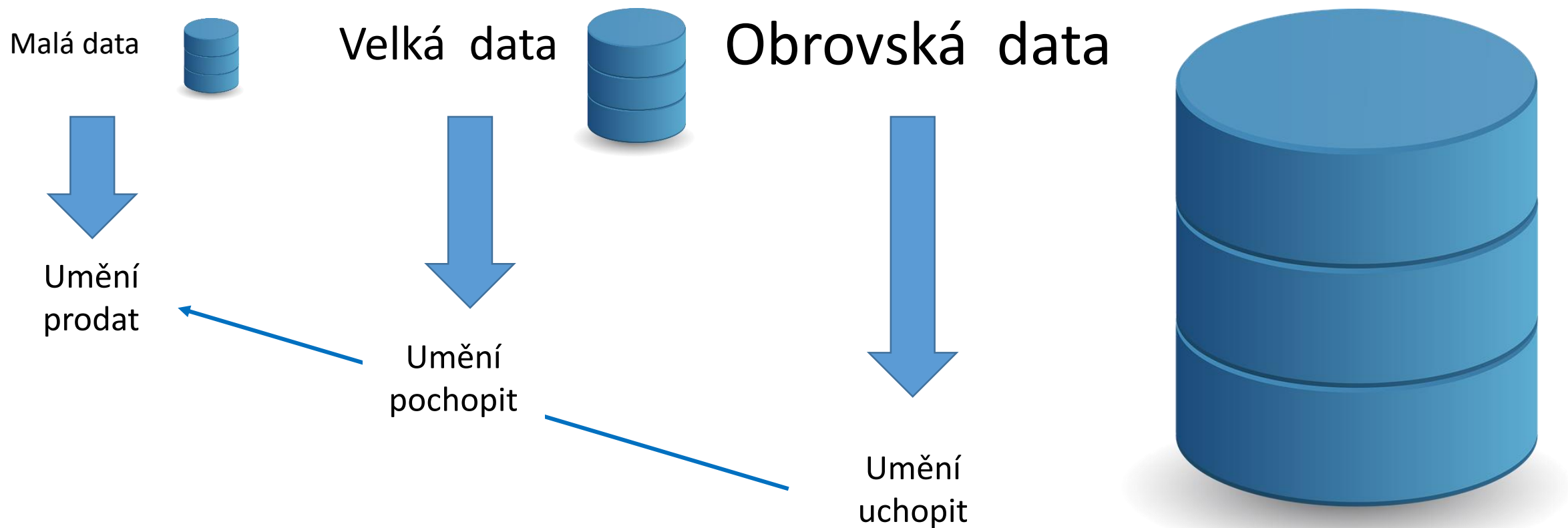
- Vzorek určité velikosti poskytuje odhad reálné hodnoty s definovanou spolehlivostí



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

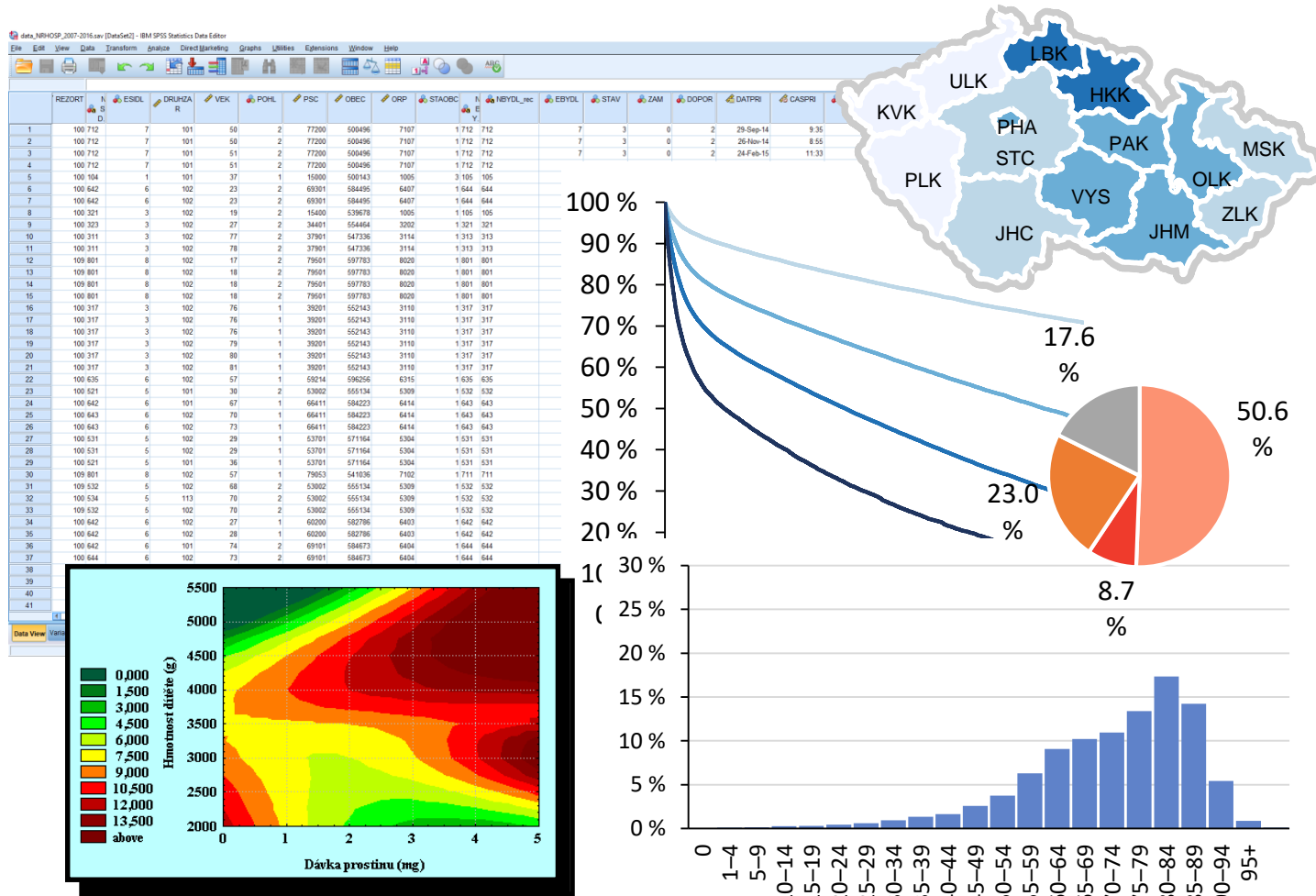
Různá velikost vzorku – různé úkoly analýzy dat

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



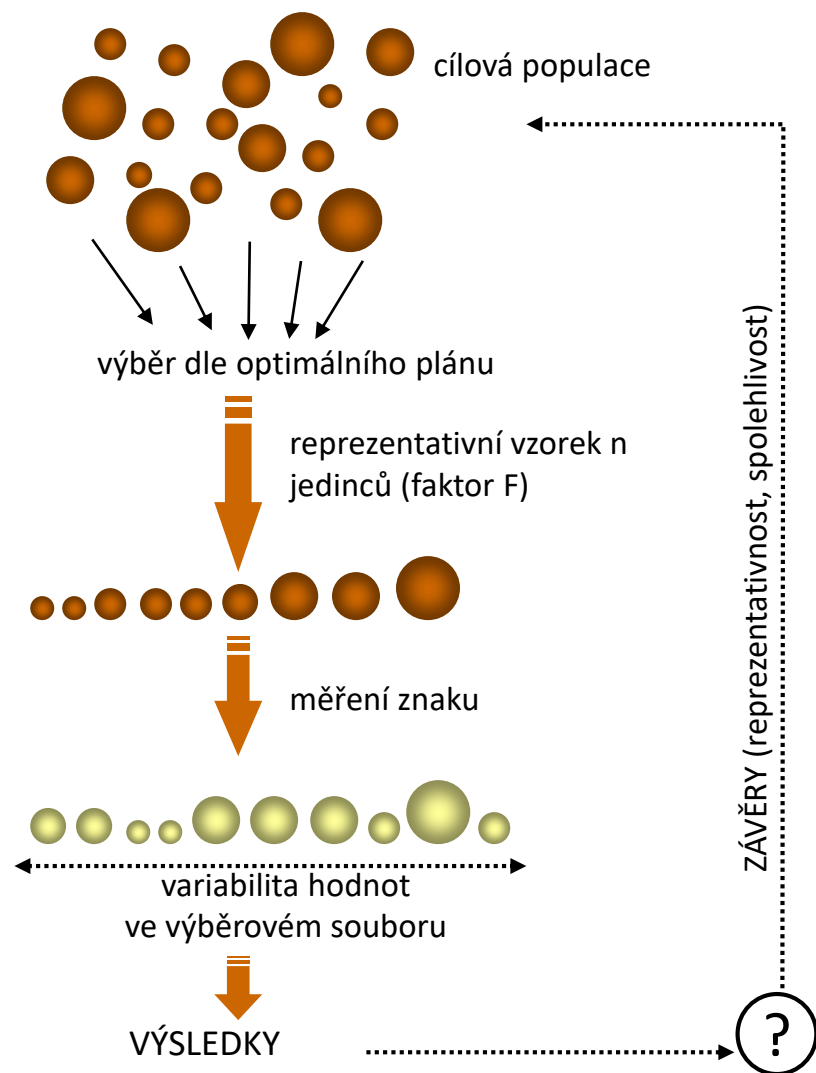
Přístup biostatistiky

- Schopnost: vidět data – komunikovat – interpretovat - prodávat



Experimentální design: nezbytná výbava biologa

Účel analýzy: Popisný

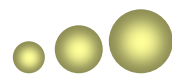


?

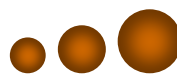
Reprezentativnost

Spolehlivost

Přesnost

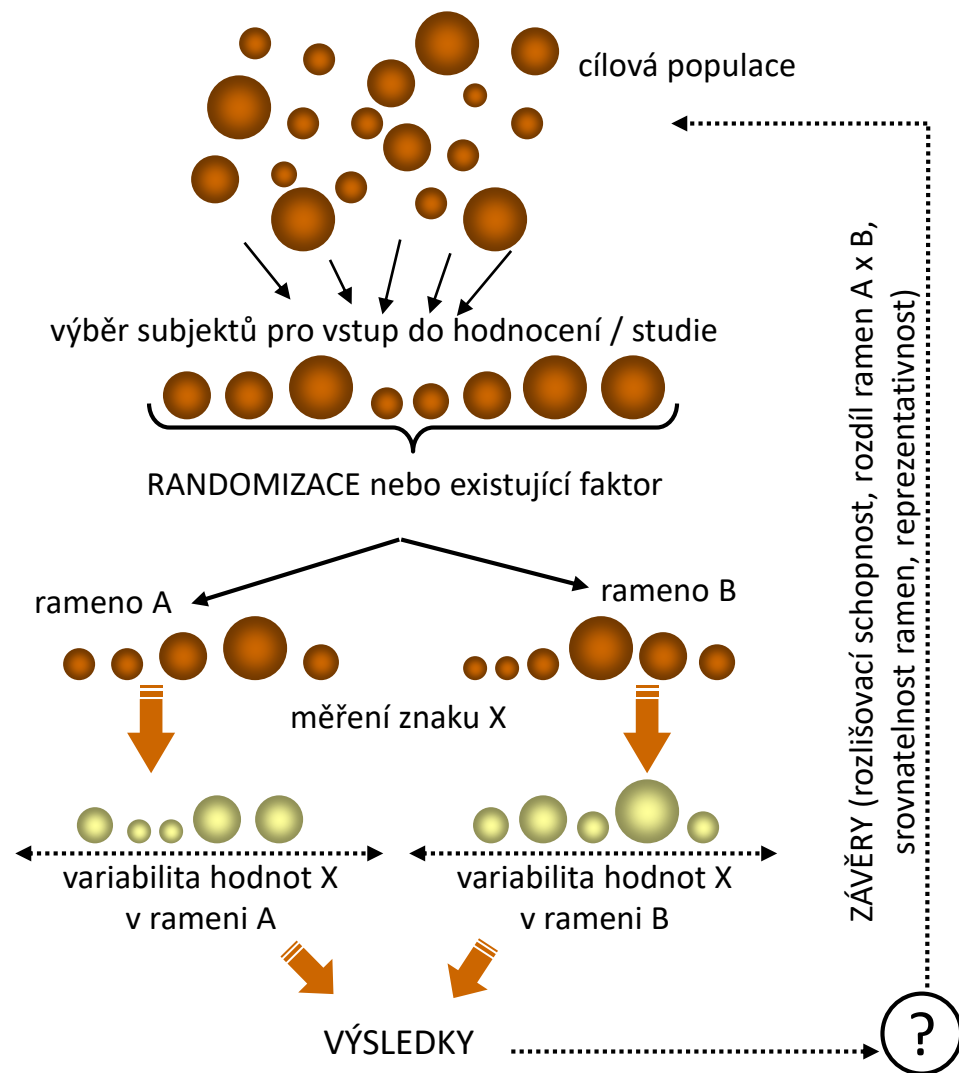


... analyzovaný znak
cílové populace (X)



... jiný významný faktor
charakterizující cílovou
populaci (F)

Experimentální design: nezbytná výbava biologa



Účel analýzy: Srovnávací (2 skupiny)

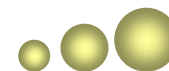
?

Reprezentativnost

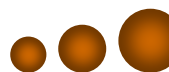
Srovnatelnost

Spolehlivost

Přesnost

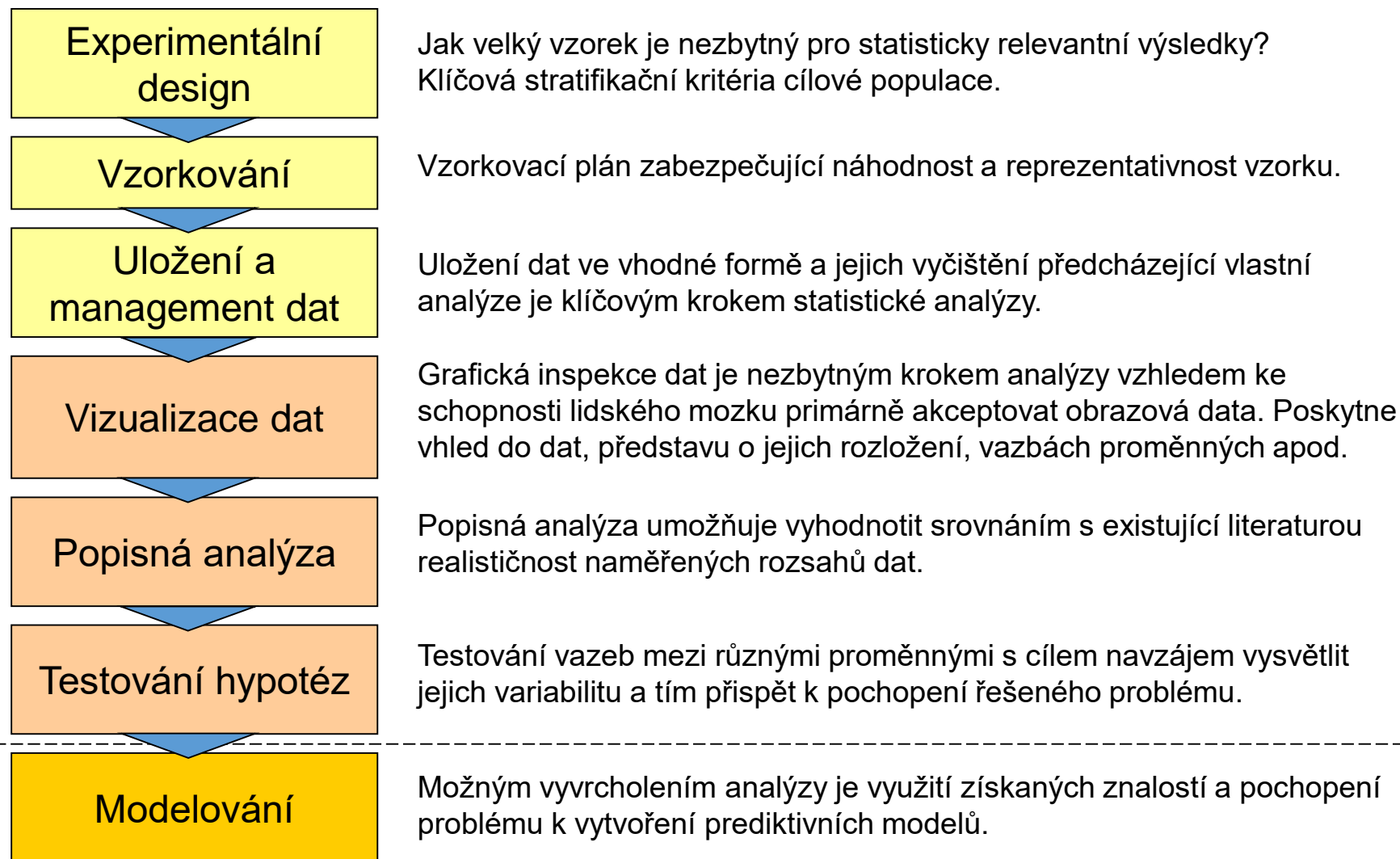


... analyzovaný znak
cílové populace (X)



... jiný významný faktor
charakterizující cílovou
populaci (F)

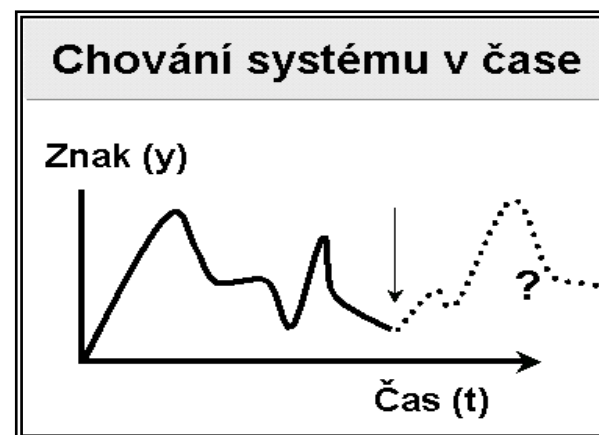
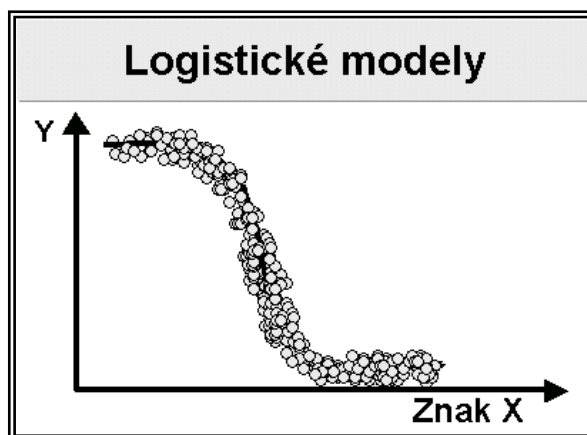
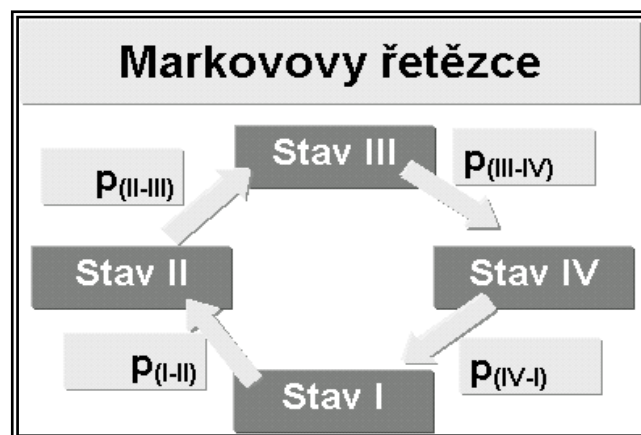
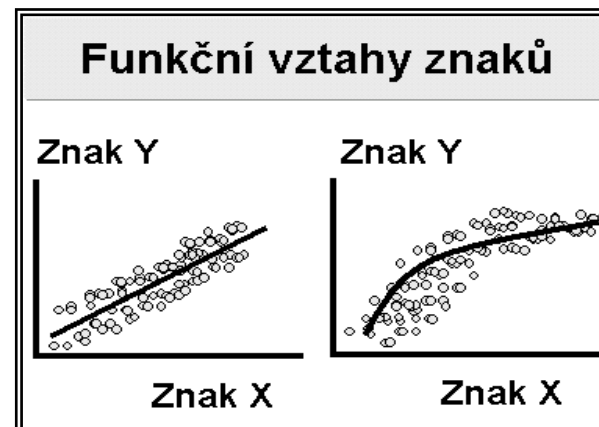
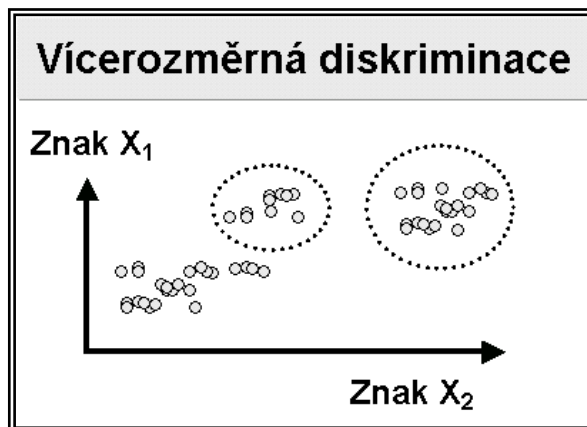
Obečné schéma využití statistické analýzy



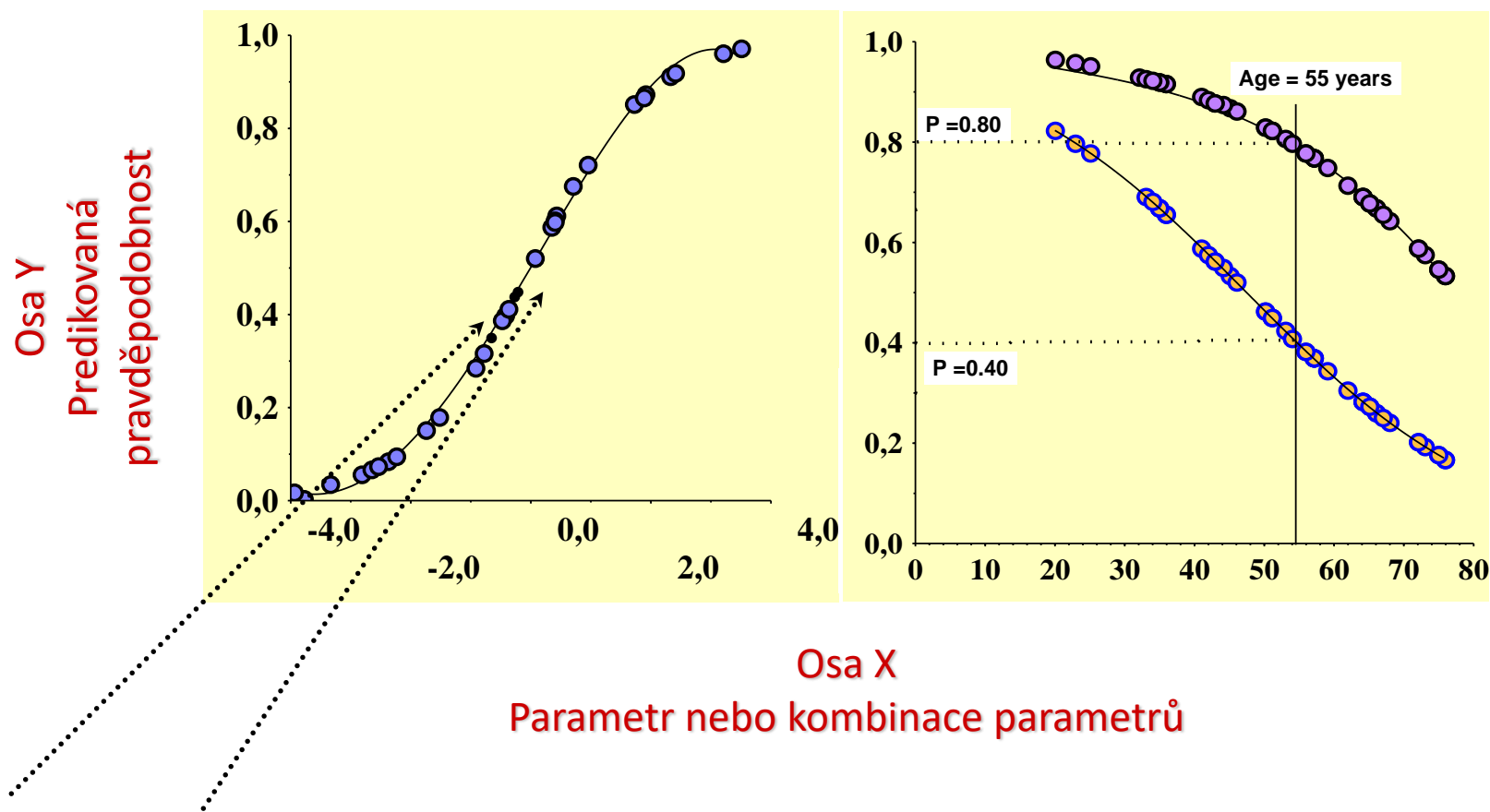
Stochastické modelování: predikce neurčitých jevů

- Prospektivně – modelově - postihuje chování jevů při respektování variability

Pravděpodobnostní vztahy					
Anamnéza x Výsledek vyšetření pacienta					
	Karcinom	Benigní léze	Benigní riziková	Zdravá	
Pozitivní anamnéza	2,22	34,44	0,00	63,33	100%
Negativní anamnéza	1,06	28,23	0,96	69,75	100%
$p < 0.05$					



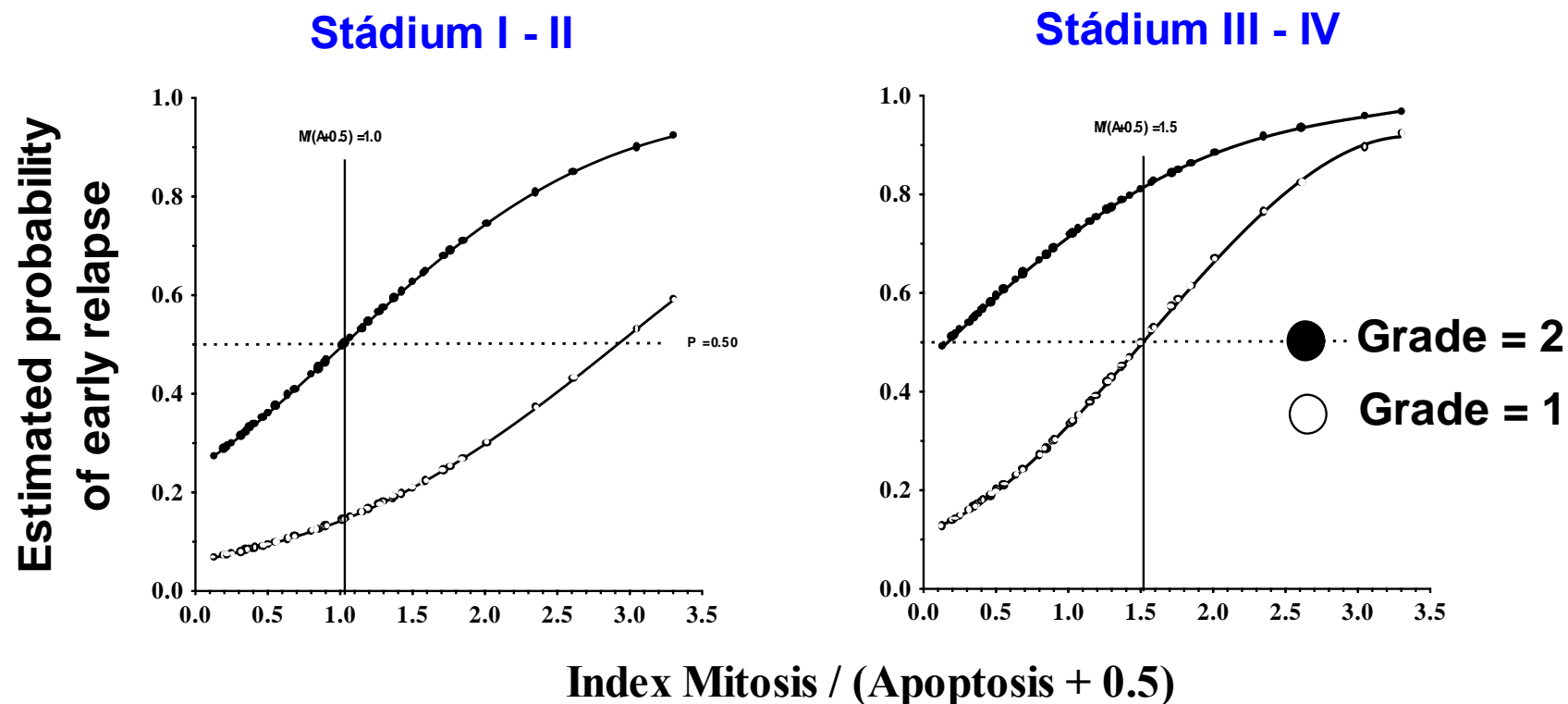
Stochastické modelování: predikce neurčitých jevů



Data konkrétních objektů k přímému
hodnocení

Stochastické modelování: predikce neurčitých jevů

- Schopnost: vytvářet prakticky využitelné nástroje



Přednáška 2

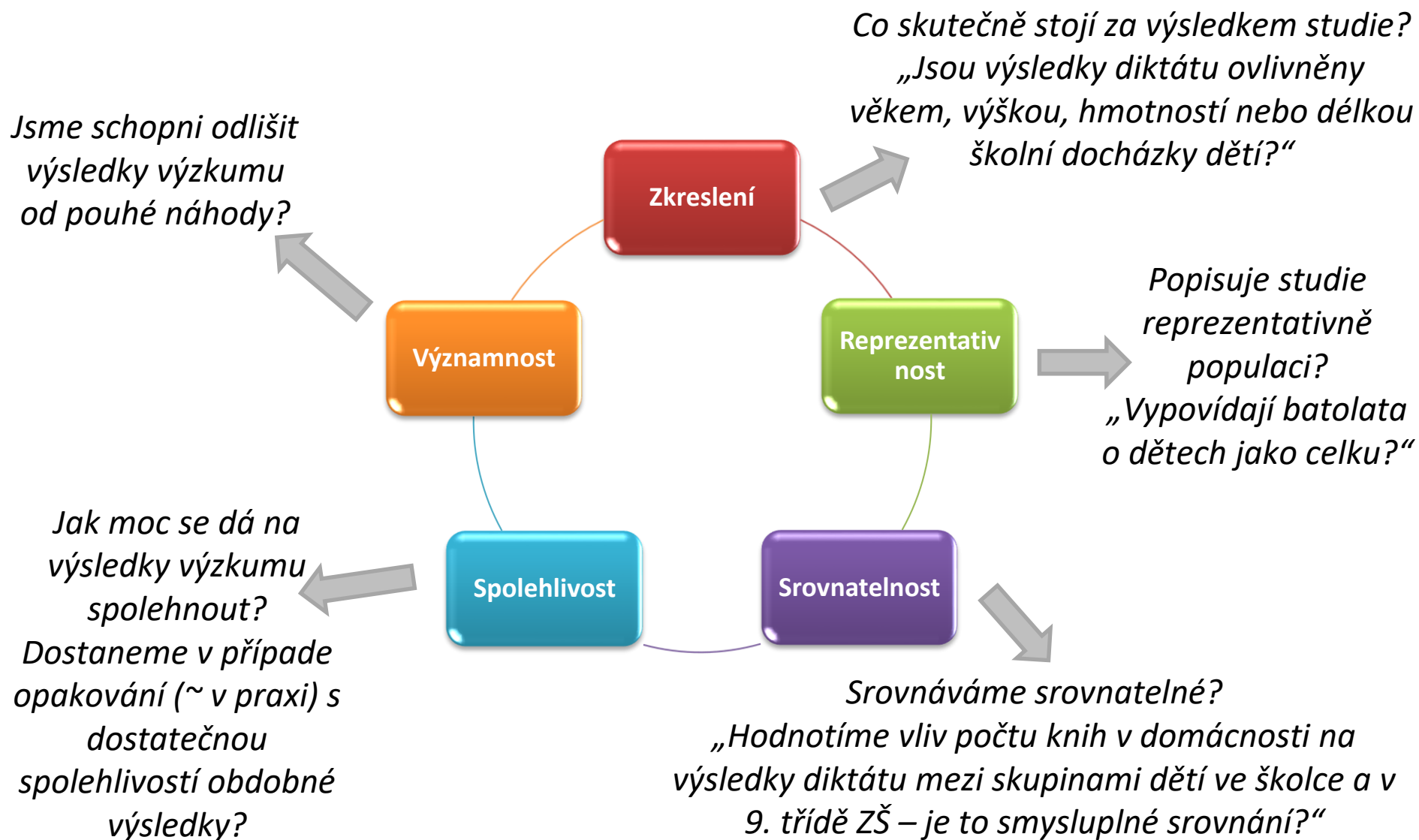
Klíčové principy biostatistiky

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

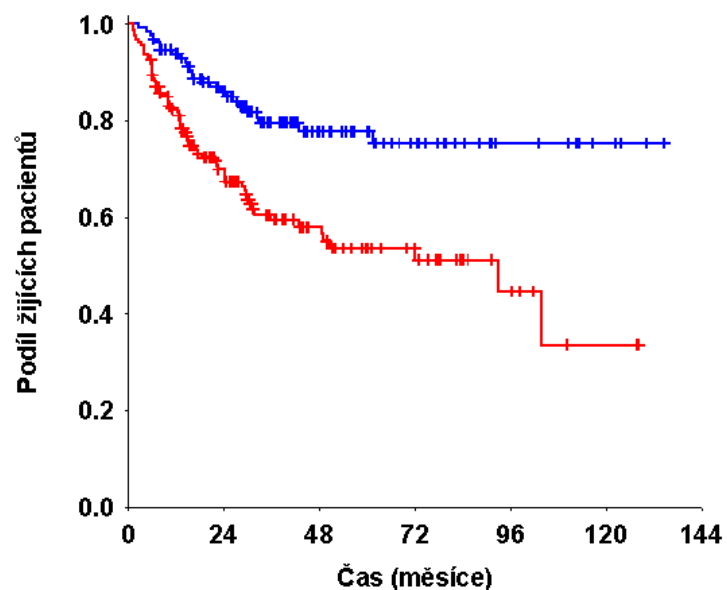


Klíčové principy – zkreslení

- 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?

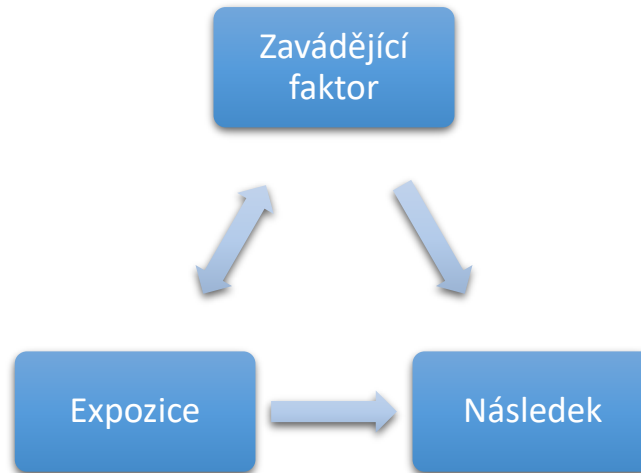
Nějaký prognostický faktor?

Stadium nemoci?

Věk?

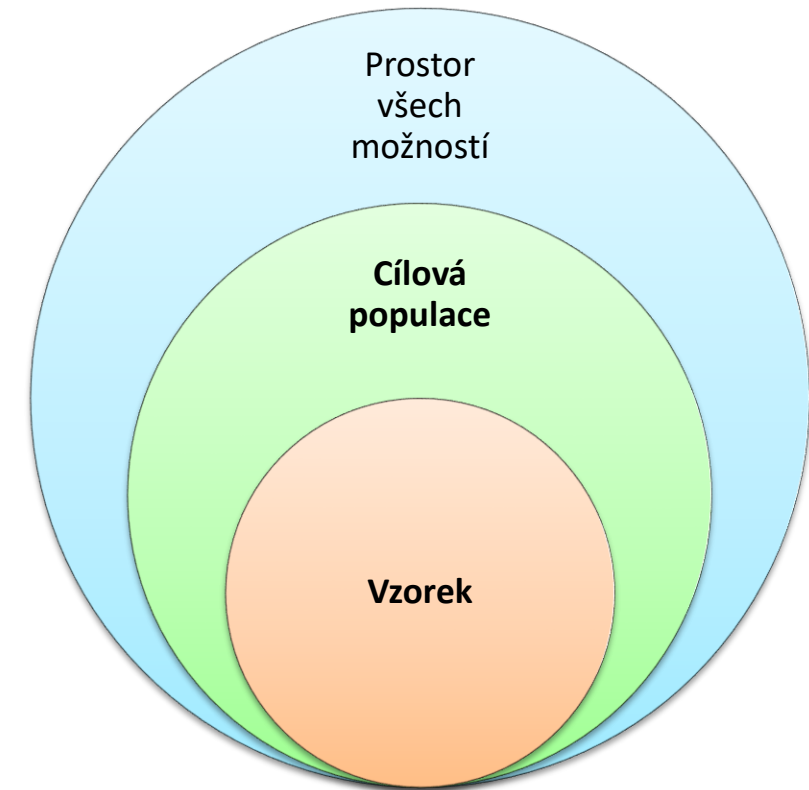
Klíčové principy – zkreslení

- 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.

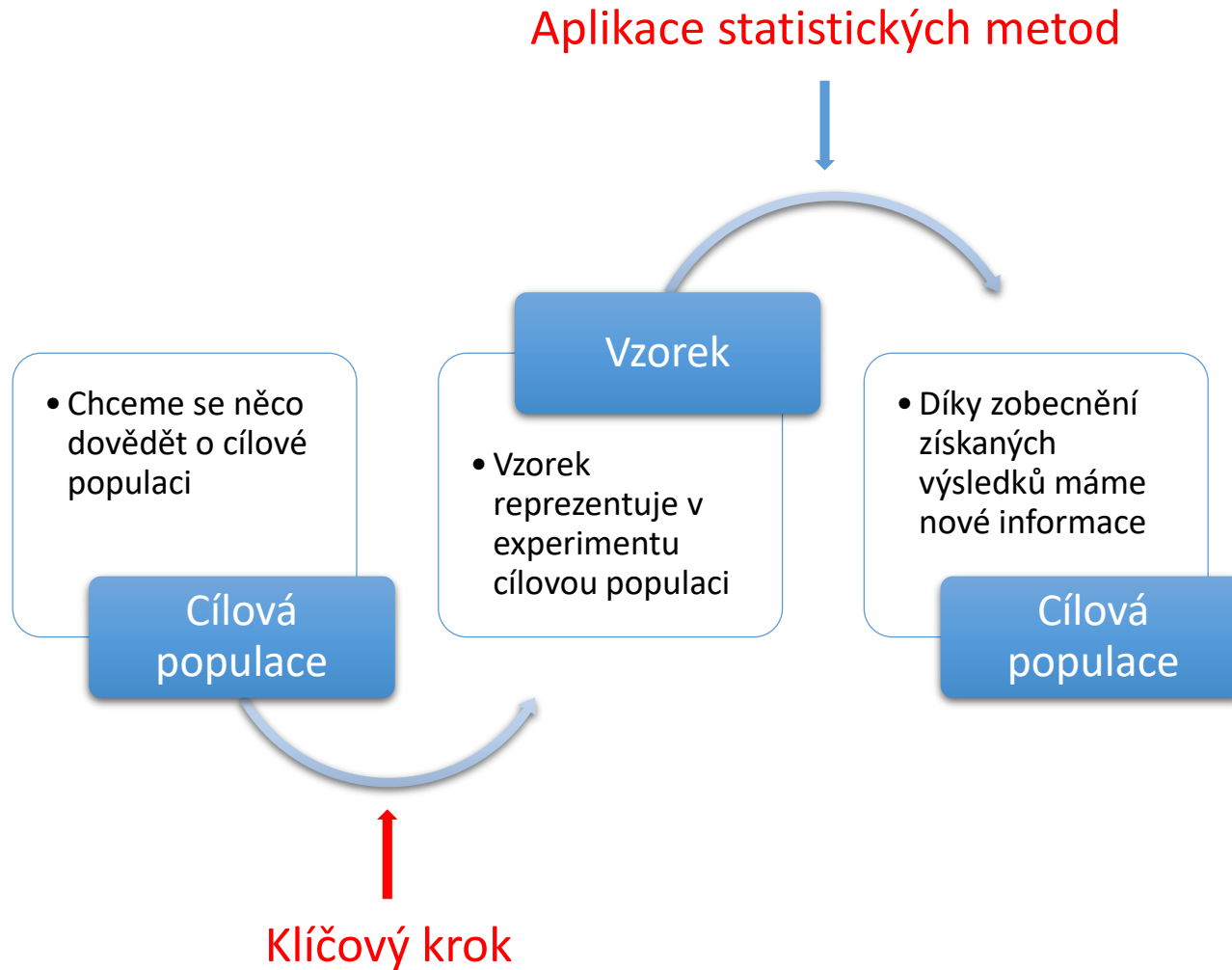


Klíčové principy – reprezentativnost

- 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

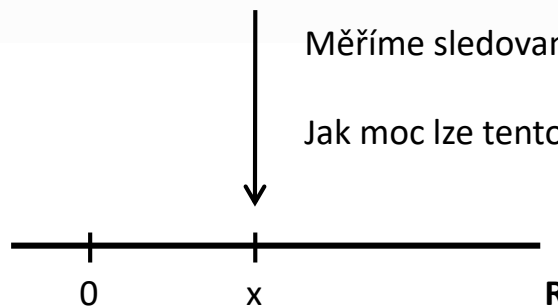
- Ve většině studií nás zajímá kvantifikace sledovaného efektu nebo charakteristiky, obecně náhodné veličiny, ve formě jednoho čísla, bodového odhadu.
- Bodový odhad je však sám o sobě nedostatečný.
- Je nutné ho doplnit intervalovým odhadem, který odpovídá pravděpodobnostnímu chování sledované veličiny, tedy odpovídá určité spolehlivosti výsledku.

Klíčové principy – spolehlivost

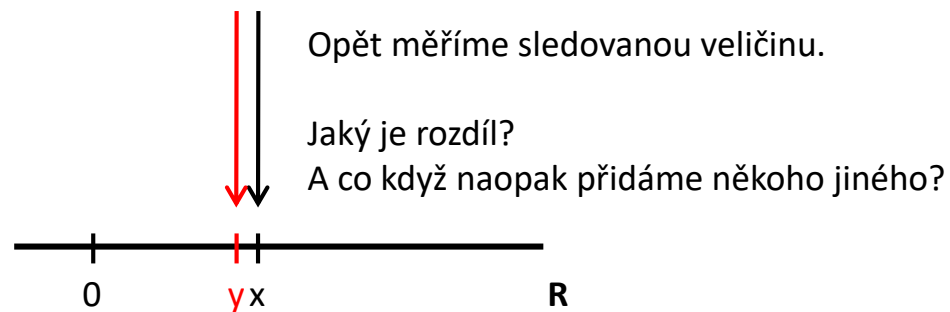


Měříme sledovanou veličinu a následně spočítáme odhad.

Jak moc lze tento bodový odhad zobecnit na cílovou populaci?

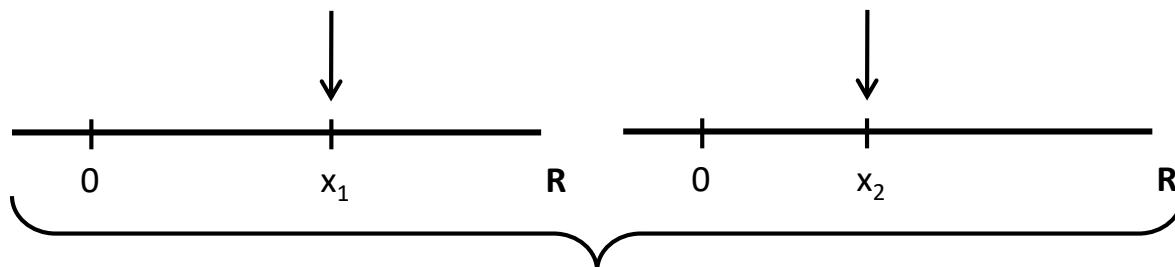


Klíčové principy – spolehlivost

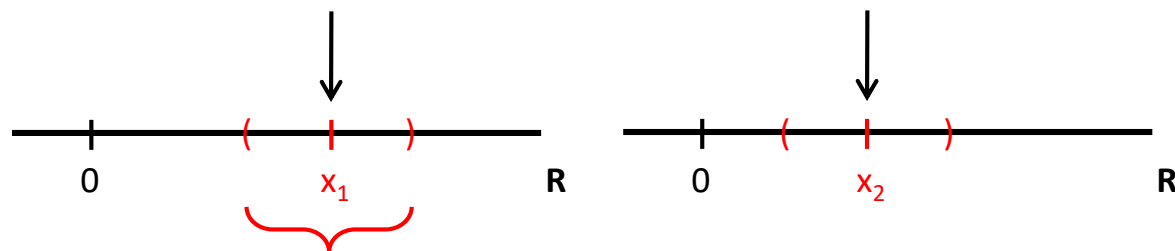


Klíčové principy – spolehlivost

Výběr číslo 1

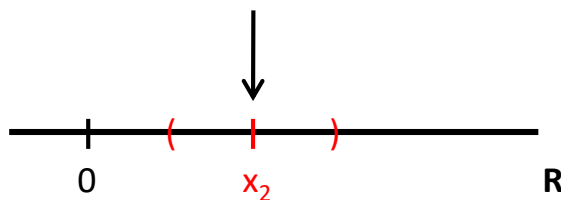
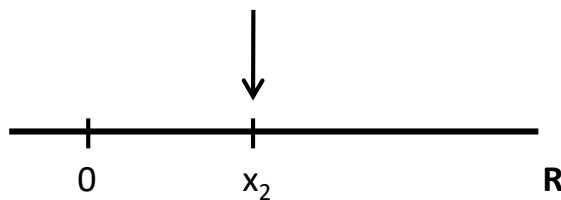


Pracujeme-li s výběrem z cílové populace, je třeba na základě variability pozorovaných dat spočítat tzv. interval spolehlivosti pro bodový odhad.

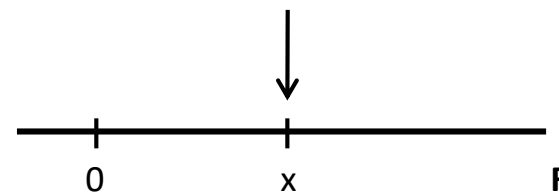


Interval spolehlivosti na základě výběru číslo 1.

Výběr číslo 2



Celá cílová populace



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

		Praktická významnost	
		ANO	NE
Statistická významnost	ANO	OK, praktická i statistická významnost jsou ve shodě.	Významný výsledek je statistický artefakt, prakticky nevyužitelný.
	NE	Výsledek může být pouhá náhoda, neprůkazný výsledek.	OK, praktická i statistická významnost jsou ve shodě.



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

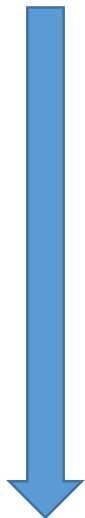
- Současná statistická analýza se neobejde bez zpracování dat pomocí statistických software.
- Předpokladem úspěchu je správné uložení dat ve formě „databázové“ tabulky umožňující jejich zpracování v libovolné aplikaci.
- Neméně důležité je věnovat pozornost čištění dat předcházející vlastní analýze.
- 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.

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

Nereálné odlehle hodnoty, pravděpodobně prohozen věk a výška.

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

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

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

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

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

- Nejrozumnější software – nejrozumnější možnosti
 - MS Office – základní grafy, snadná editovatelnost, lze invenčně upravit, snadná replikovatelnost výměnou dat
 - R – různé knihovny (např. ggplot) – vyšší vstupní investice, nejrozumnější 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

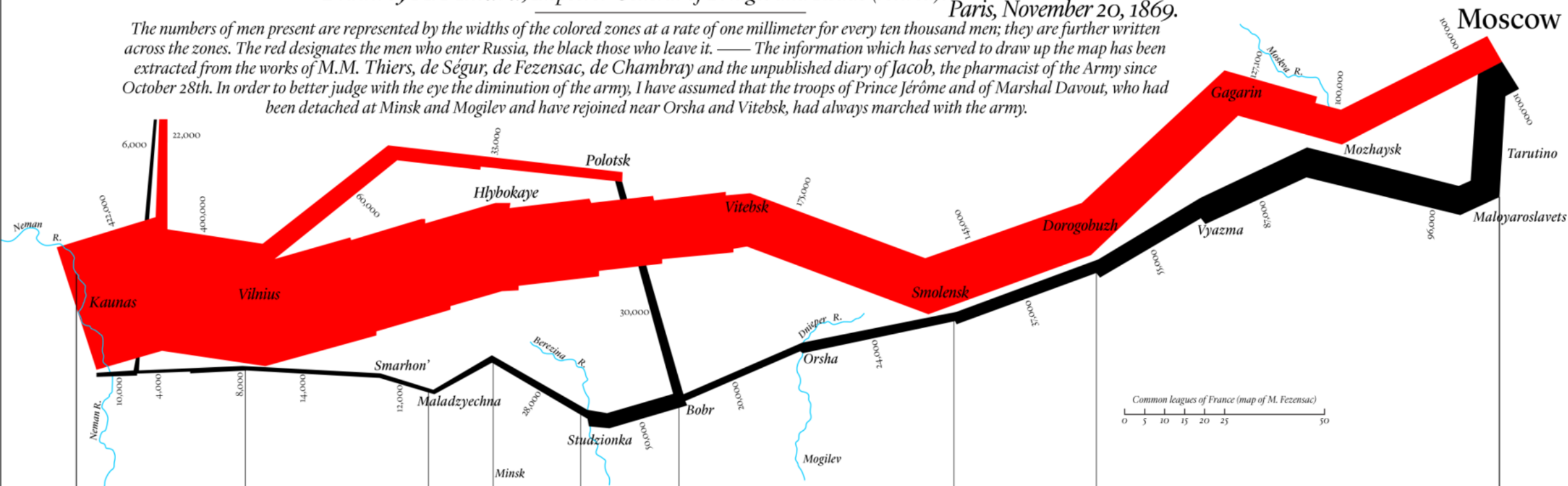
Slavné grafy: Charles Joseph Minard – Napoleonovo tažení do Ruska

Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812 ~ 1813

Drawn by M. Minard, Inspector General of Bridges and Roads (retired).

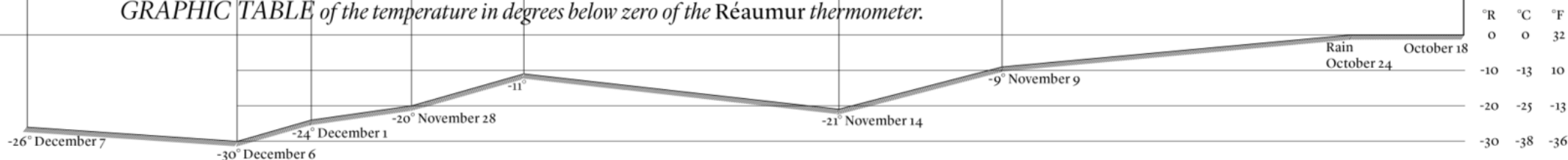
Paris, November 20, 1869.

The numbers of men present are represented by the widths of the colored zones at a rate of one millimeter for every ten thousand men; they are further written across the zones. The red designates the men who enter Russia, the black those who leave it. — The information which has served to draw up the map has been extracted from the works of M.M. Thiers, de Ségur, de Fezensac, de Chambray and the unpublished diary of Jacob, the pharmacist of the Army since October 28th. In order to better judge with the eye the diminution of the army, I have assumed that the troops of Prince Jérôme and of Marshal Davout, who had been detached at Minsk and Mogilev and have rejoined near Orsha and Vitebsk, had always marched with the army.



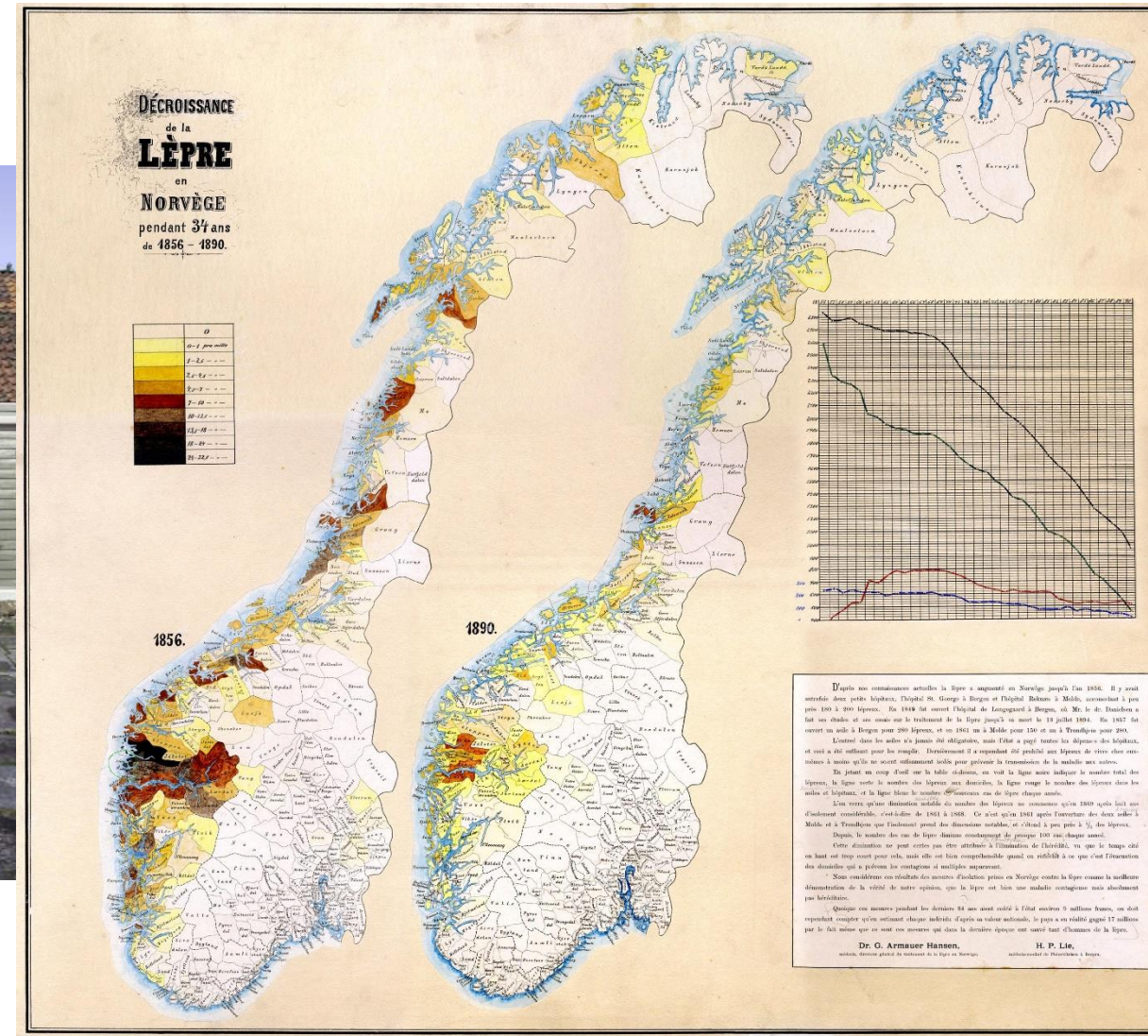
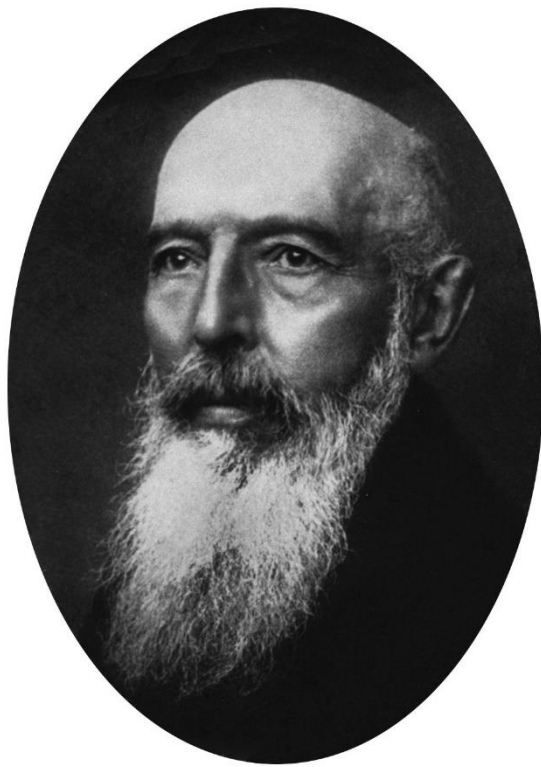
GRAPHIC TABLE of the temperature in degrees below zero of the Réaumur thermometer.

The Cossacks pass the frozen Neman at a gallop.



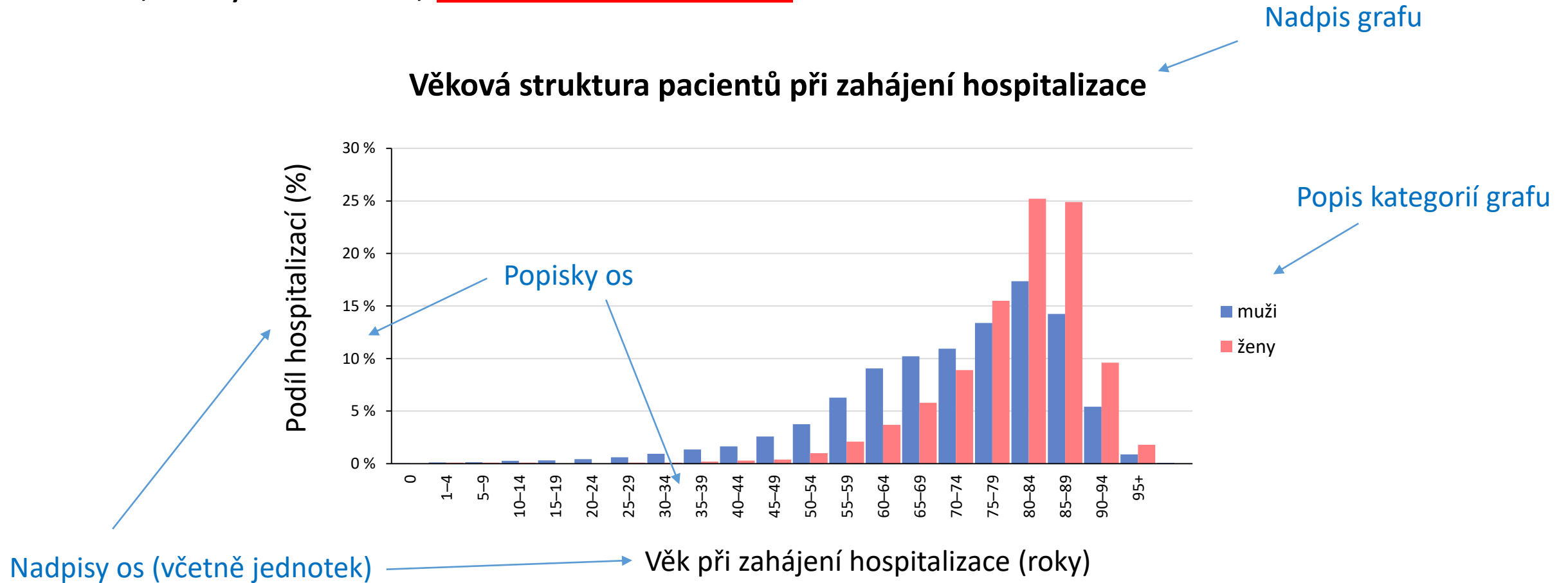
Slavné grafy: Eradikace lepry v Norsku

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



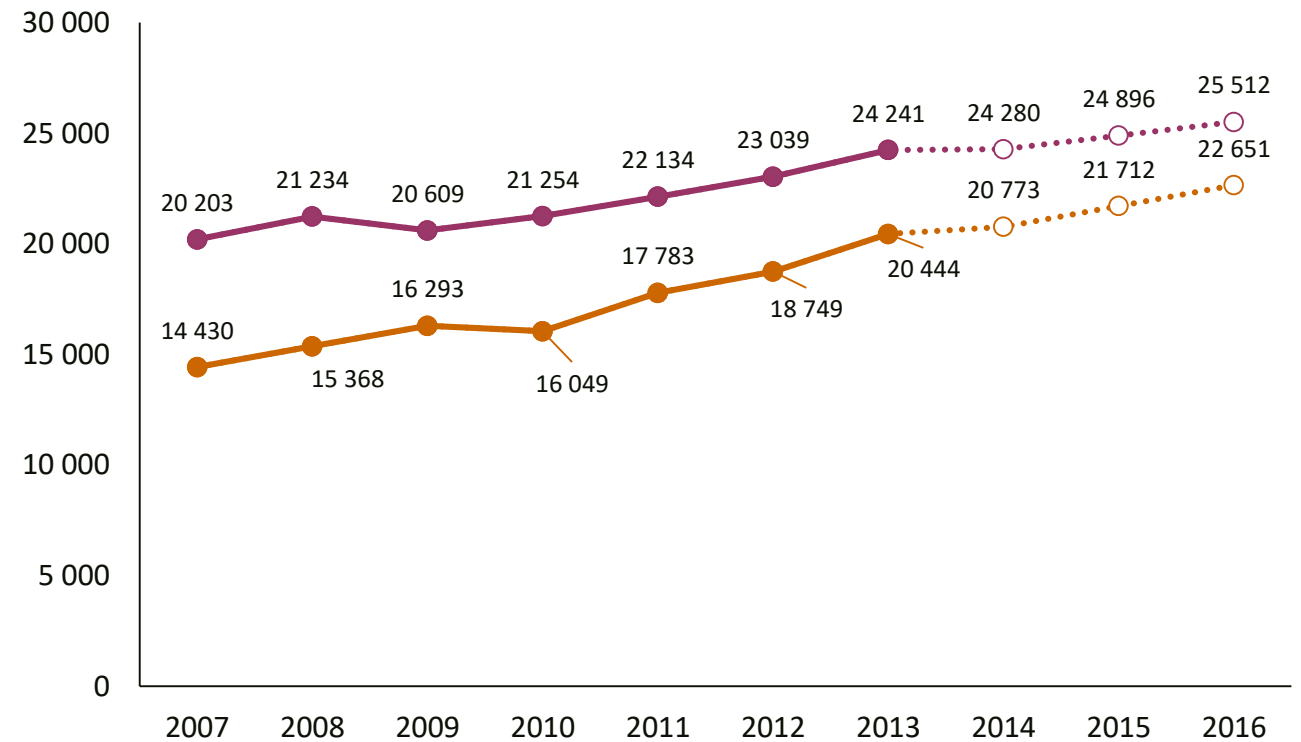
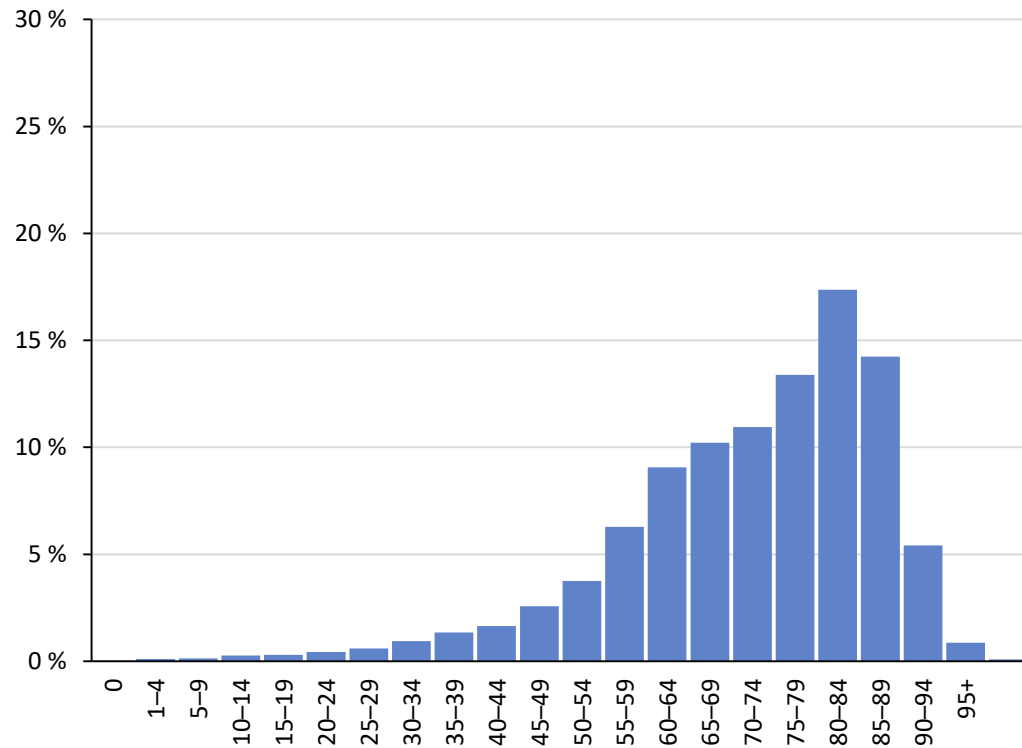
Co nesmí chybět na grafu

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



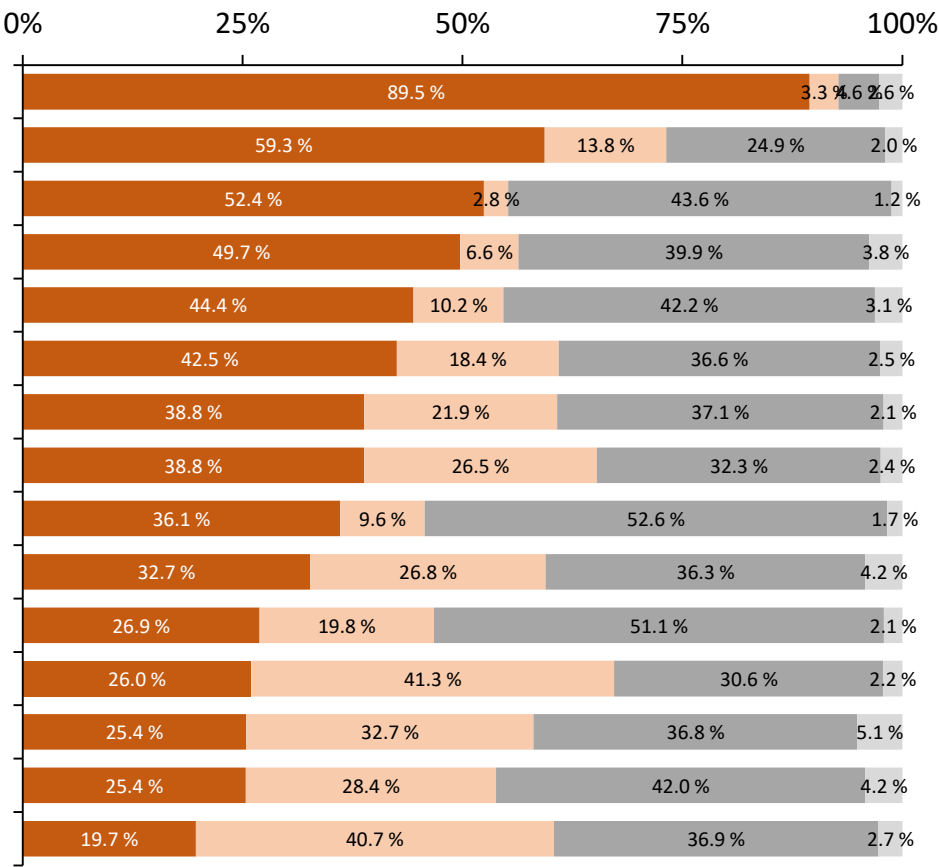
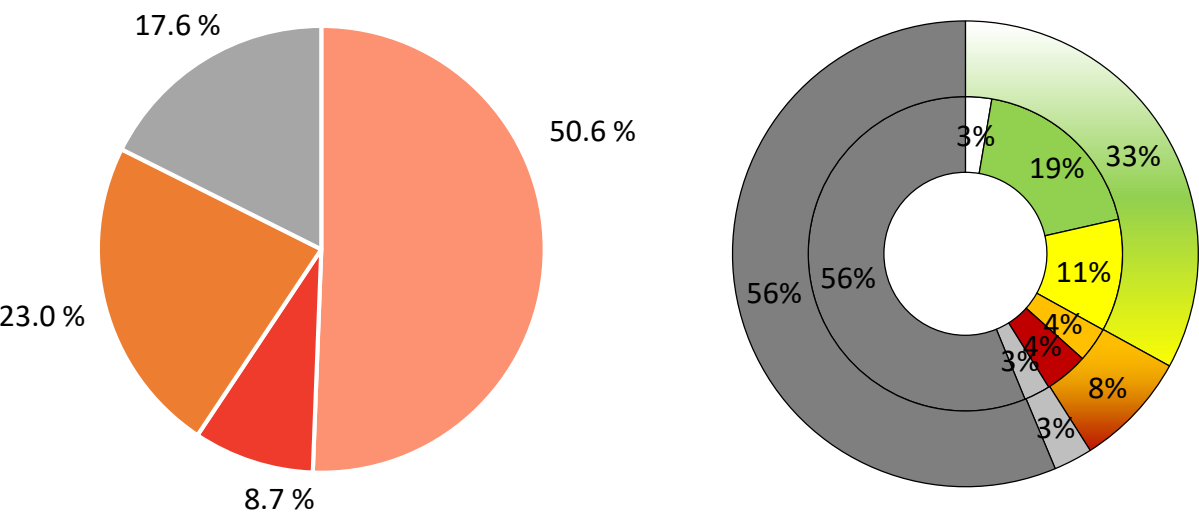
Sloupcové a čárové grafy

- Jednoduchá tvorba, vizualizace absolutních hodnot nebo procent



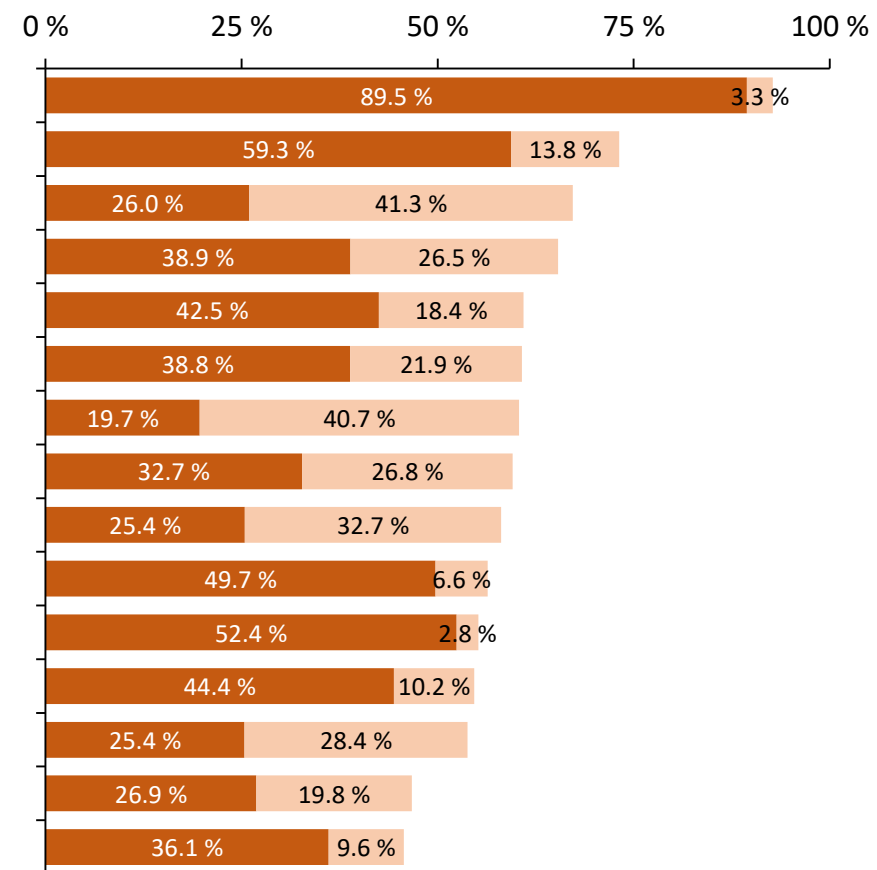
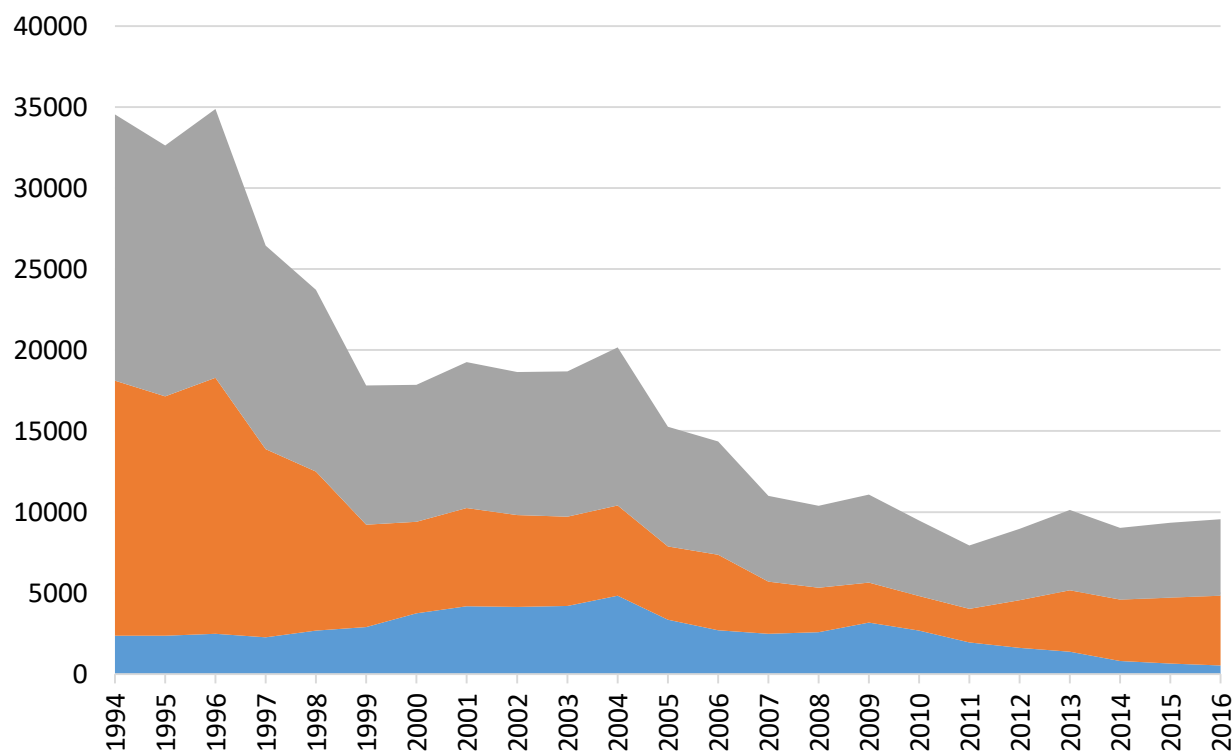
Koláčové a páskové grafy

- Jednoduchá tvorba, vizualizace procent



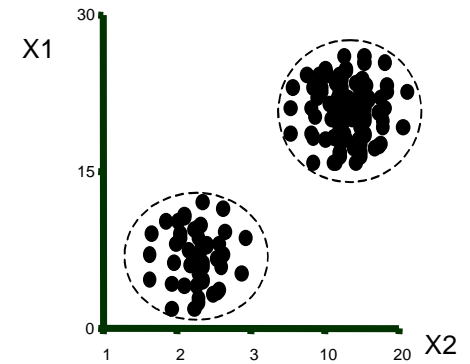
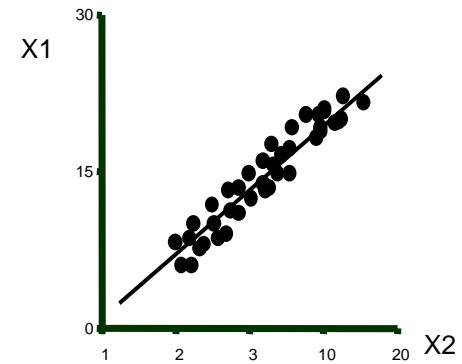
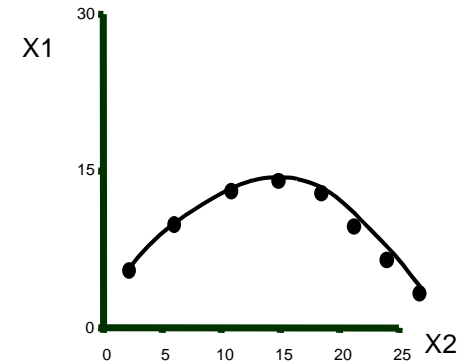
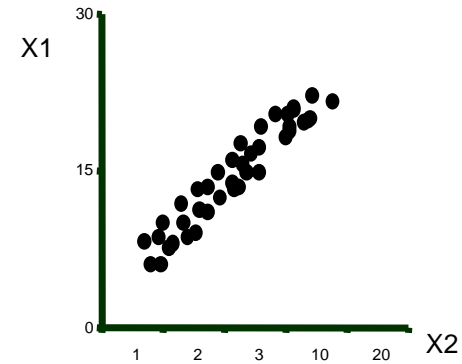
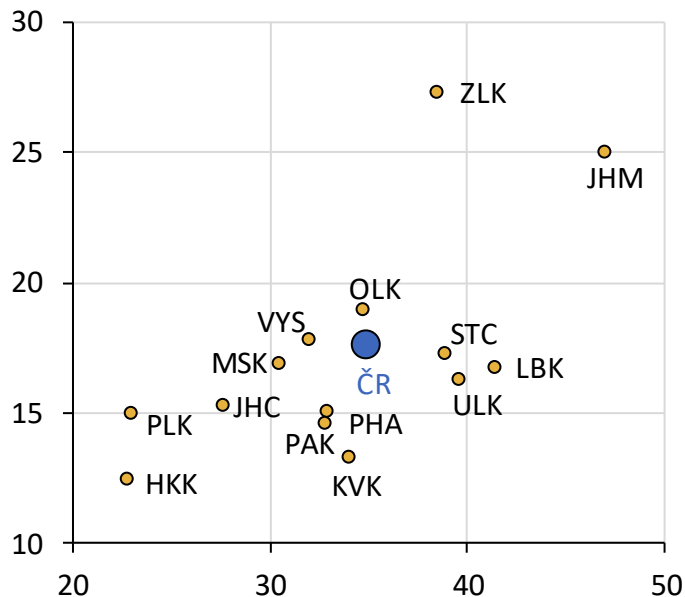
Skládané grafy

- Kumulativní zobrazení více informací



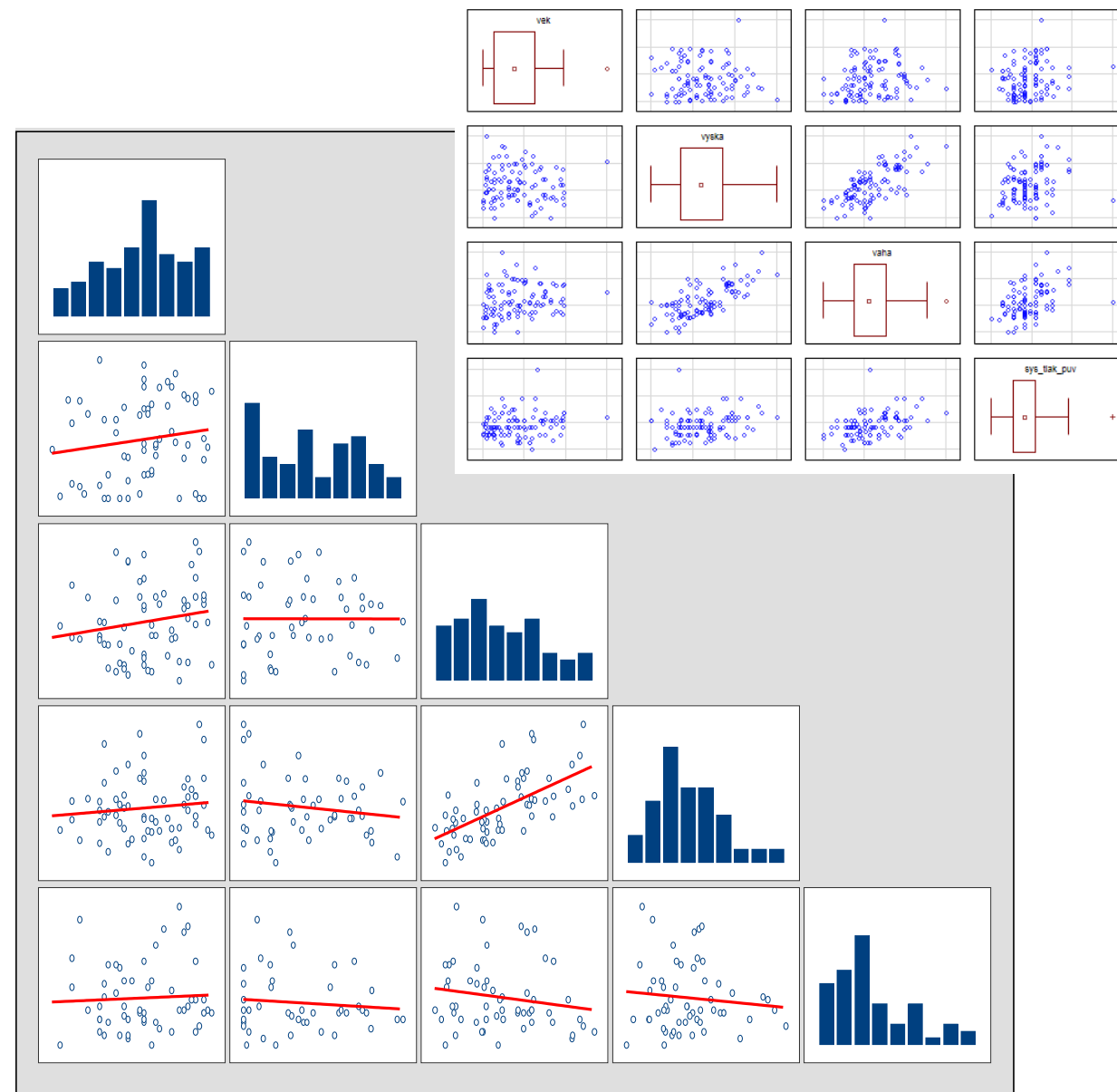
XY graf (scatter plot)

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



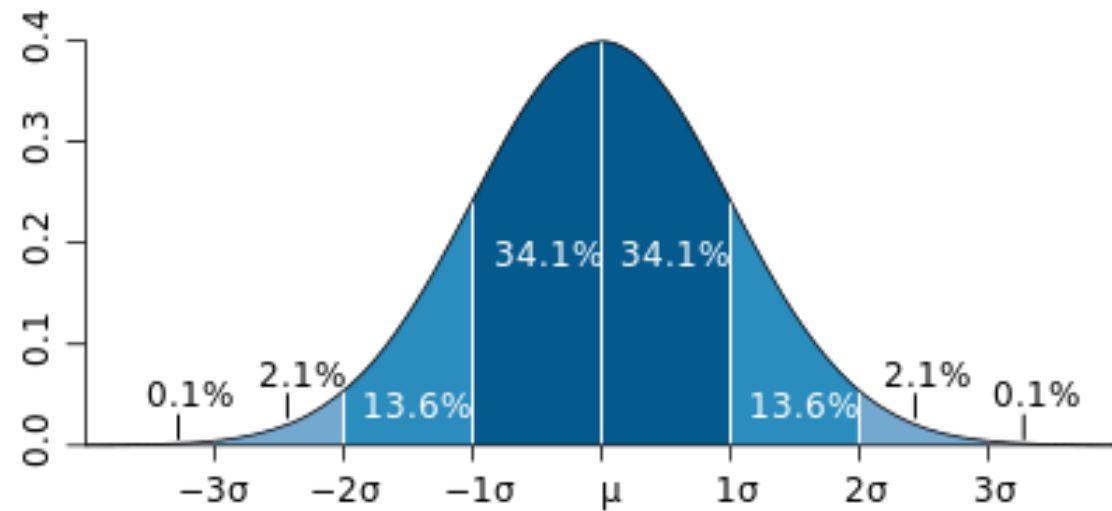
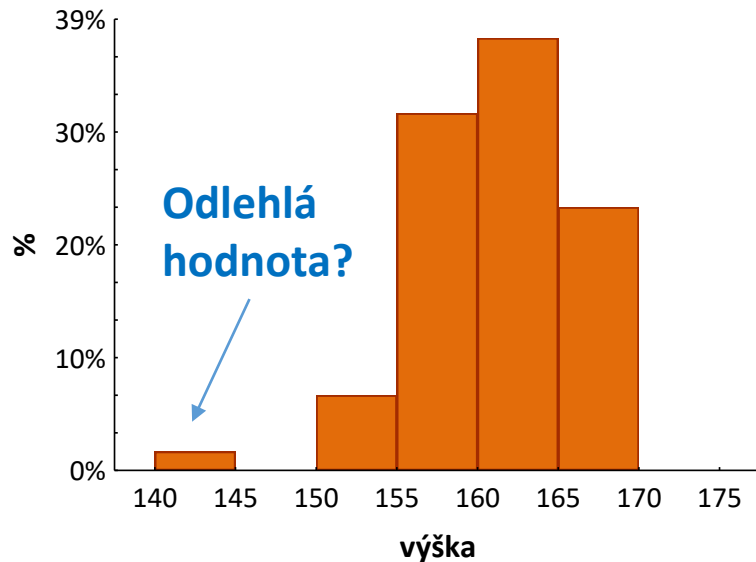
Maticový graf

- 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



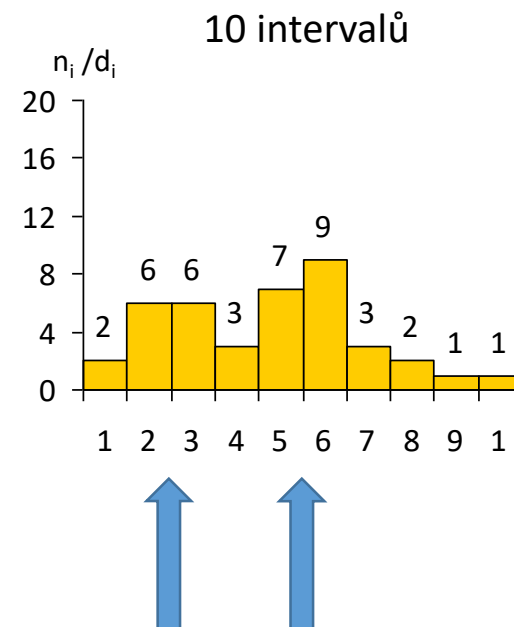
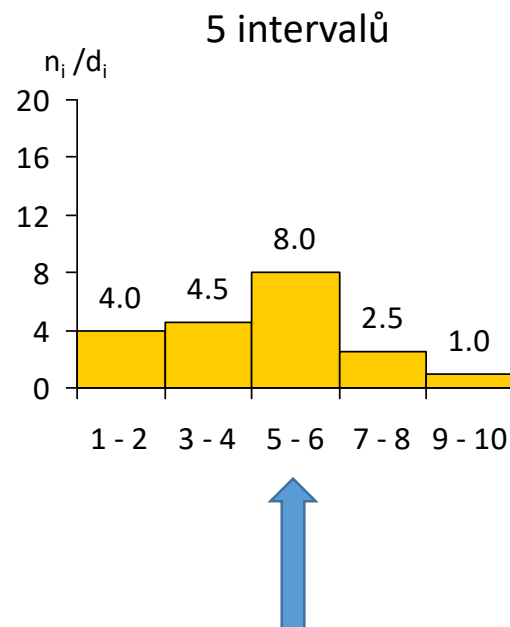
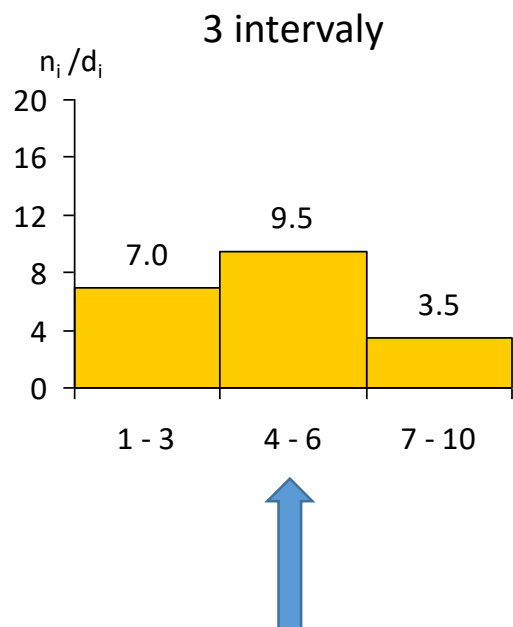
Histogram

- Graf sumarizující rozložení hodnot spojitých proměnných, úzce spjat s teorií statistických rozdělání
- 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



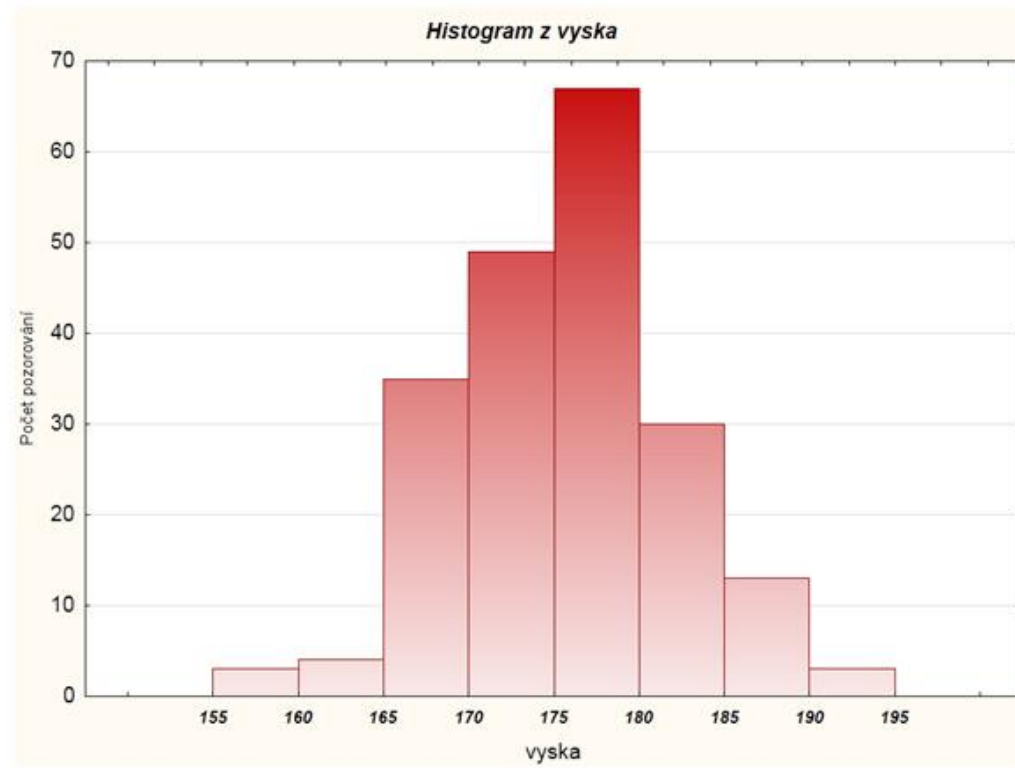
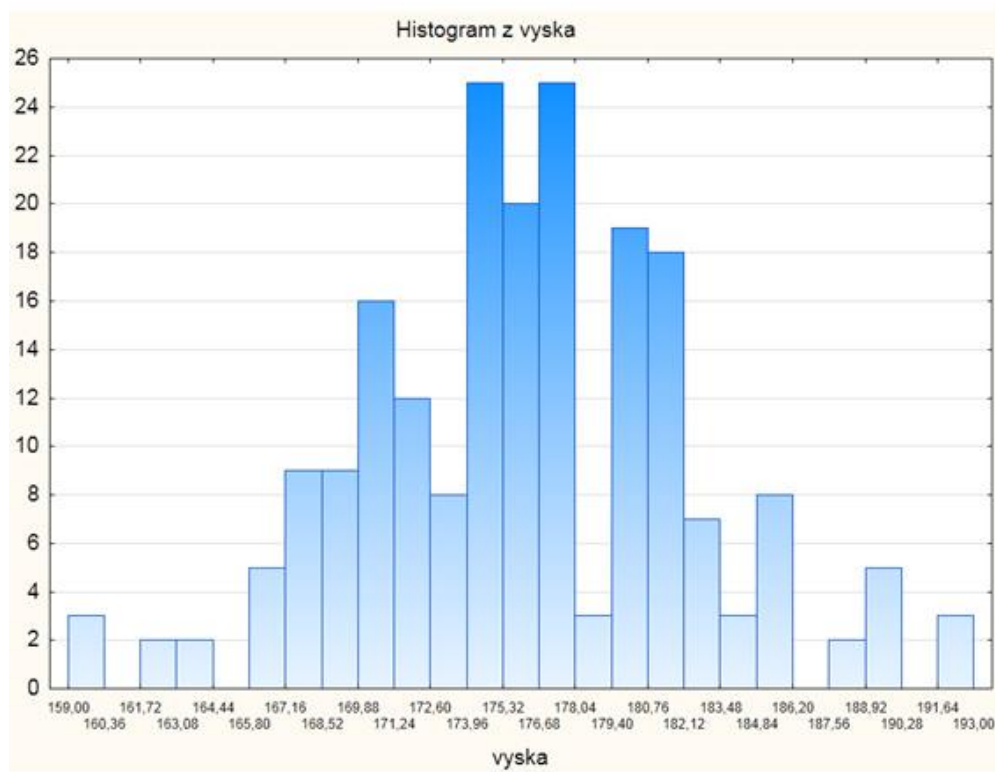
Histogram: vliv kategorizace dat

- 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 roztržštěná.



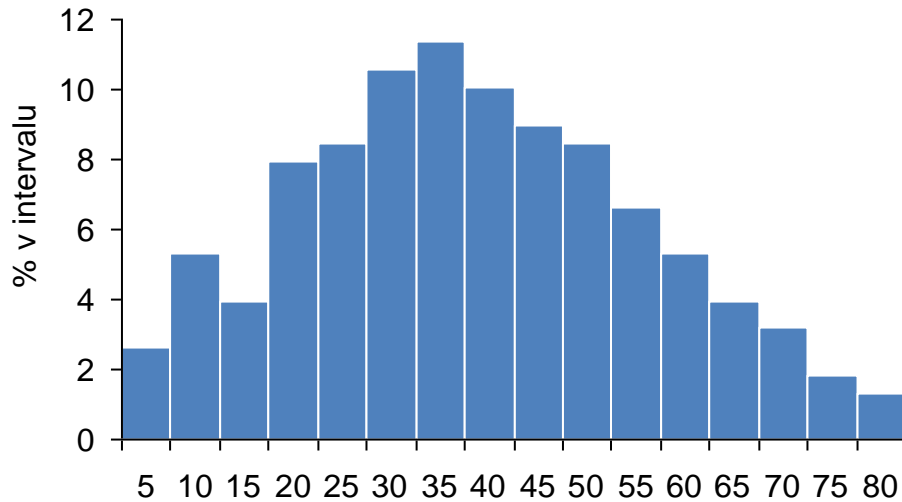
Histogram: vliv kategorizace dat

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

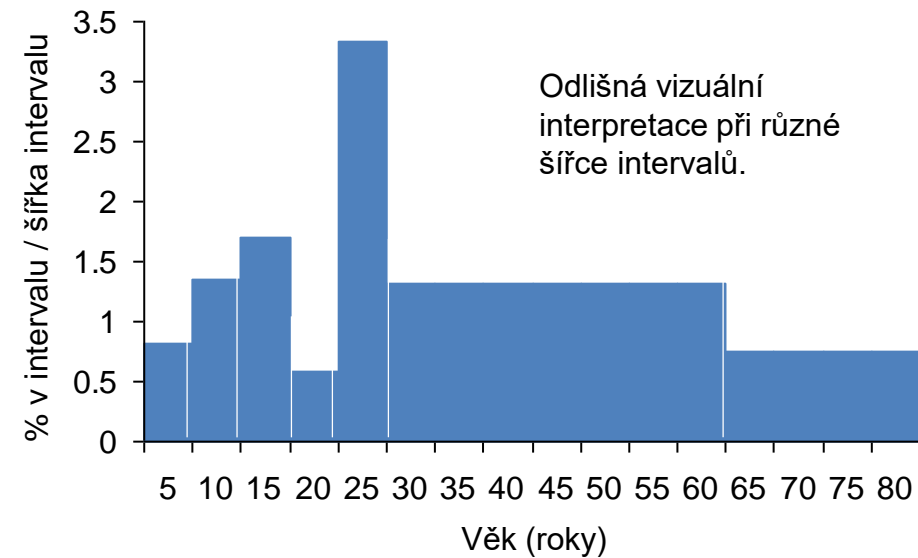
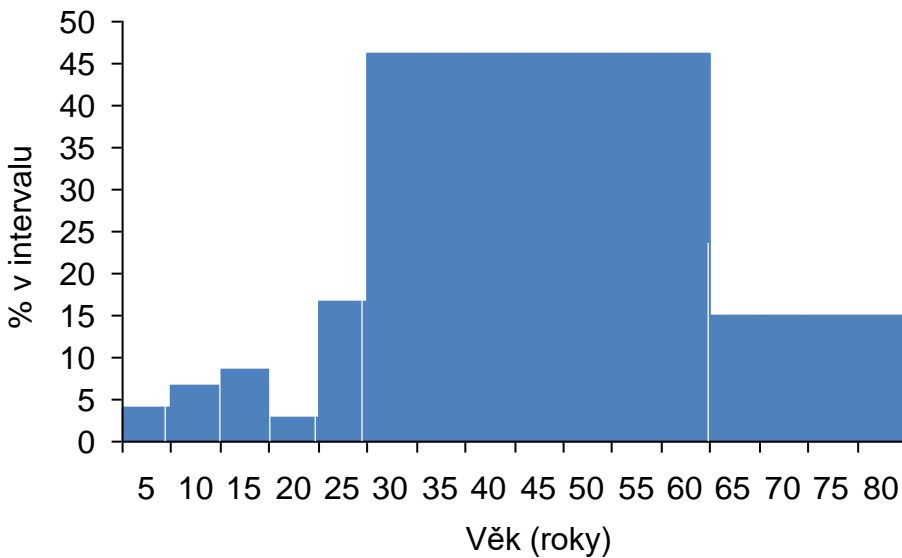
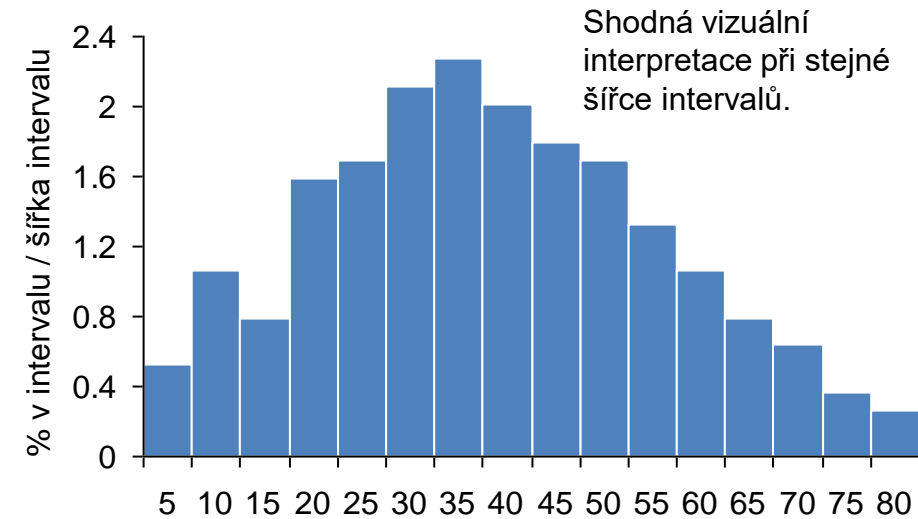


Histogram a sloupcový graf

Sloupcový graf

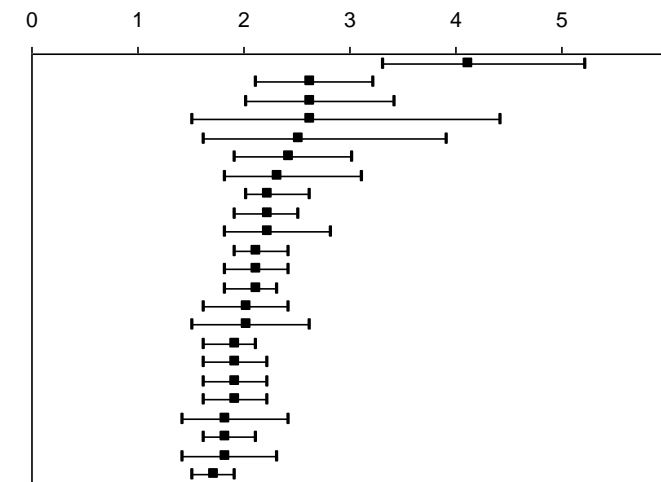
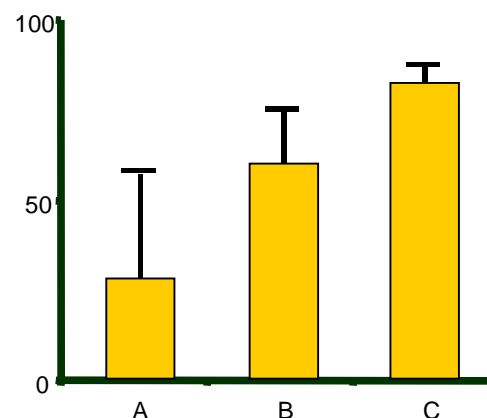
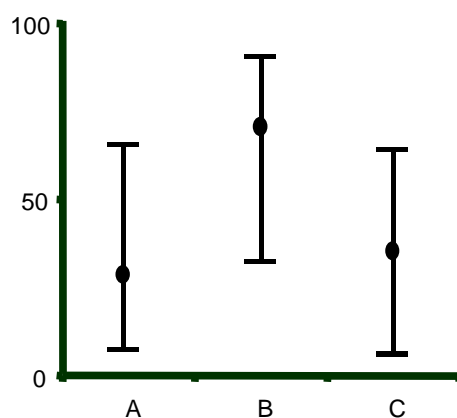
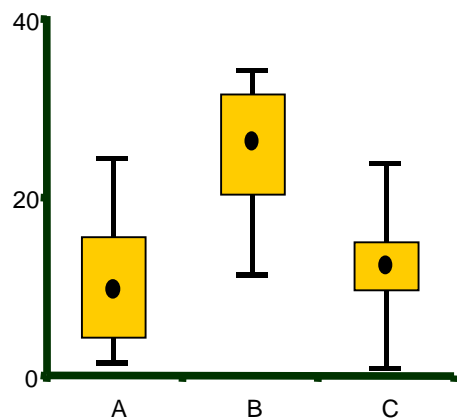


Histogram

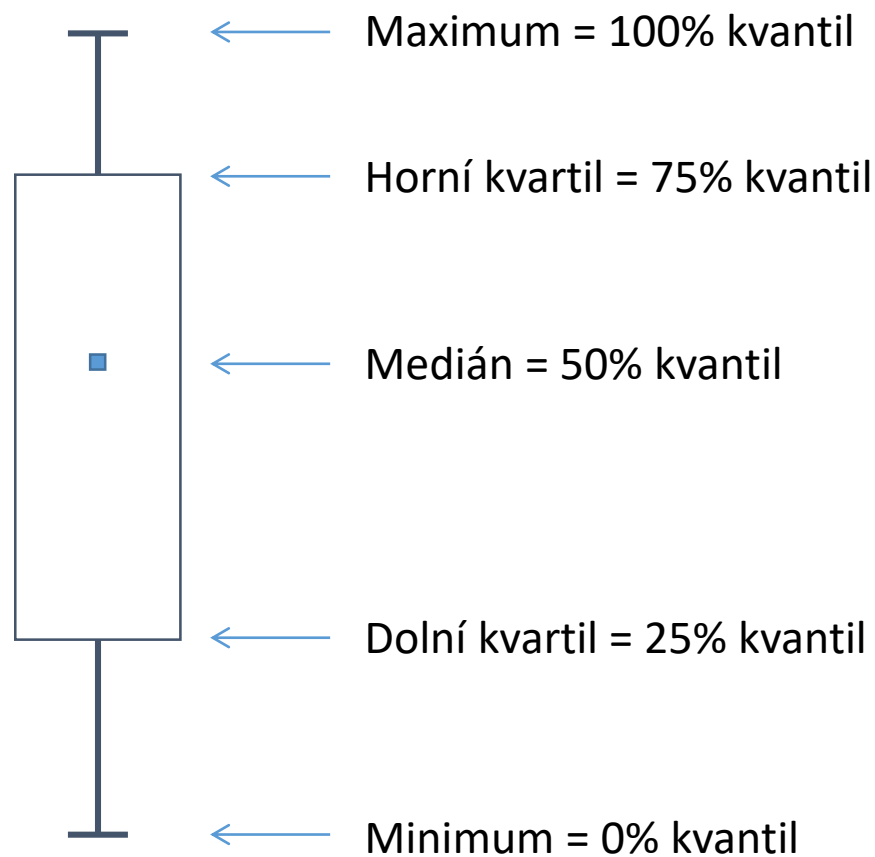


Krabicový graf – box and whisker plot: co to je?

- 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 ratio, risk ratio, hazard ratio atd.)
- Obrovské množství variant



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

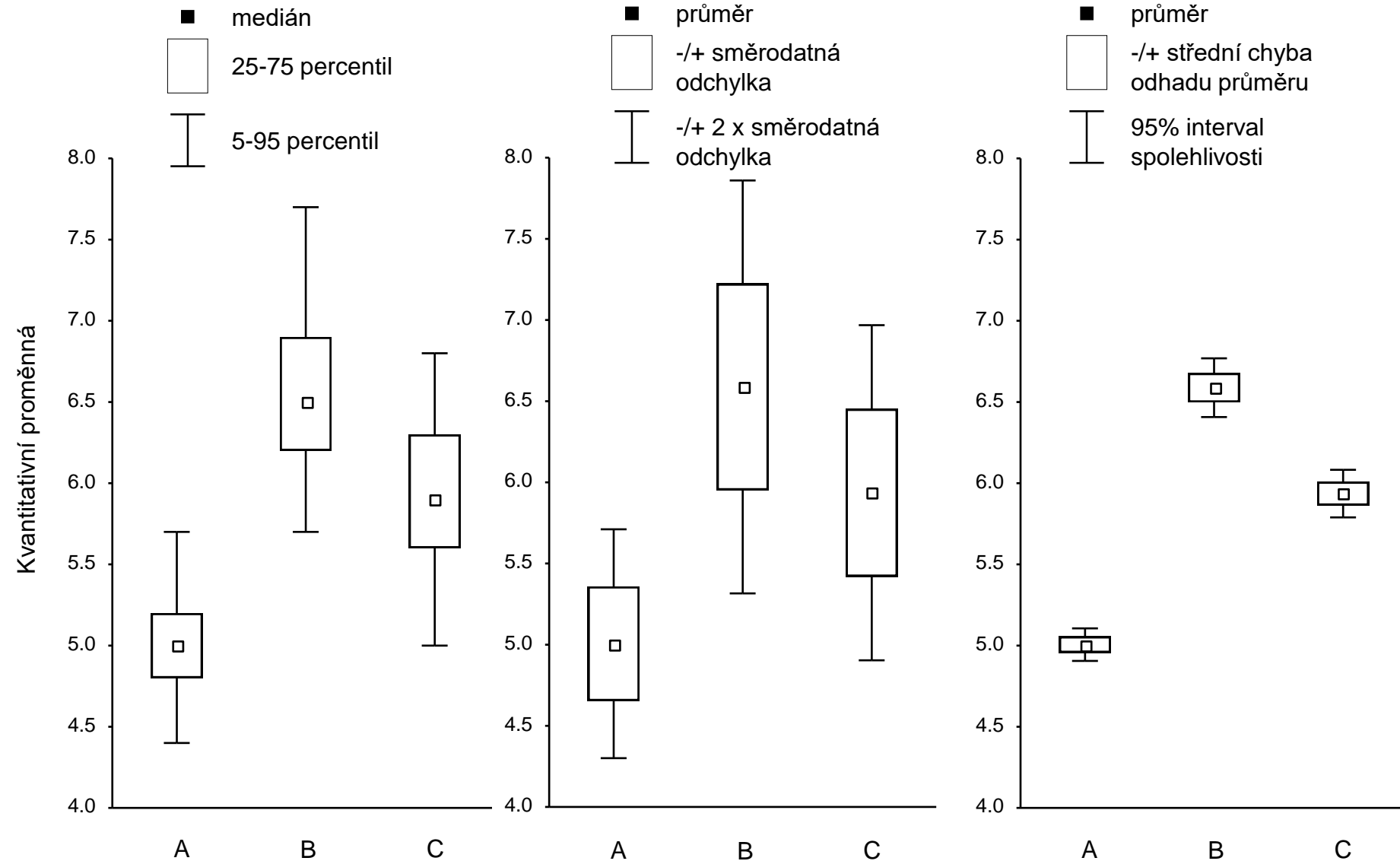


Jednotlivé body grafů mohou obsahovat libovolné popisné statistiky – průměry, směrodatné odchylky, intervaly spolehlivosti, odds ratia, hazard ratia atd.

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

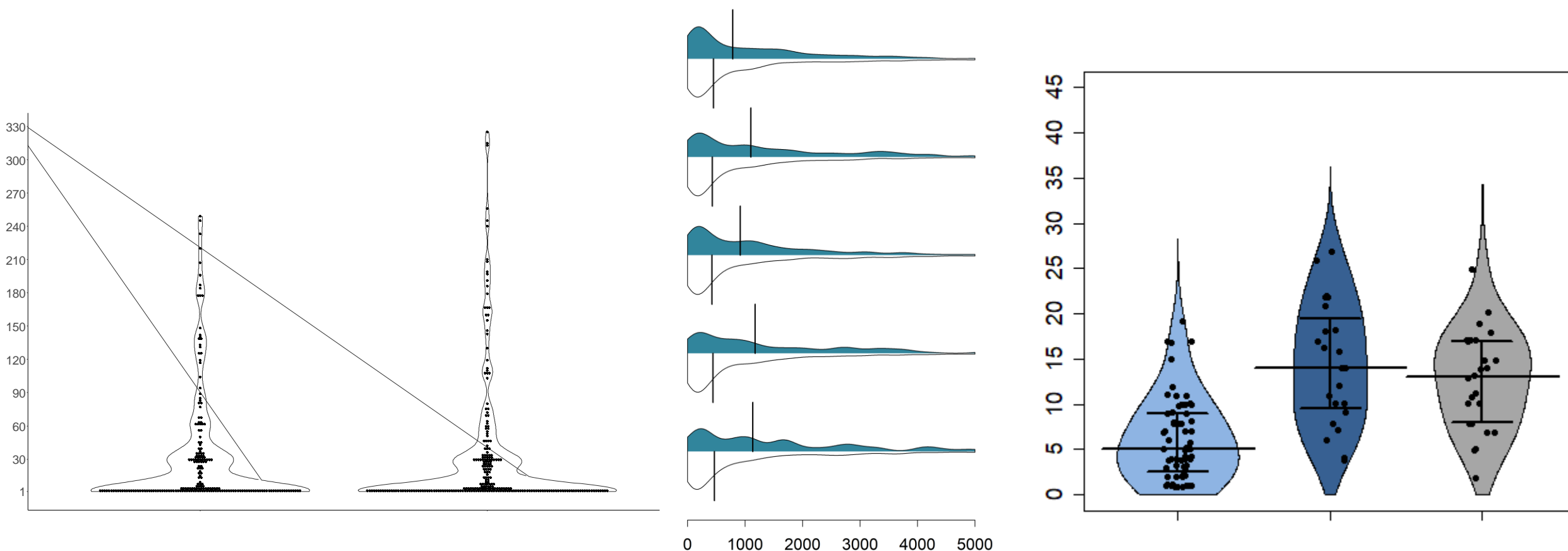
Box and whisker plot a jeho různé varianty I

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



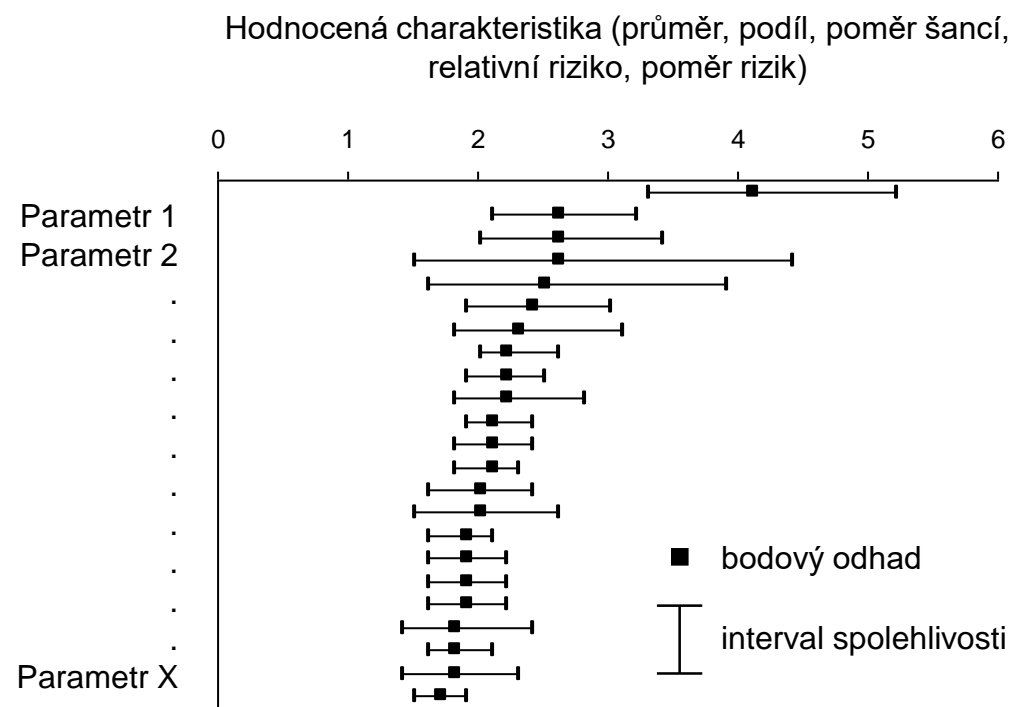
Box and whisker graf a jeho různé varianty II: Violin plot a Beanplot

- 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



Variable	Subgroup	N		Median PFS (months)		HR
		Placebo-Rd	IRd	Placebo-Rd	IRd	
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
	Naïve	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

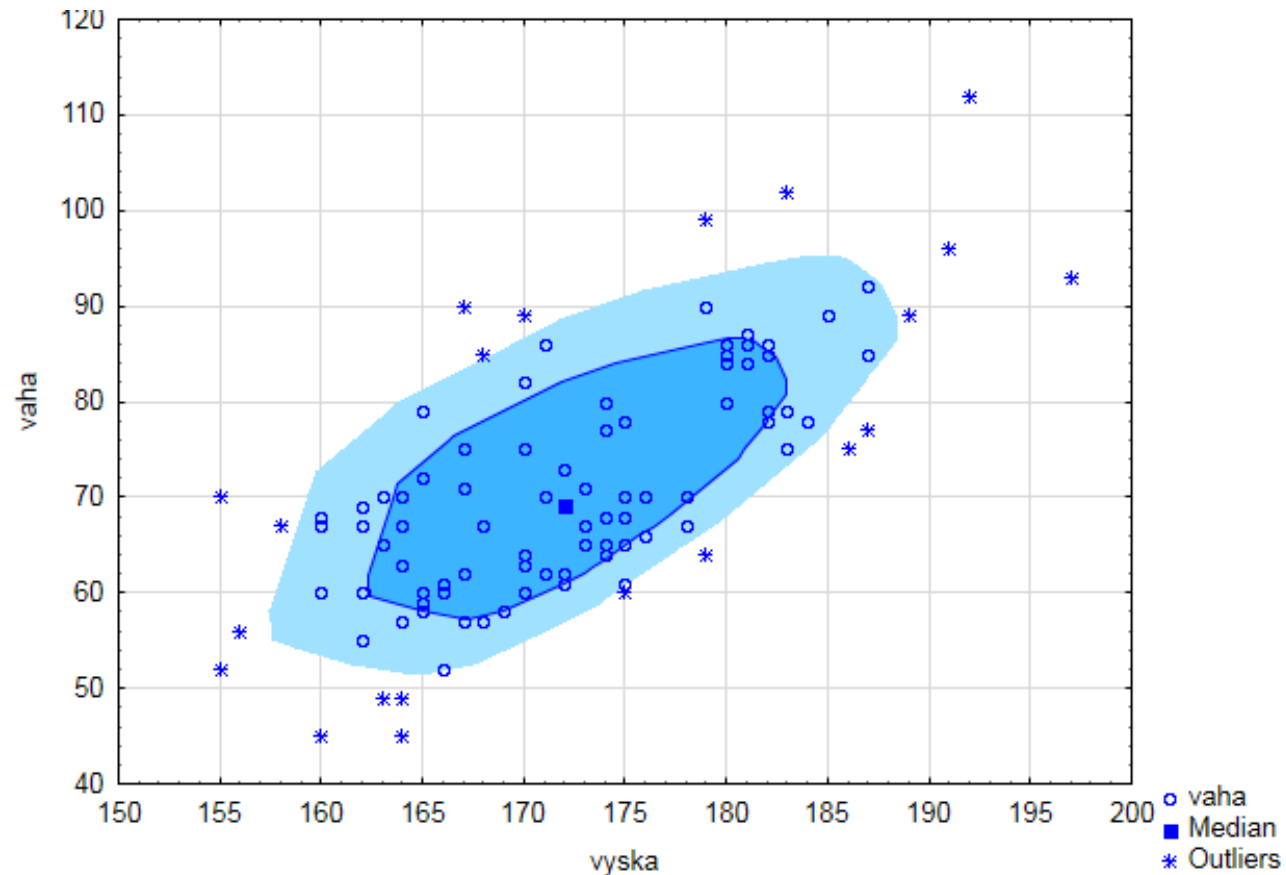
0.250 0.500 1.000 2.000

Favors IRd ← → Favors placebo-Rd

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

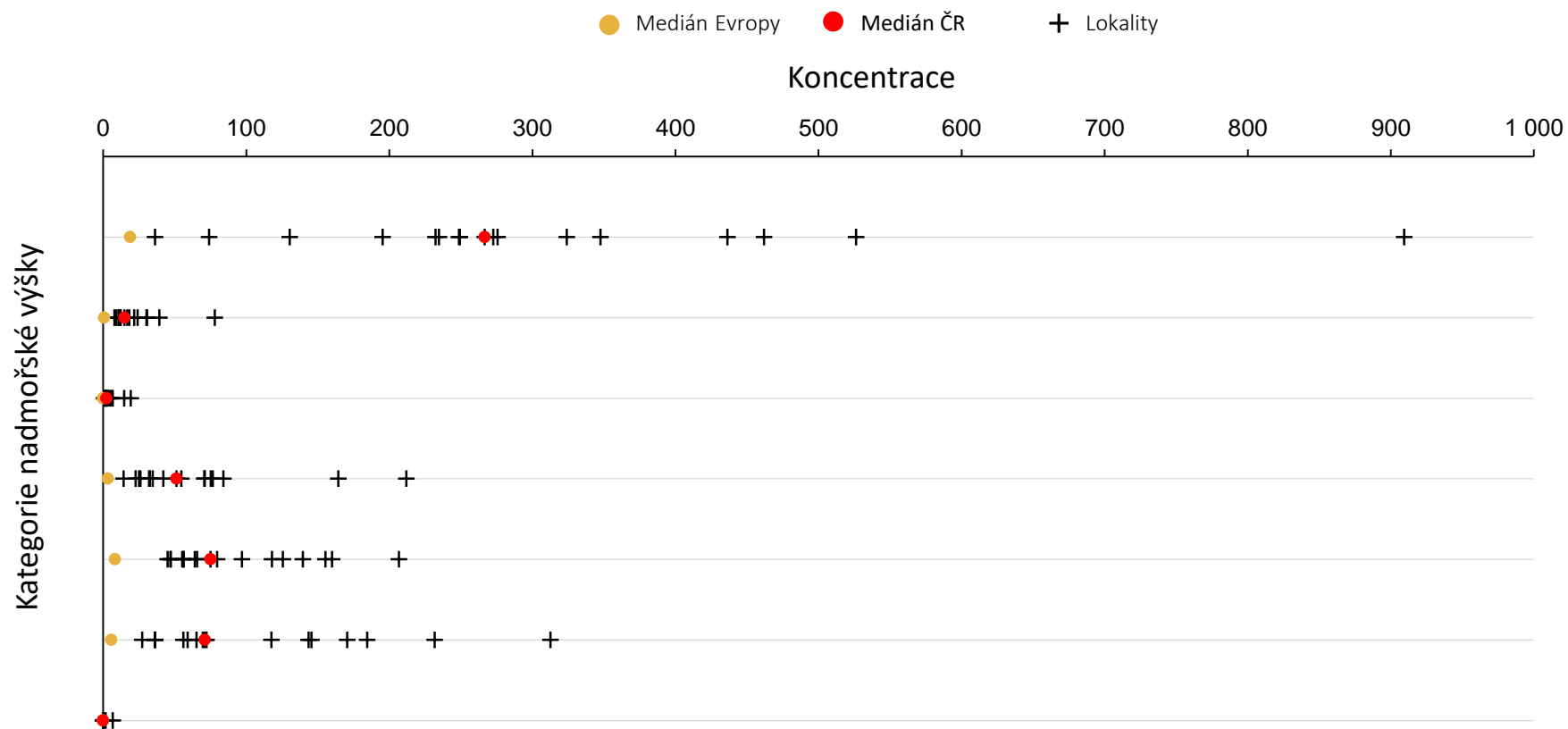
Box and whisker graf a jeho různé varianty IV: Bagplot

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



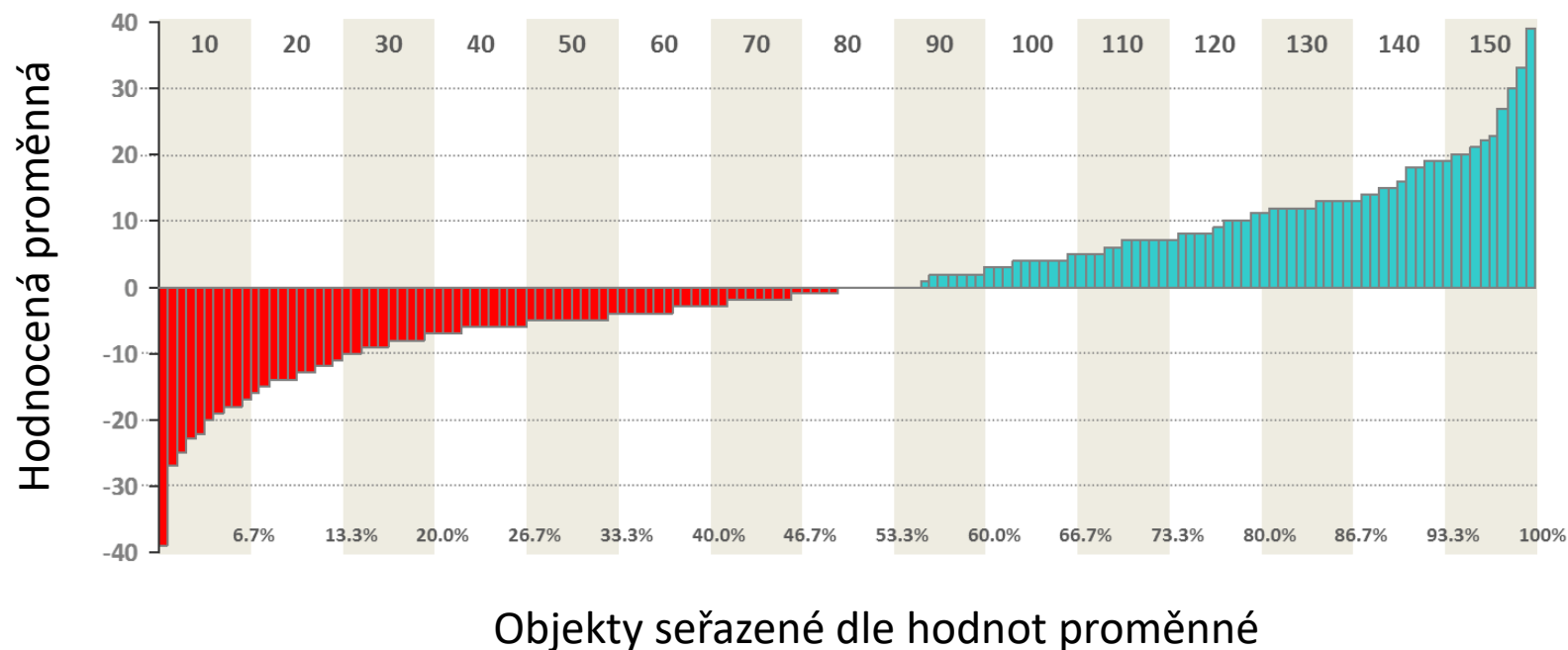
Invenční využití jednoduchých grafů: Korálkový graf

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



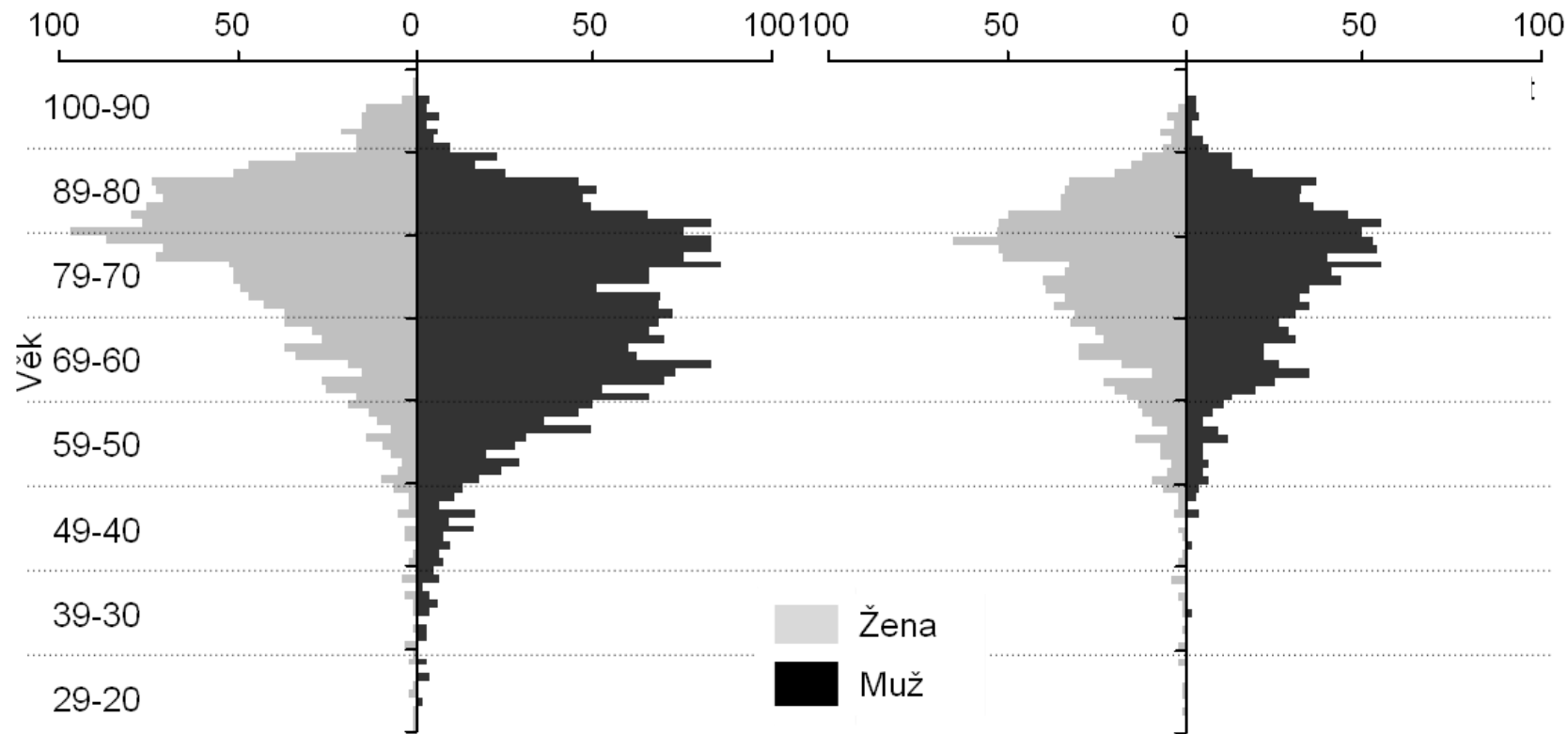
Invenční využití jednoduchých grafů: Waterfall plot

- 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.



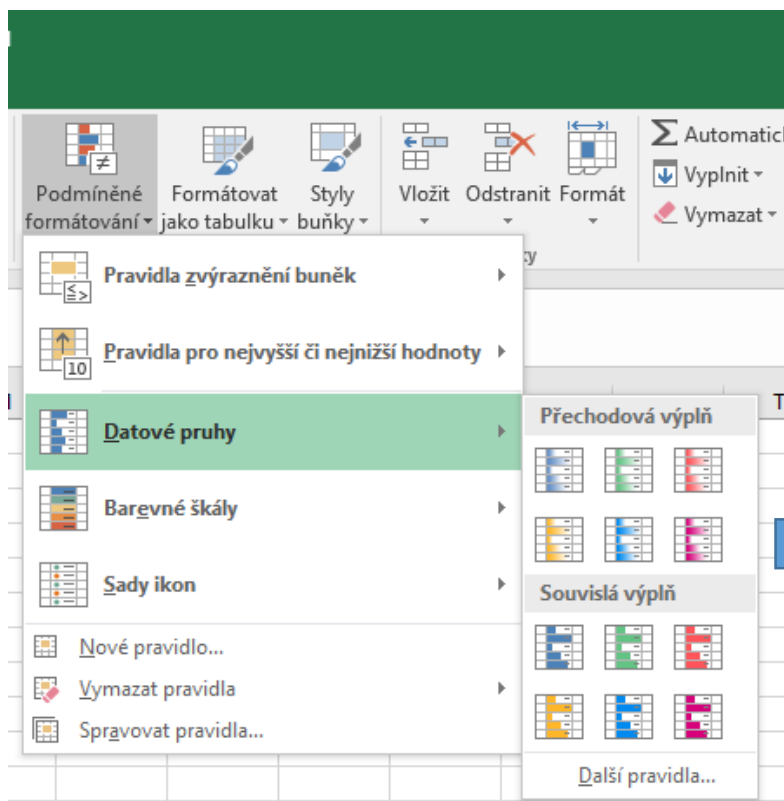
Invenční využití jednoduchých grafů: Demografická pyramida

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



Excel – podmíněné formátování jako grafy

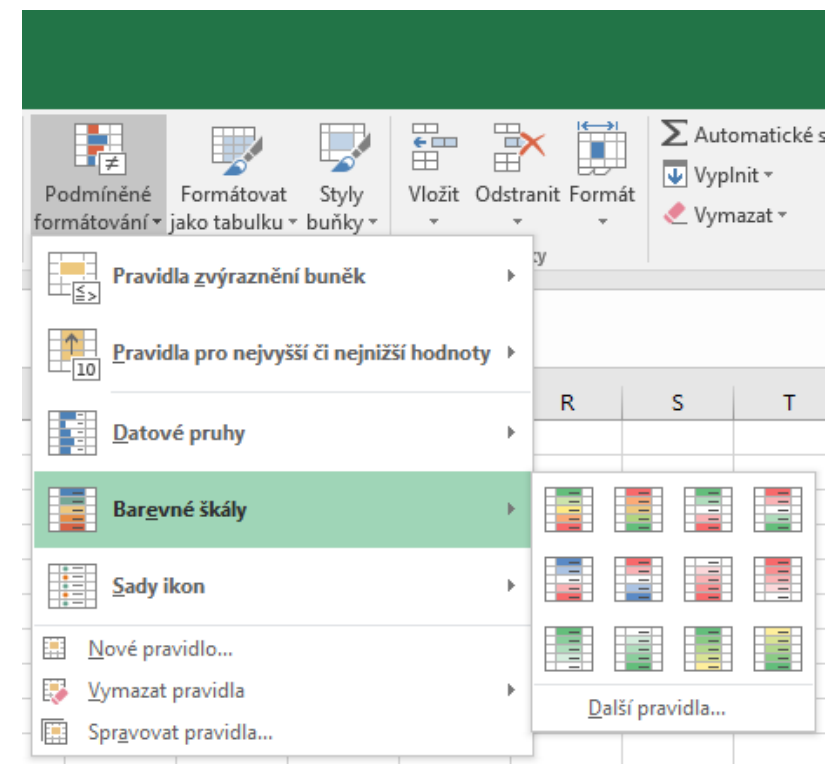
- Pro zpřehlednění excelových tabulek je možné využít grafické prvky v jeho buňkách
- Datové pruhy a barevné škály



The screenshot shows an Excel spreadsheet with a table of data. The table has columns M through U and rows 1 through 6. The values in the table are as follows:

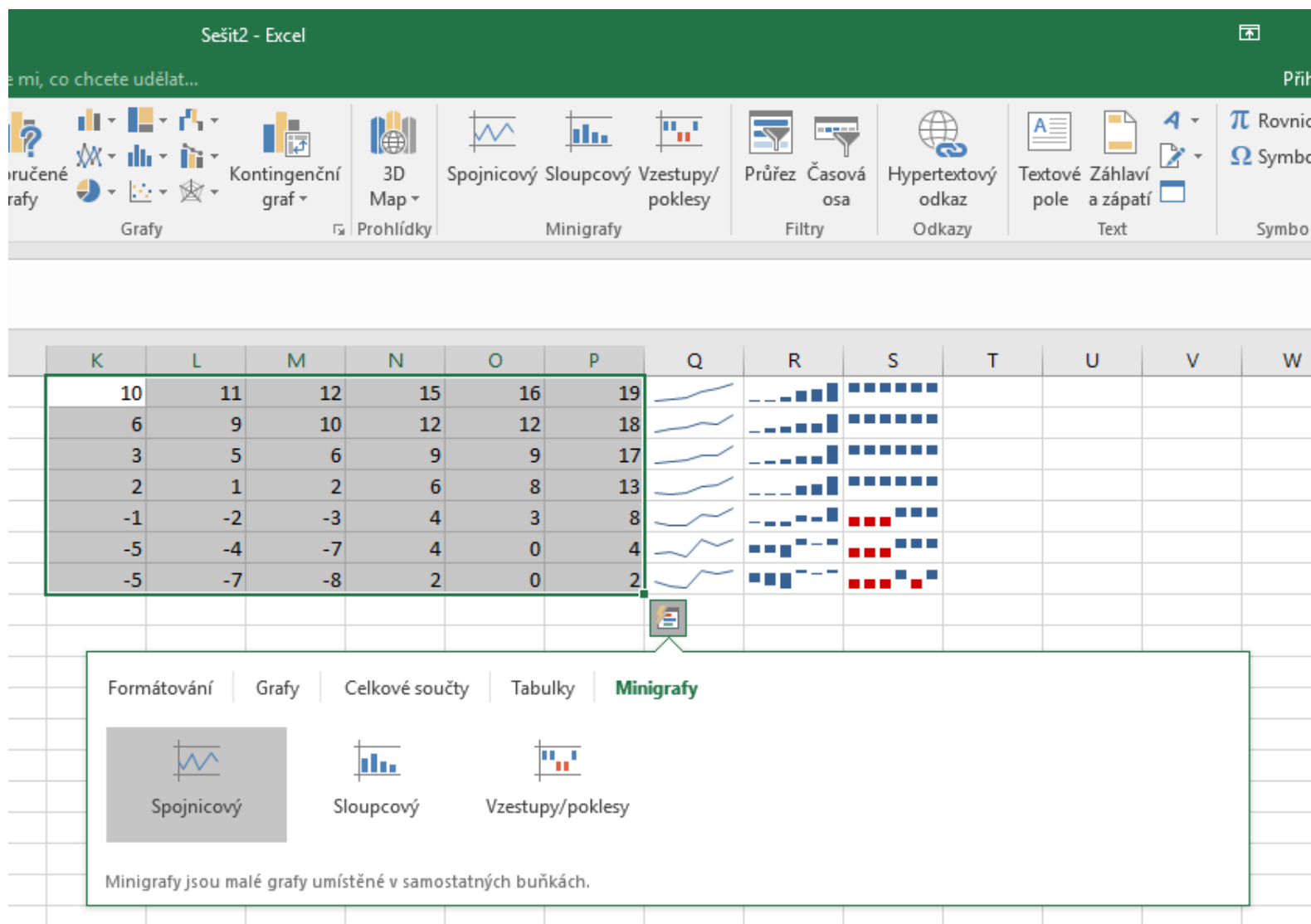
	M	N	O	P	Q	R	S	T	U
1	10		1	2	3	4	5	6	
2	15		2	3	4	5	6	7	
3	1		3	4	5	6	7	8	
4	5		4	5	6	7	8	9	
5	6		5	6	7	8	9	10	
6	7		6	7	8	9	10	11	
7	1								
8	22								

Blue arrows point from the 'Datové pruhy' menu in the left screenshot to the data bars in this table, and from the 'Barevné škály' menu in the right screenshot to the color scale in this table.



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

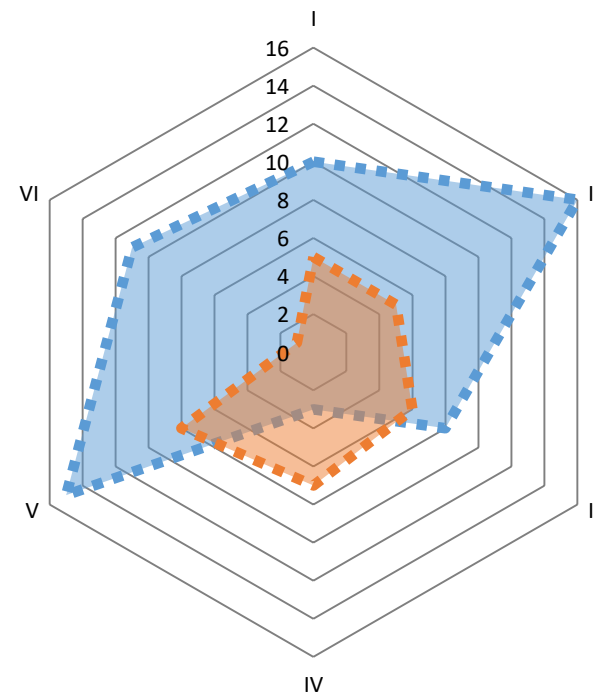
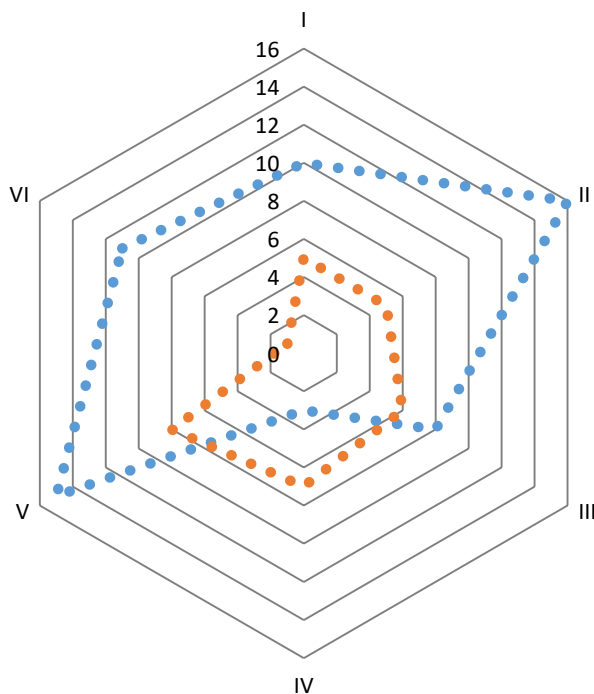
- Druh 3D grafu – osy tvoří dvě proměnné, barva třetí proměnnou
- Lze vytvořit v excelu pomocí podmíněného formátování
- Často ve vícerozměrné analýze pro vizualizaci asociačních matic

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%

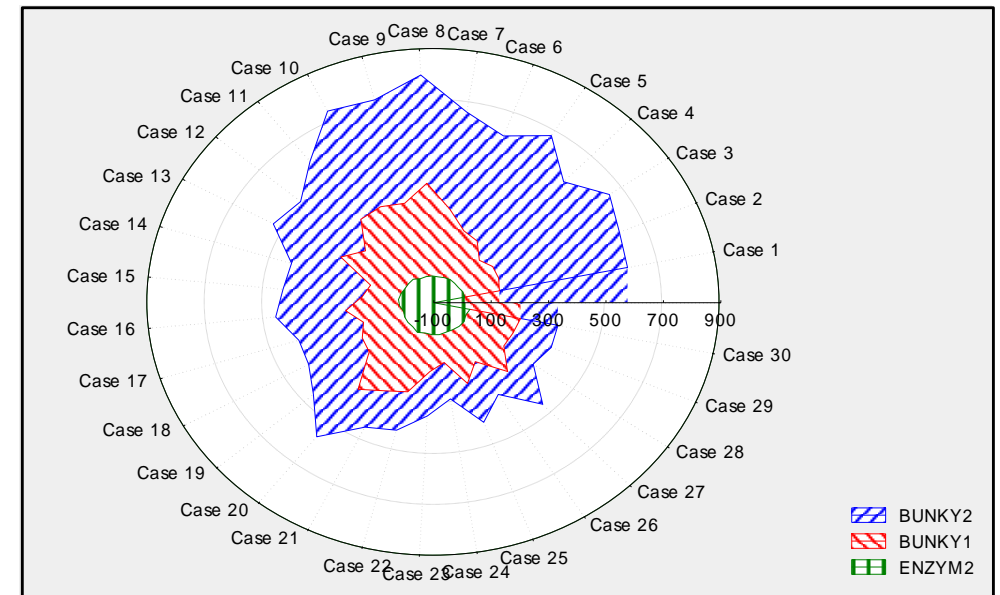
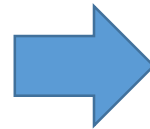
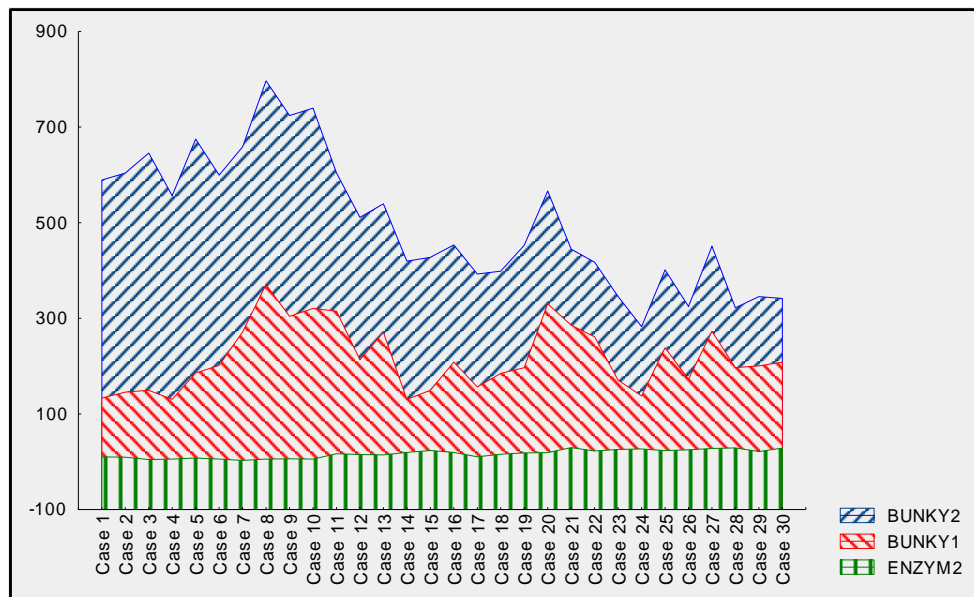
Pavoučí / paprskové grafy

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



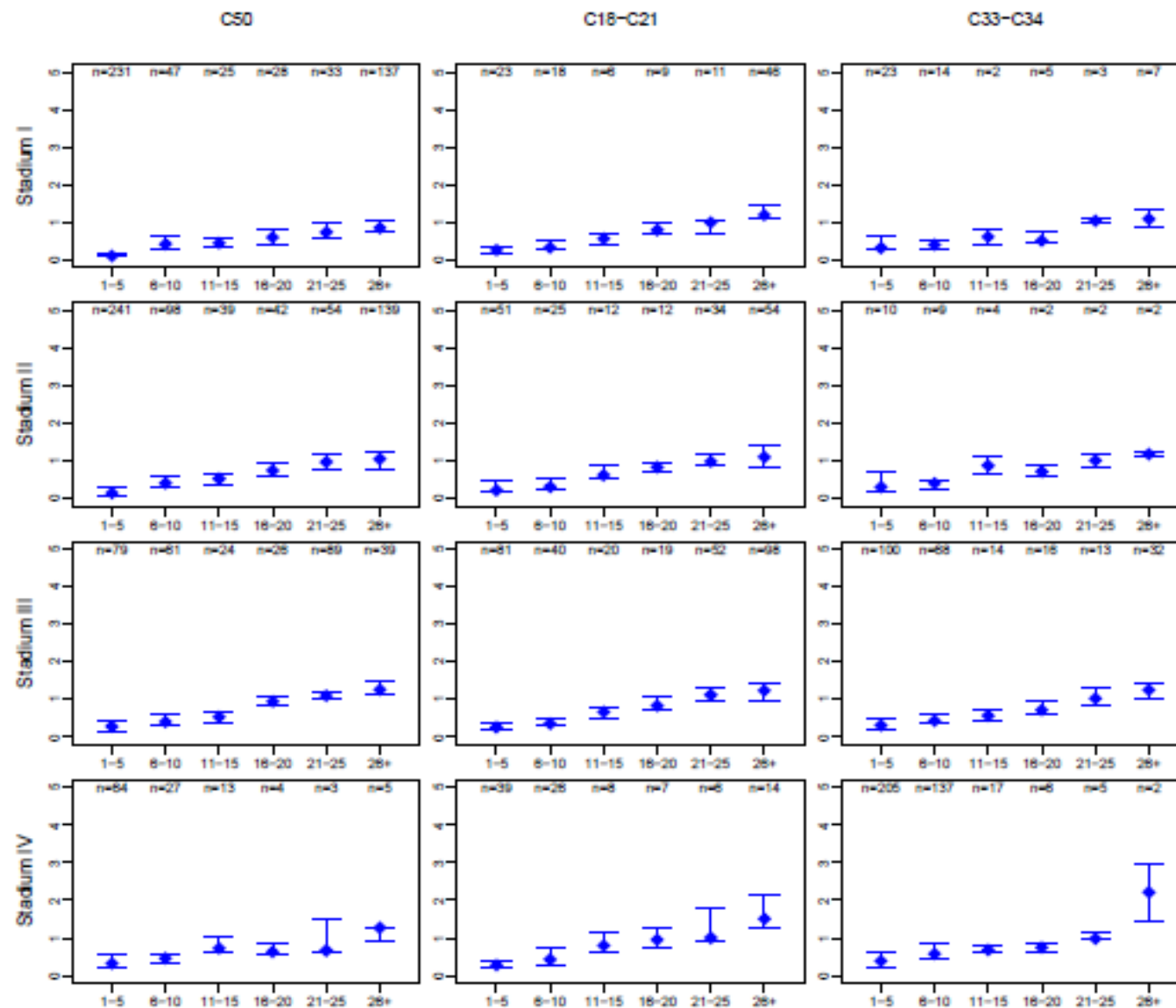
Polární graf

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



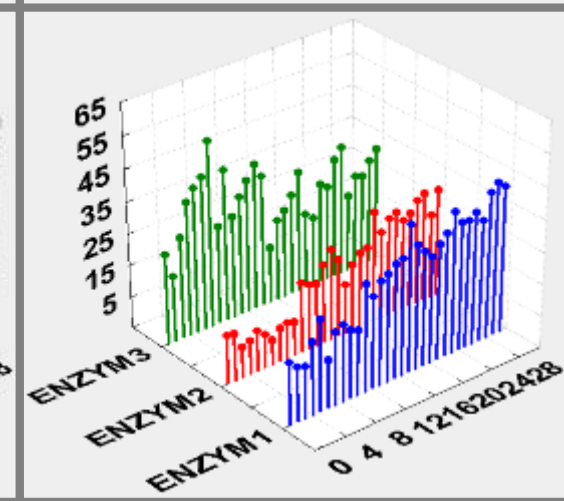
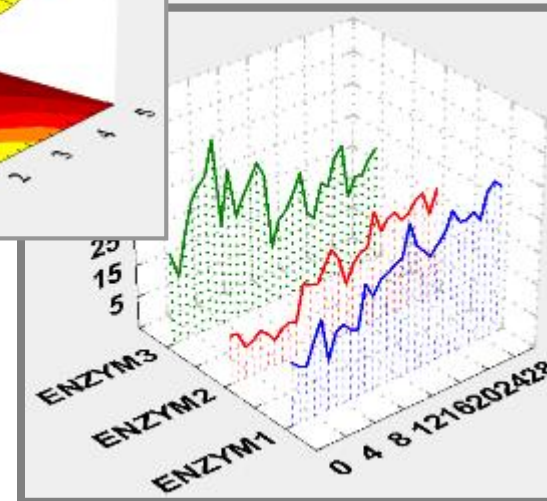
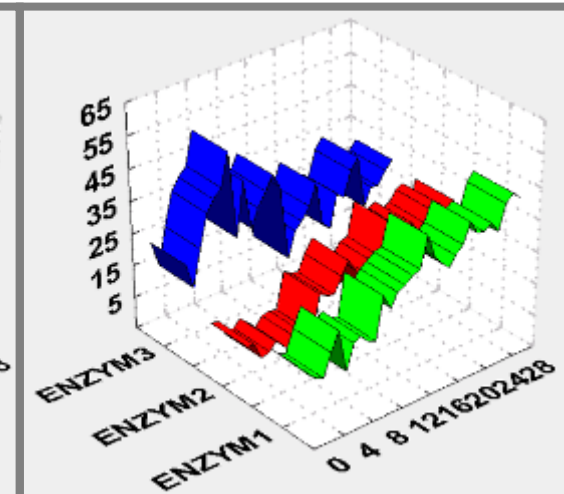
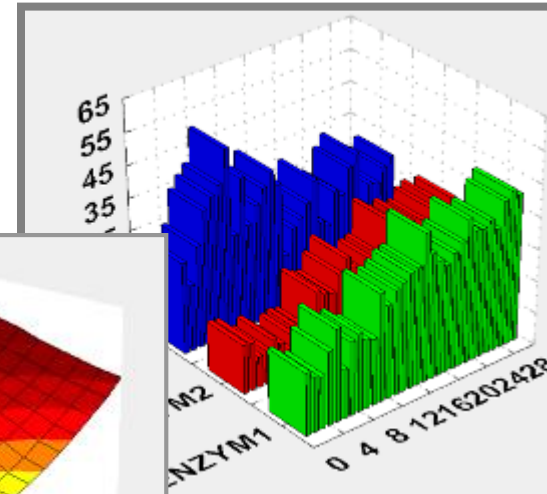
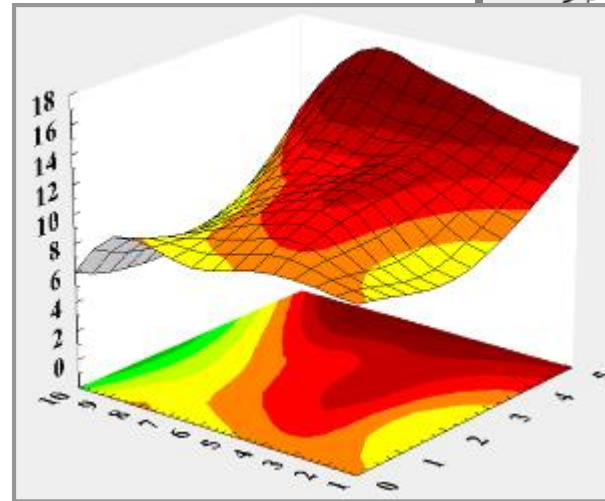
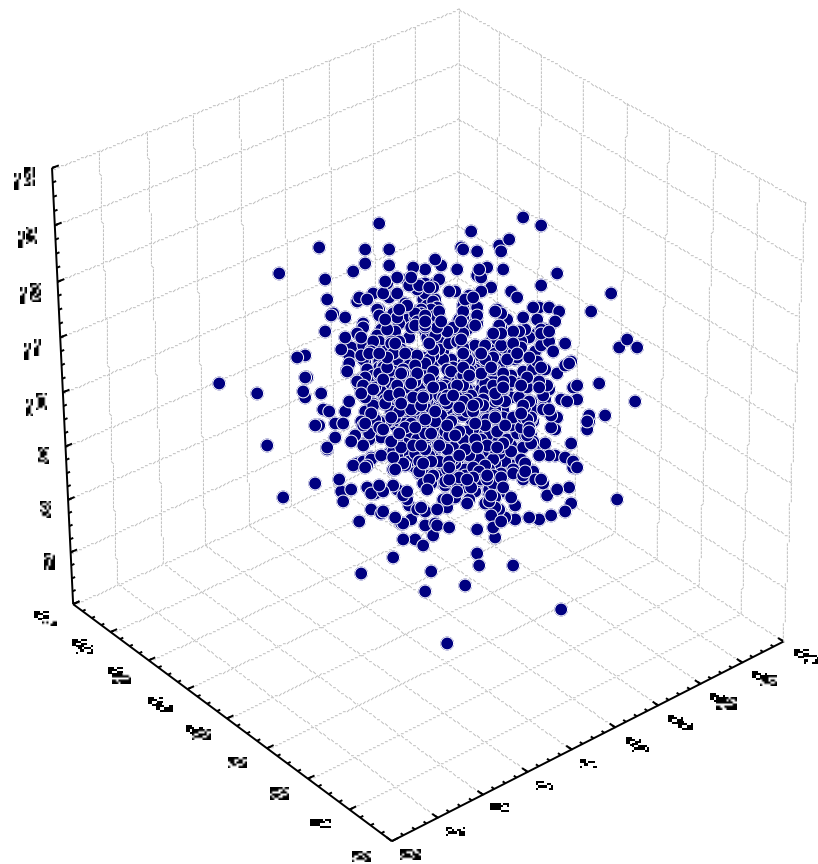
Grafické tabule

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



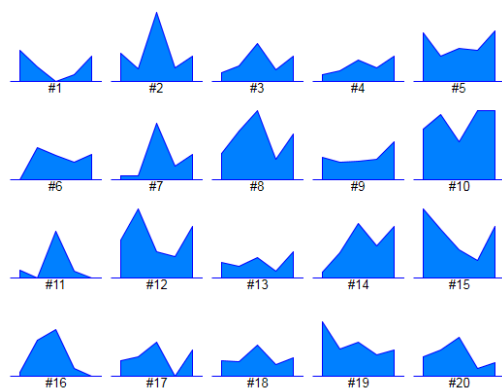
3D grafy

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

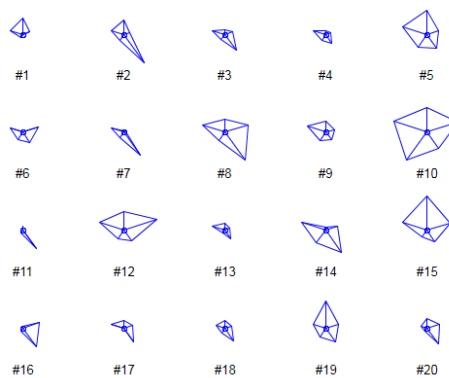


Chernoffovy tváře (ikonové grafy)

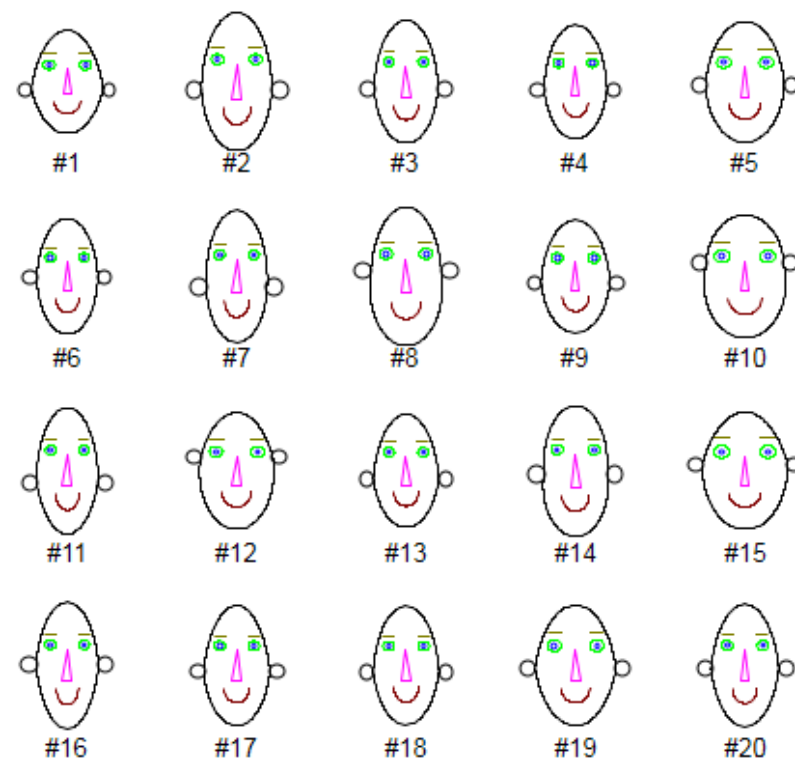
- Jednotlivé proměnné jsou zobrazeny jako rysy tváře
- Patří mezi tzv. ikonové grafy
 - hodnoty znaků znázorněny jako geometrické útvary či symboly
 - každému objektu (subjektu) odpovídá jeden obrazec složený z těchto geometrických útvarů či symbolů
 - umožní vizuálně porovnat, které objekty (subjekty)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



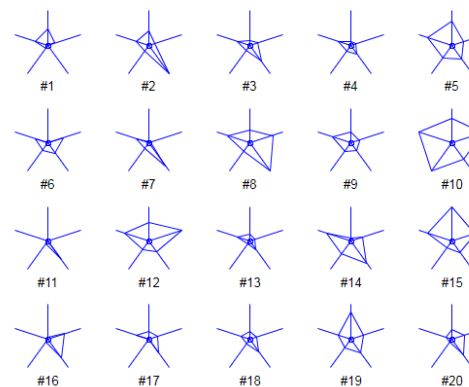
Left to right:
vek
cel_cholesterol
vaha
sys_tlak
dia_tlak



Clockwise:
vek
cel_cholesterol
vaha
sys_tlak
dia_tlak



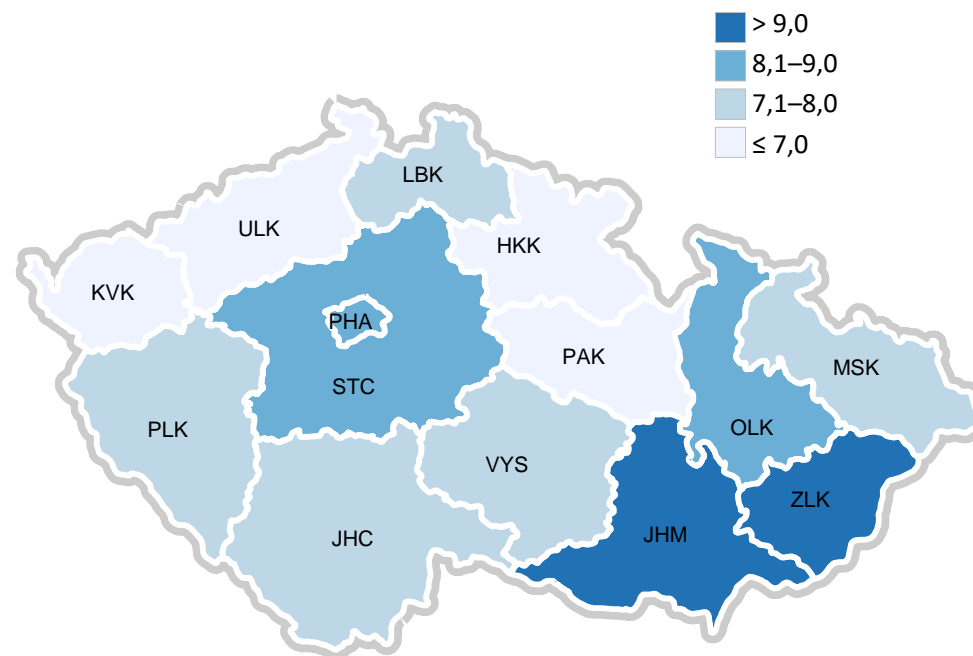
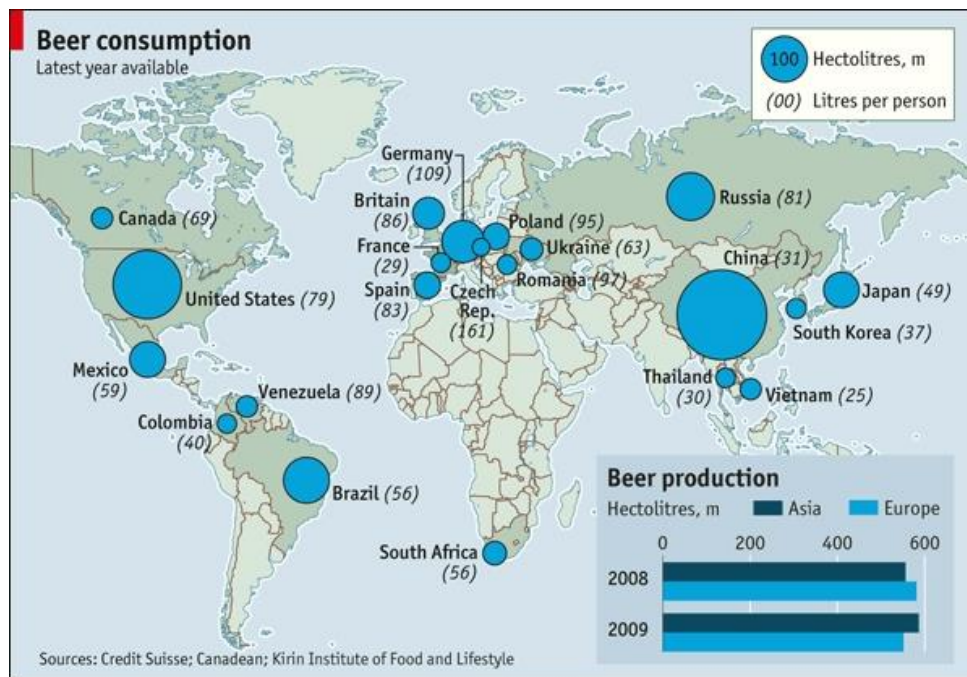
— face/w = vek
— ear/lev = cel_cholesterol
— halfface/h = vaha
— upface/ecc = sys_tlak
— loface/ecc = dia_tlak



Clockwise:
vek
cel_cholesterol
vaha
sys_tlak
dia_tlak

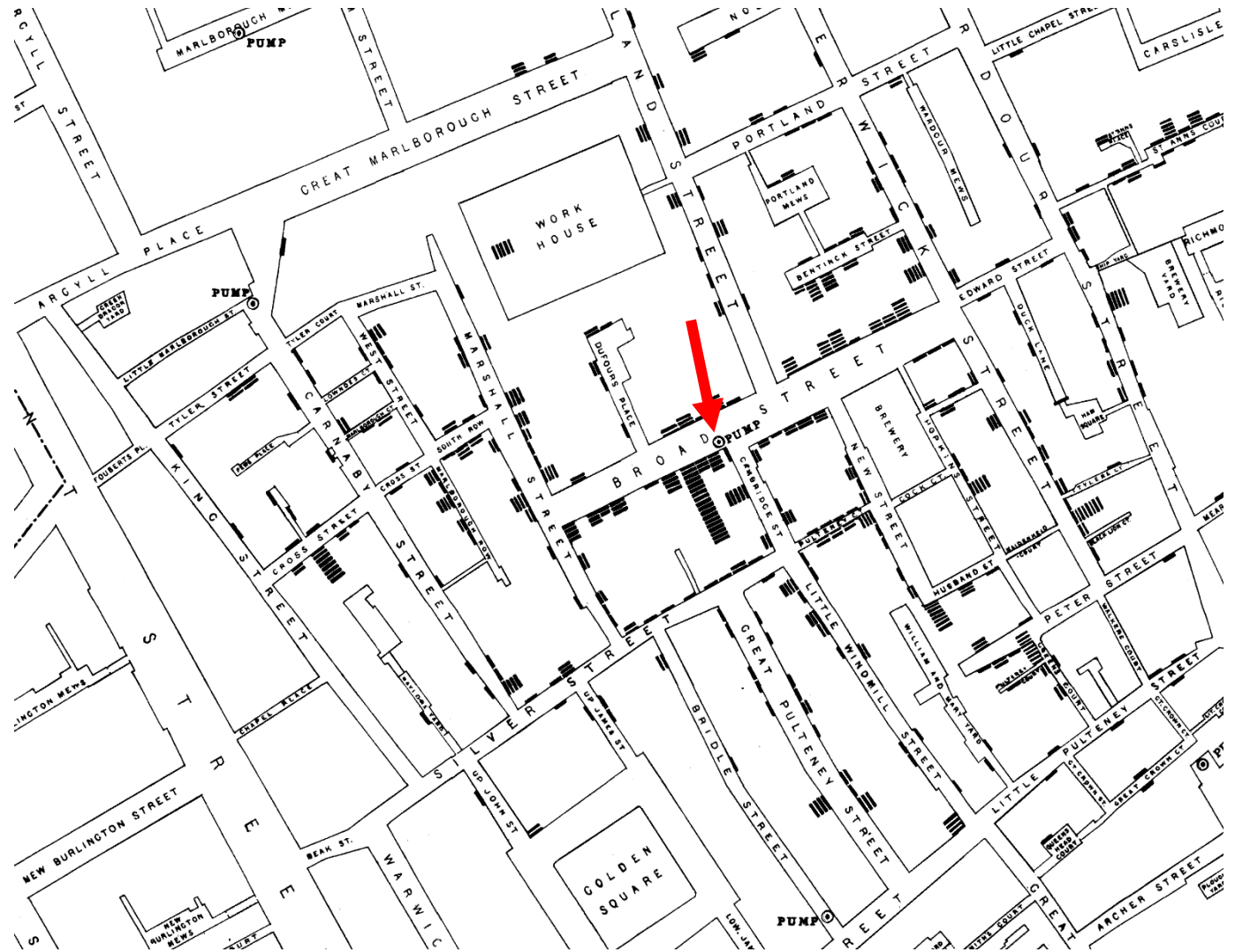
Mapy jsou také grafy

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

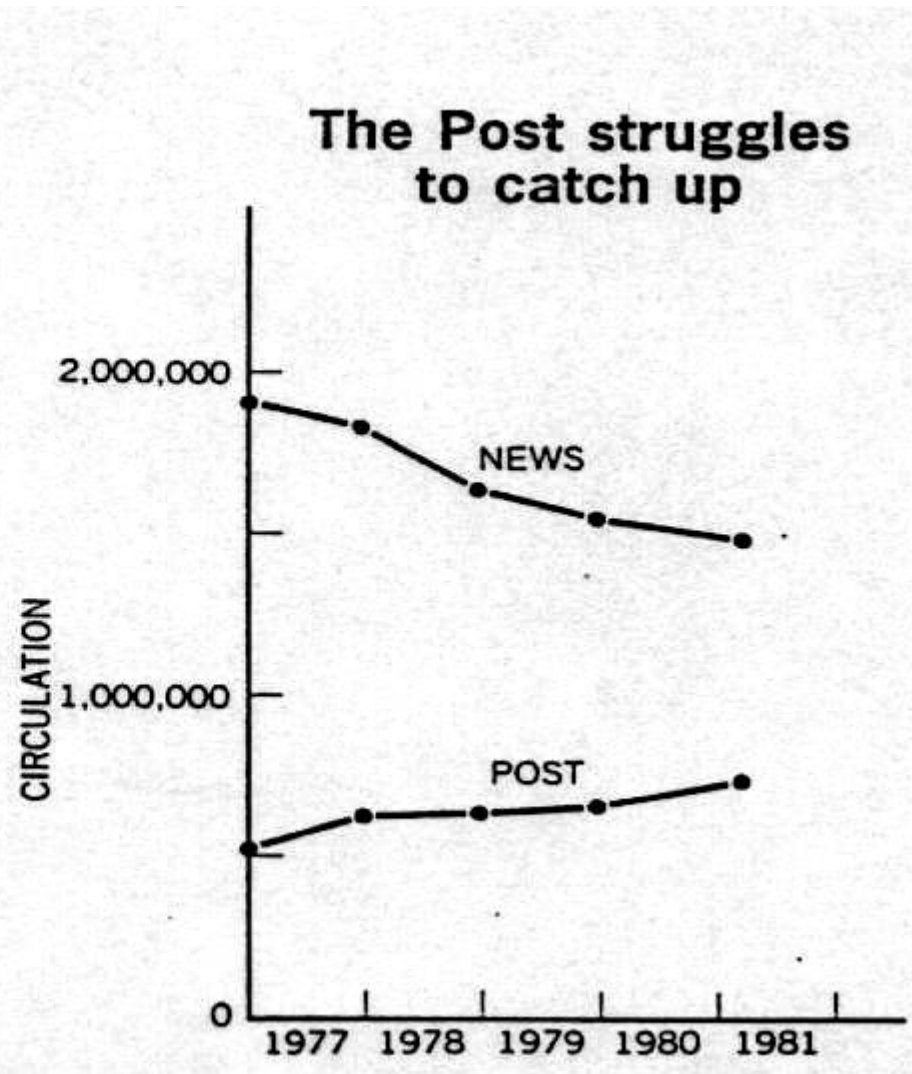
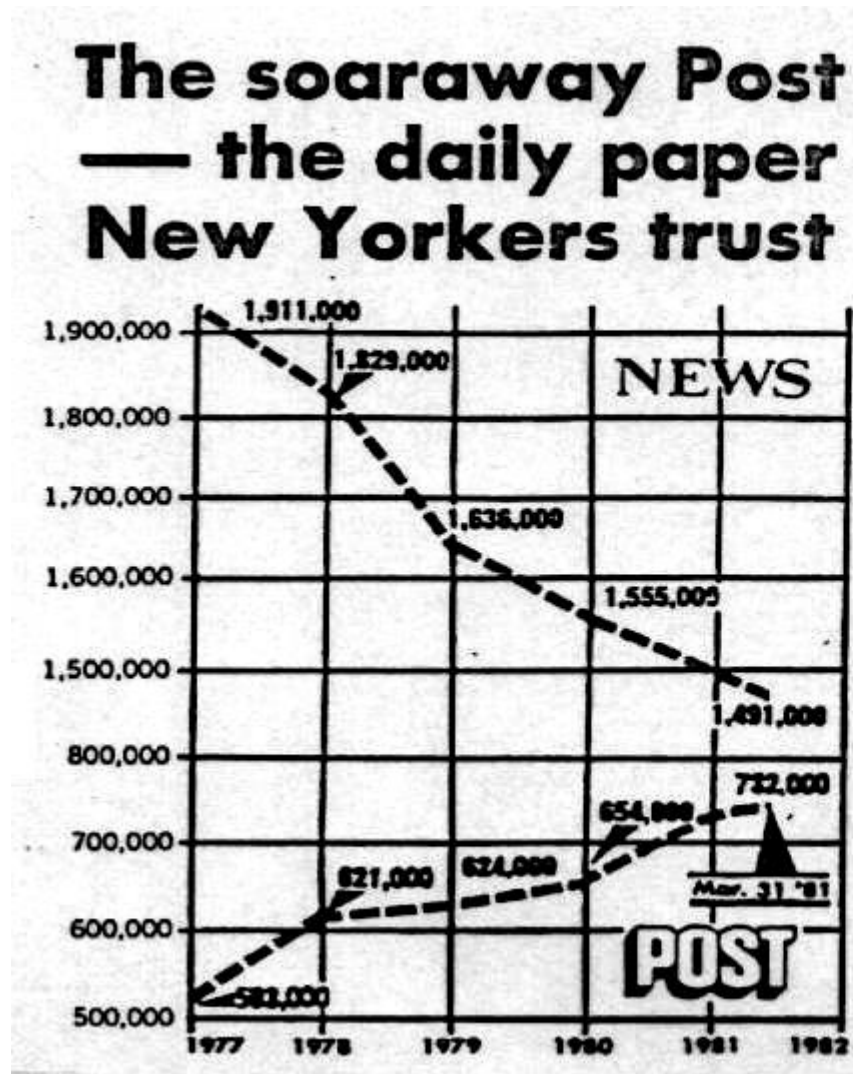


Slavné mapy: John Snow – cholera v Londýně

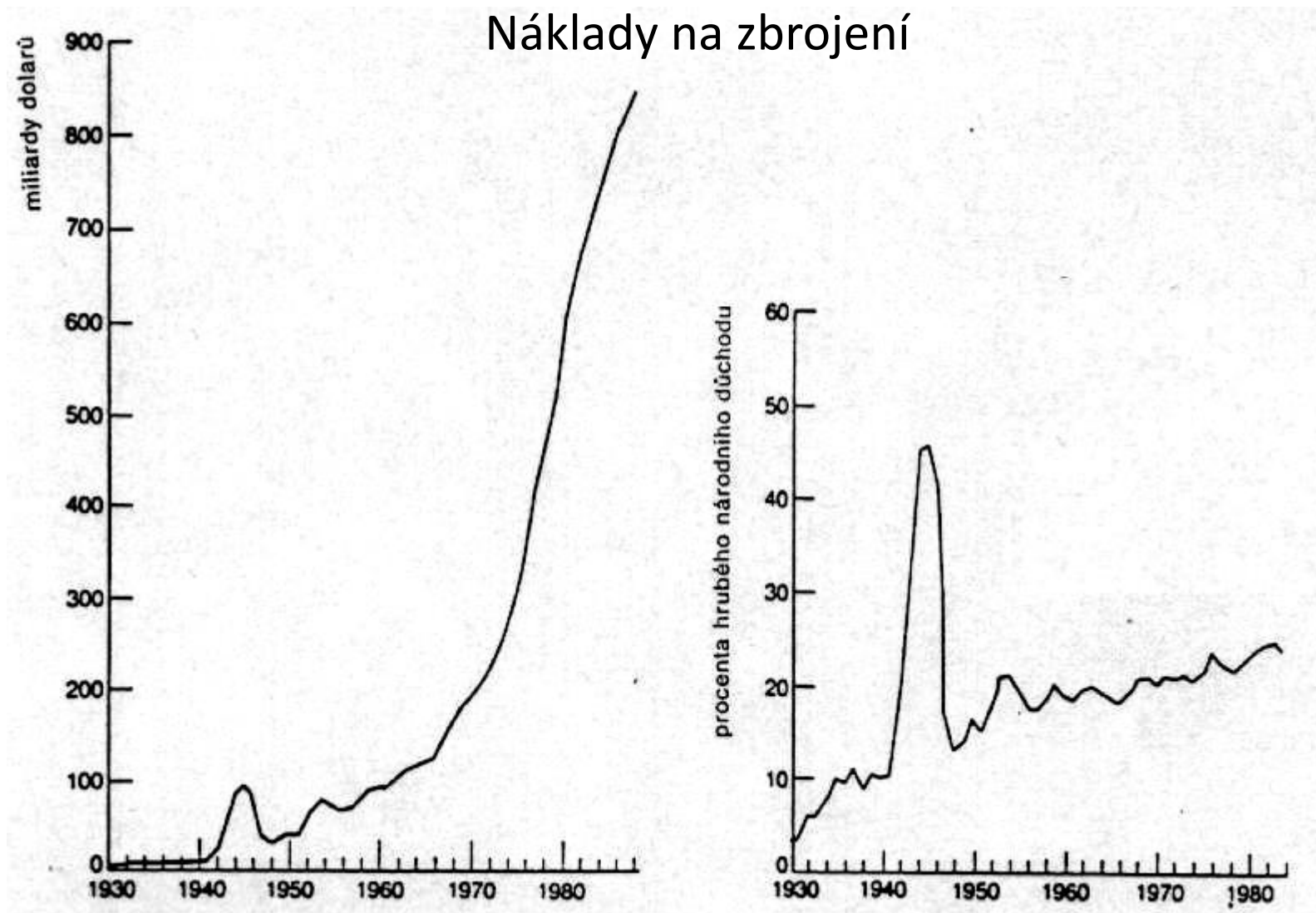
- 1854 Broad Street cholera outbreak
- Počty případů vyneseny jako černé sloupce dle bydliště obětí
- Identifikace zdroje nákazy – kontaminovaná studně
- Jeden z prvních příkladů prostorové analýzy dat a epidemiologického mapování



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

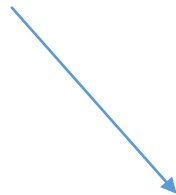
- Základním principem statistiky je pravděpodobnost výskytu nějaké události.
- Prostřednictvím vzorkování se snažíme odhadnout skutečnou pravděpodobnost událostí.
- Klíčovou otázkou je velikost vzorku, čím větší vzorek, tím větší šance na projevení se skutečné pravděpodobnosti výskytu jevu.

Vznik informací: pojmy I

Skutečnost



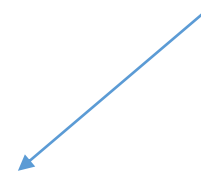
Jev - podmnožina všech možných výsledků pokusu/děje, o které lze říct, zda nastala nebo ne



Pozorovatel



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)**

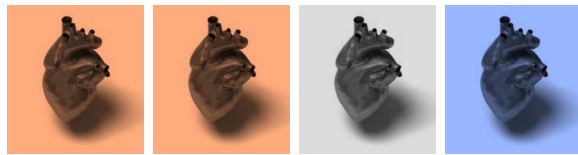
Vznik informací: vzorkování

Statistika hovoří o realitě prostřednictvím výběru z cílové populace

Statistické předpoklady korektního vzorkování je nutné dodržet

Náhodný výběr z cílové populace

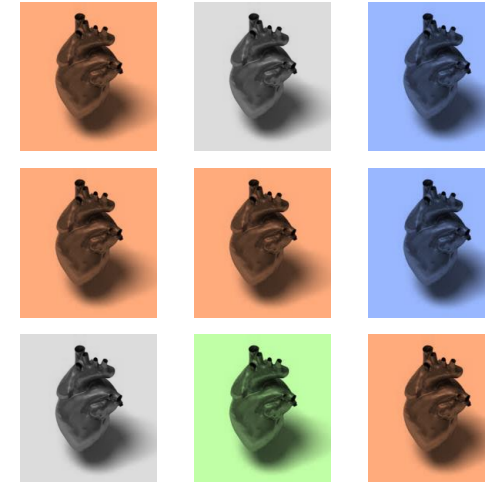
Representativnost: struktura vzorku musí maximálně reflektovat realitu



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



Cílová populace



Příklad vzorkování

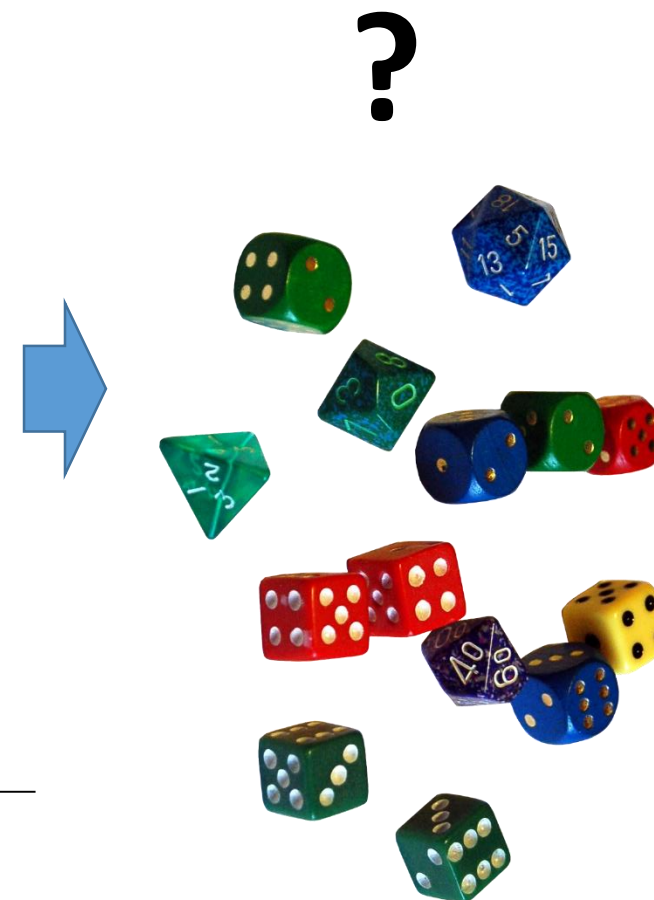
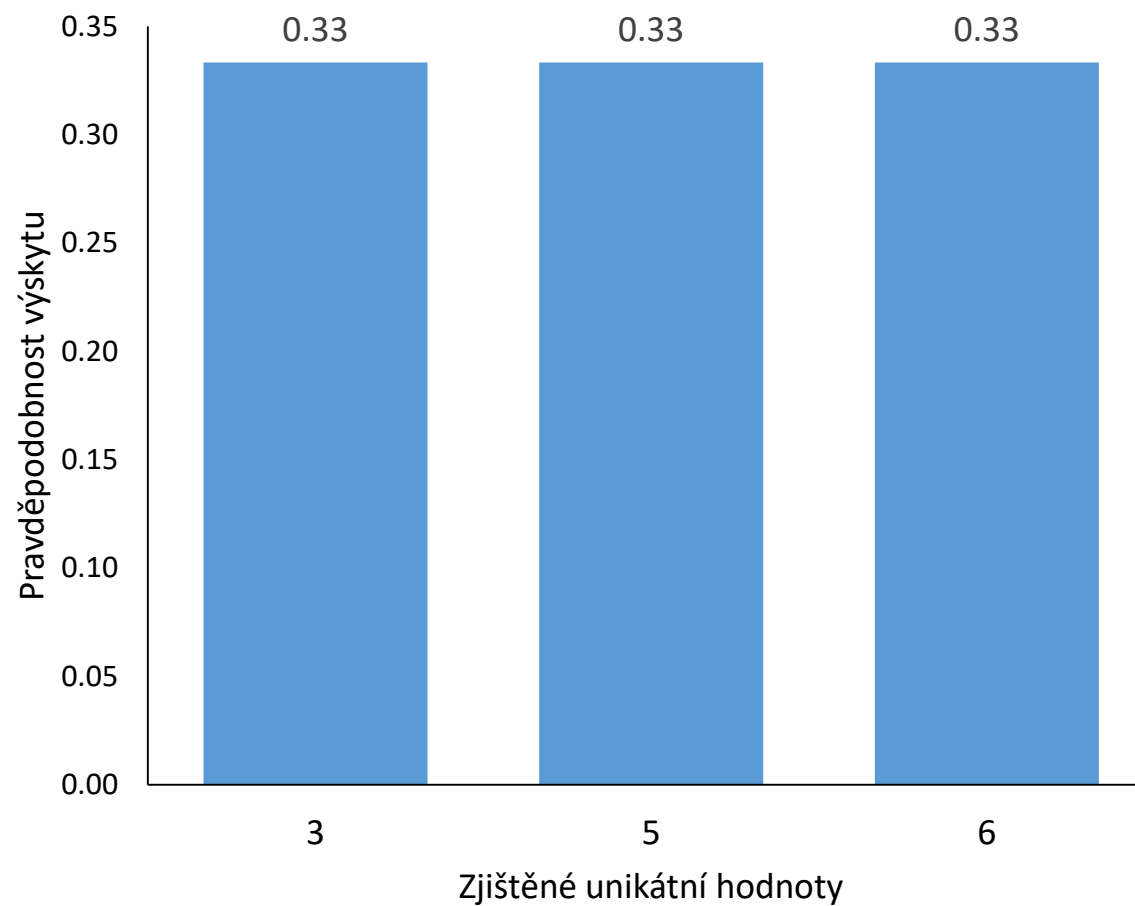
- Na základě vzorkování chceme zjistit vlastnosti nějakého jevu
- Naší cílovou populací budou hody kostkou s neznámými vlastnostmi



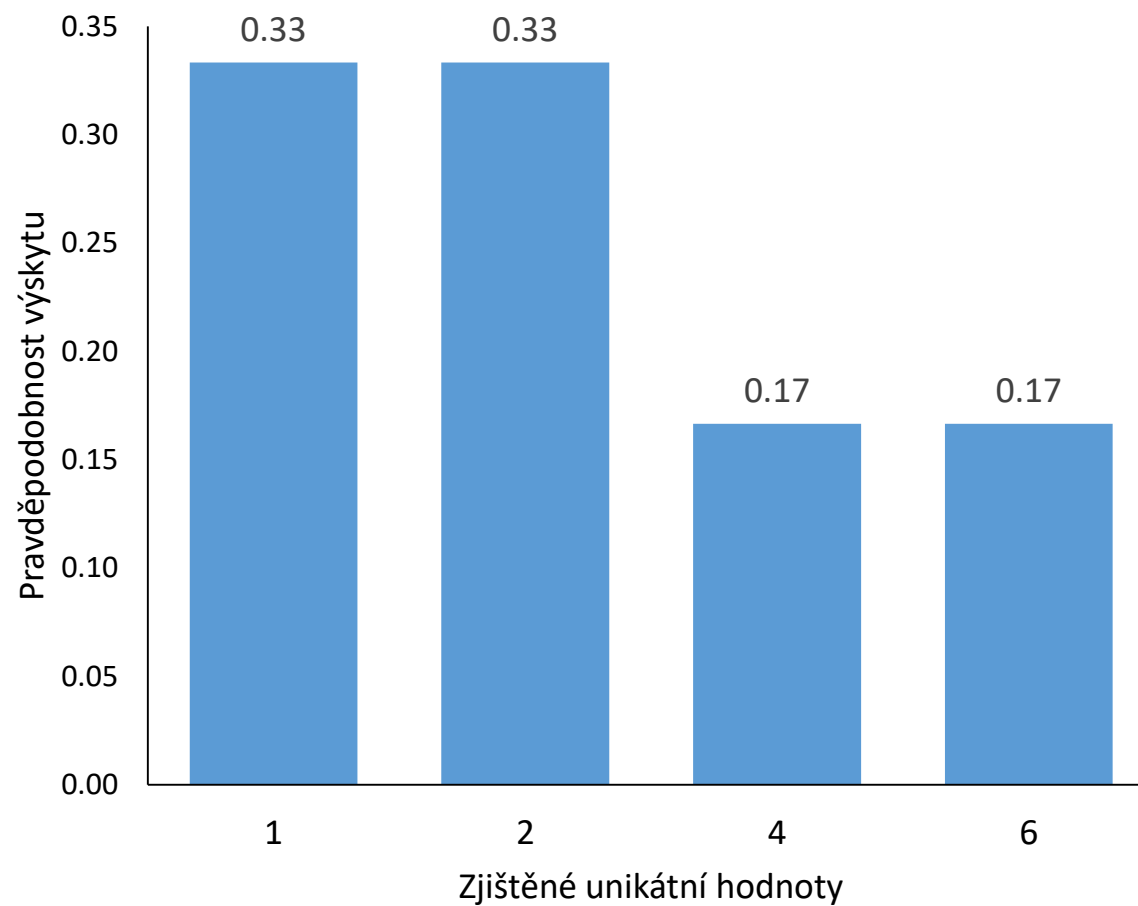
- Chceme zjistit vlastnosti neznámé použité kostky



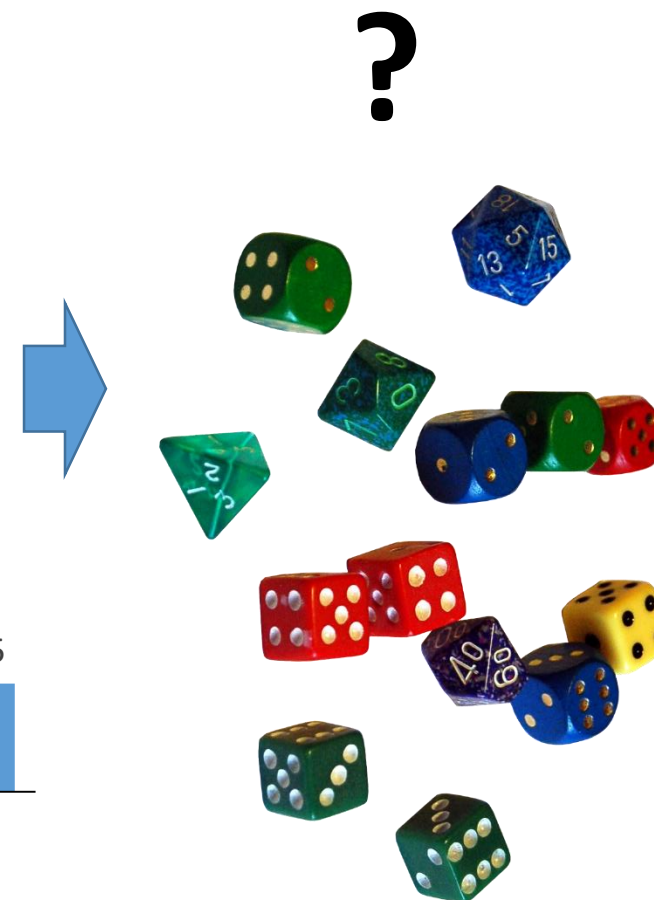
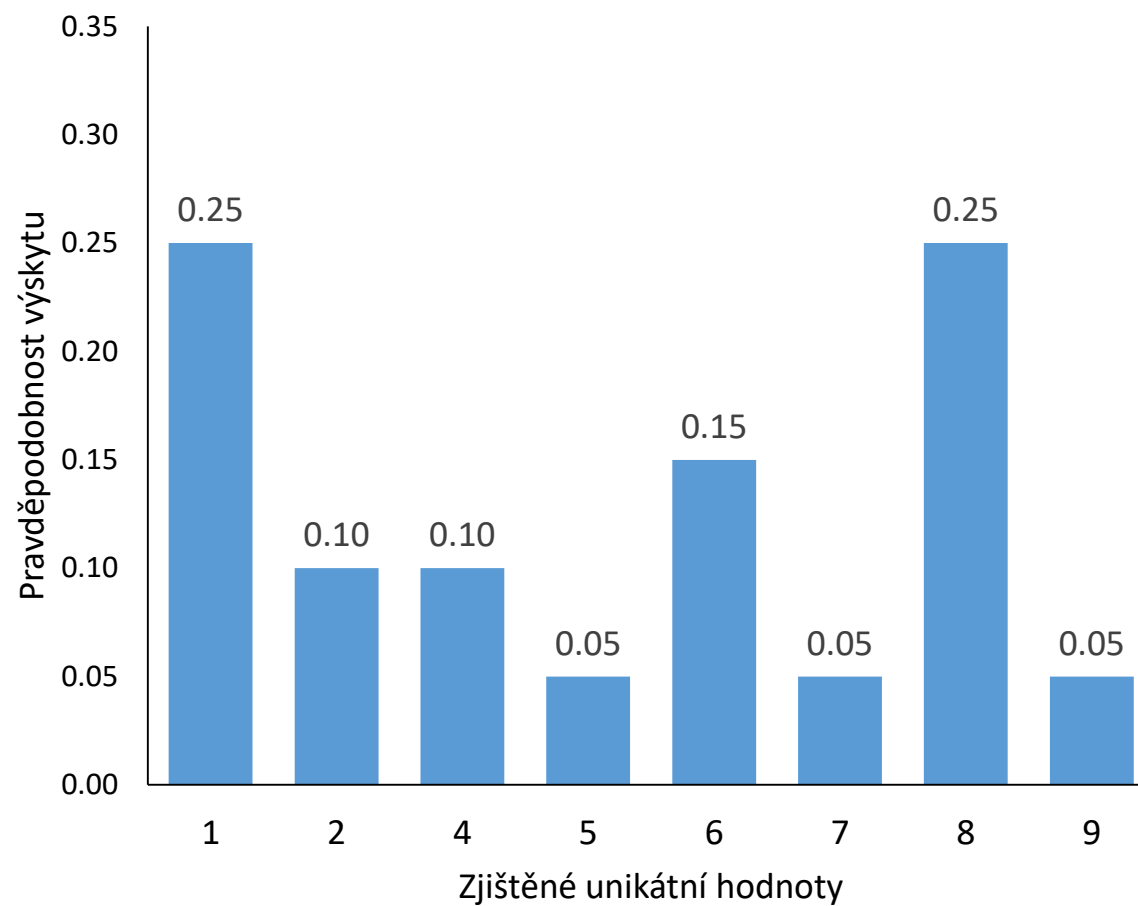
Příklad vzorkování: $N=3$



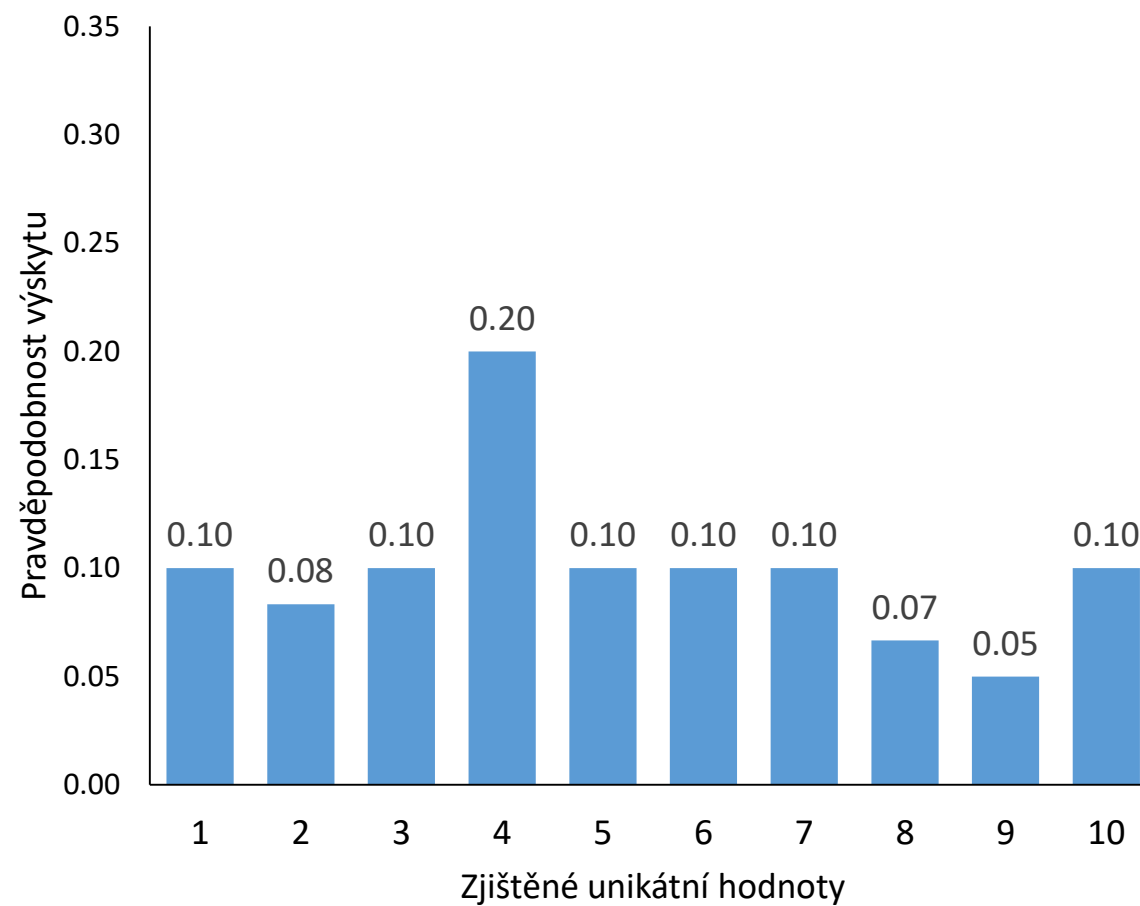
Příklad vzorkování: $N=6$



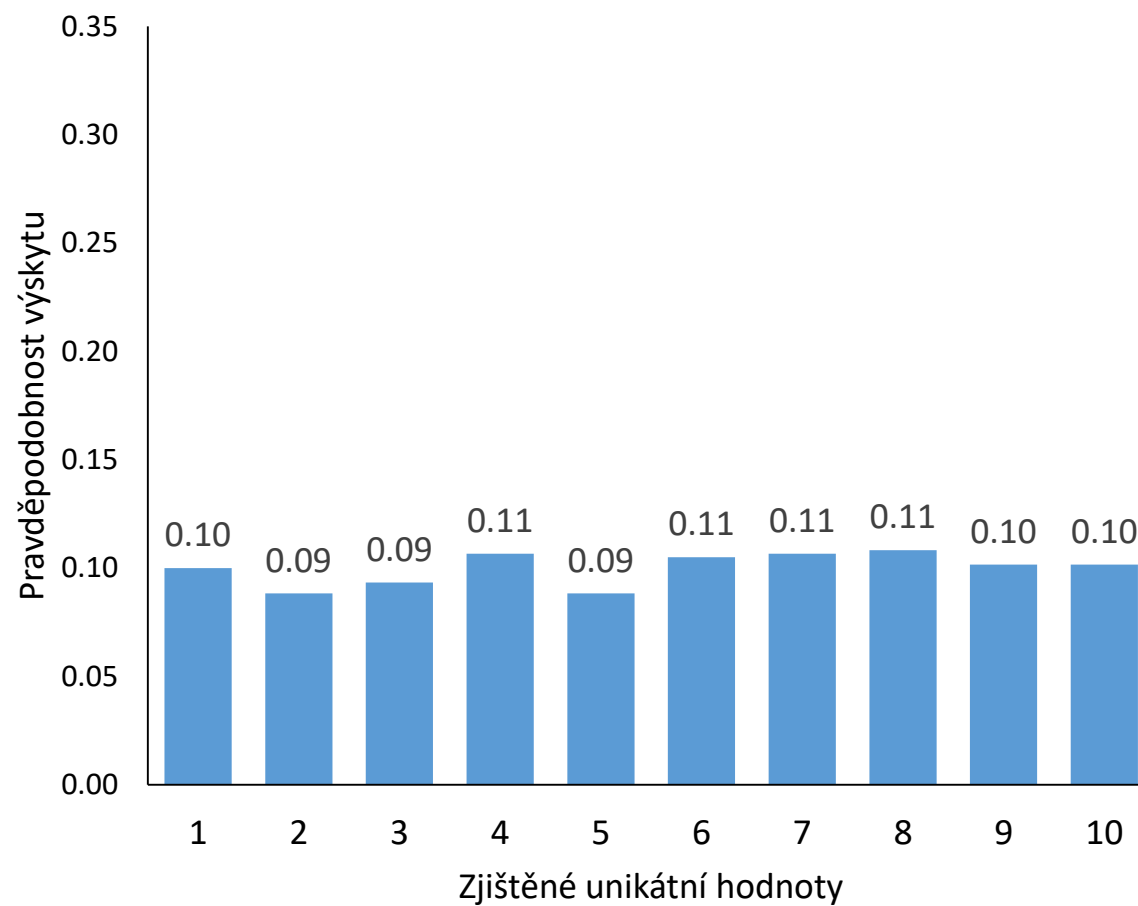
Příklad vzorkování: N=20



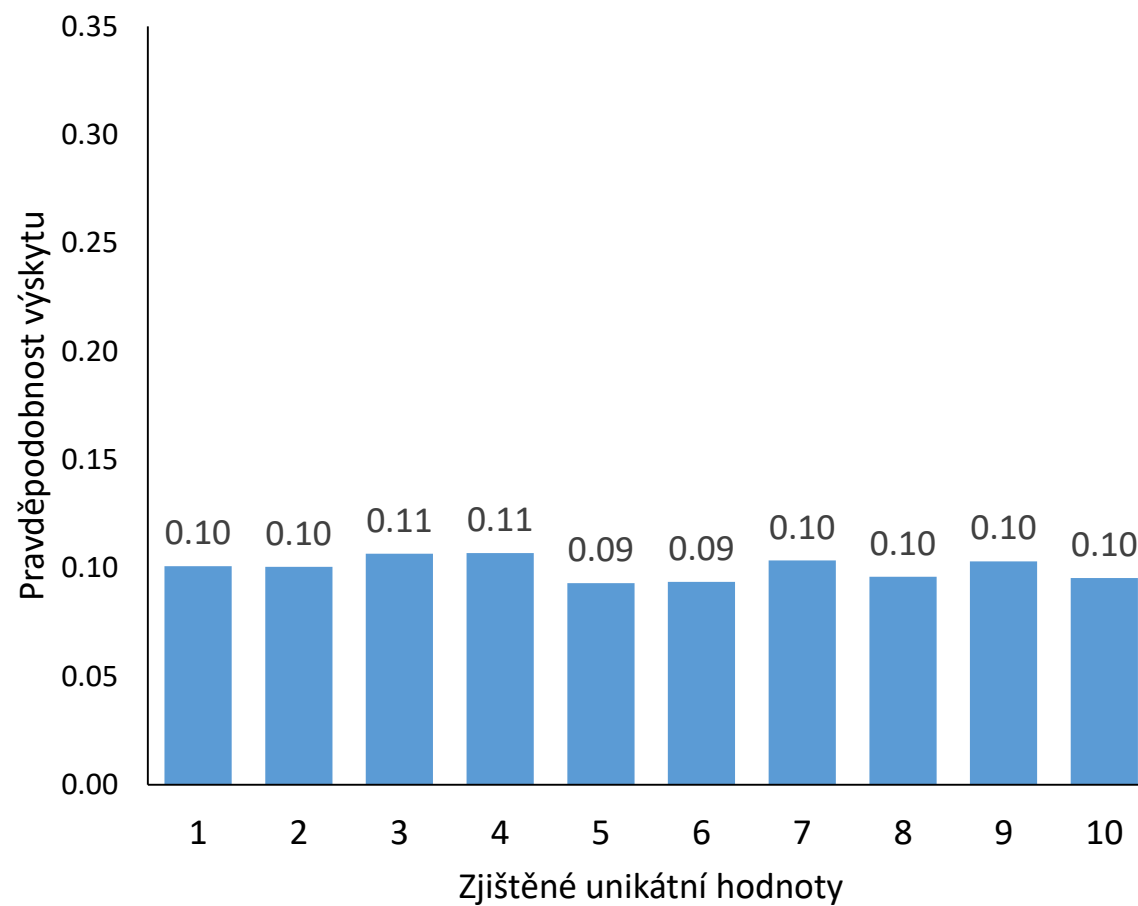
Příklad vzorkování: N=60



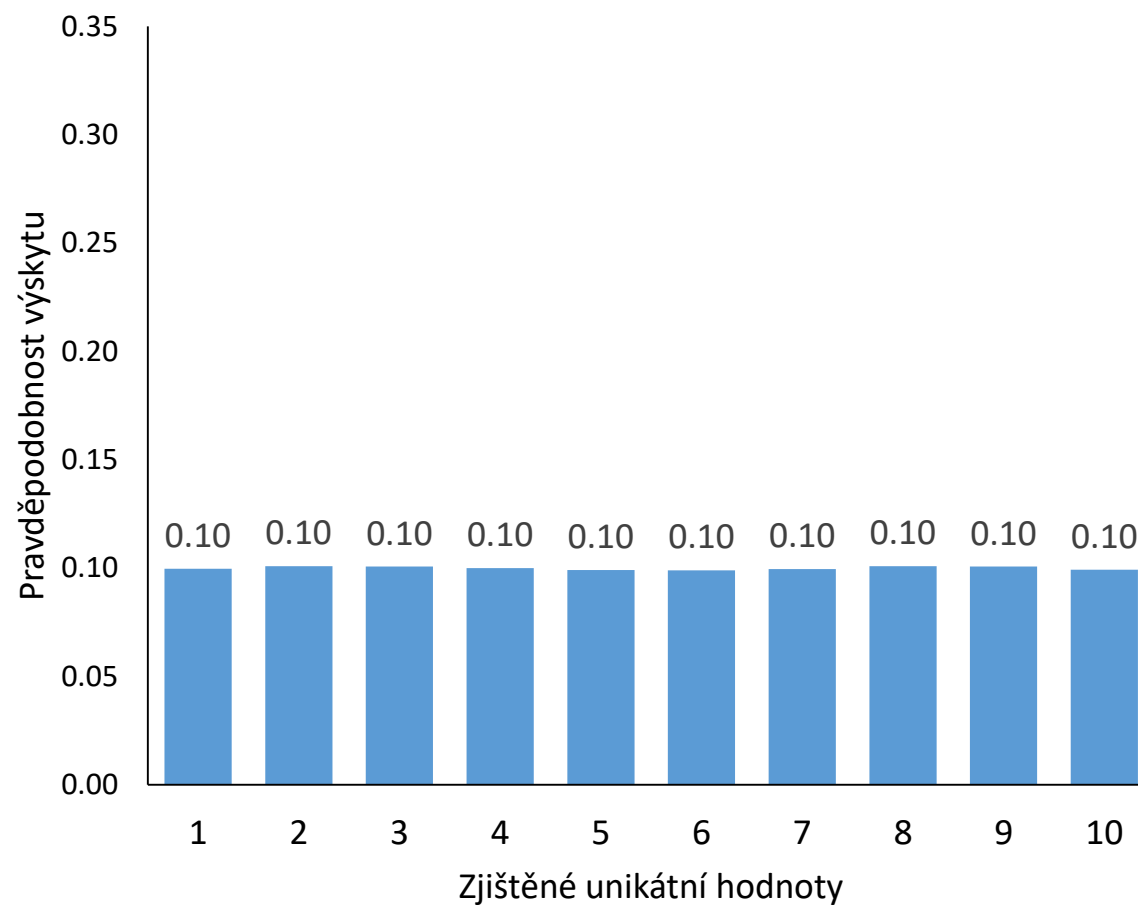
Příklad vzorkování: N=600



Příklad vzorkování: N=6 000



Příklad vzorkování: $N=60\ 000$



?

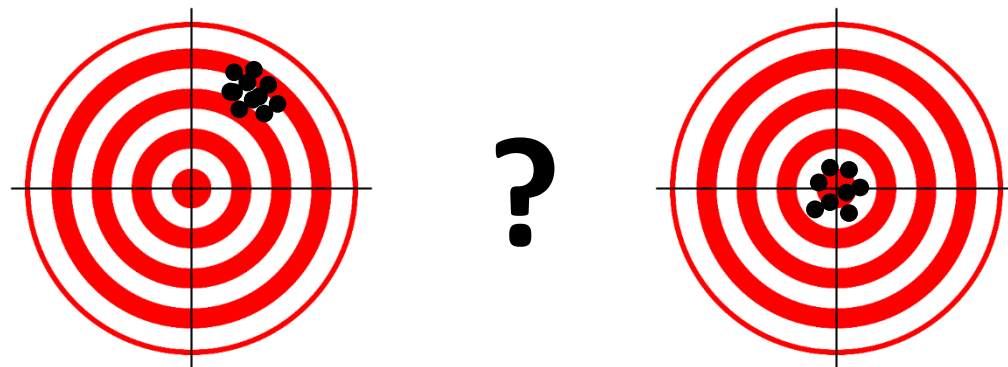


Příklad vzorkování: závěr

- Sledovaný jev má pravděpodobně tvar desetistěnné kostky



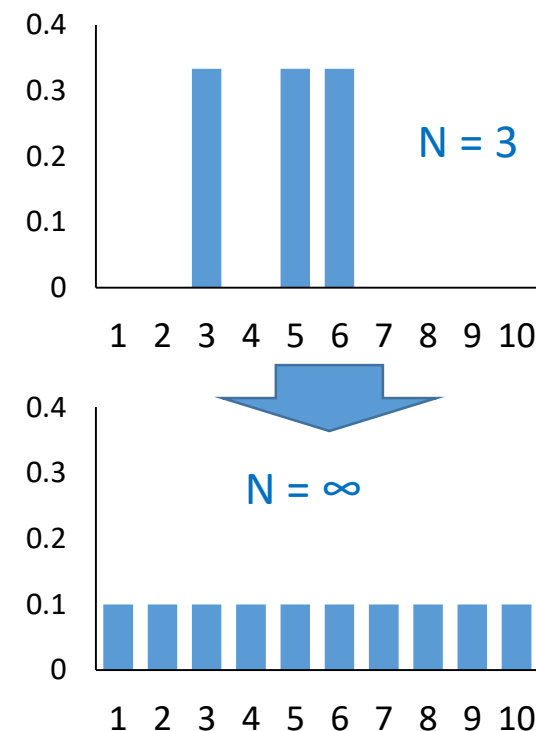
- U složitých stochastických systémů se pravda získá až po odvedení značného množství experimentální práce: musíme dát systému šanci se projevit
- Při realizaci náhodného experimentu roste se zvyšujícím se počtem opakování pravdivá znalost systému (výsledky se stávají stabilnější a spolehlivější)
- Diskutabilní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



Empirický zákon velkých čísel

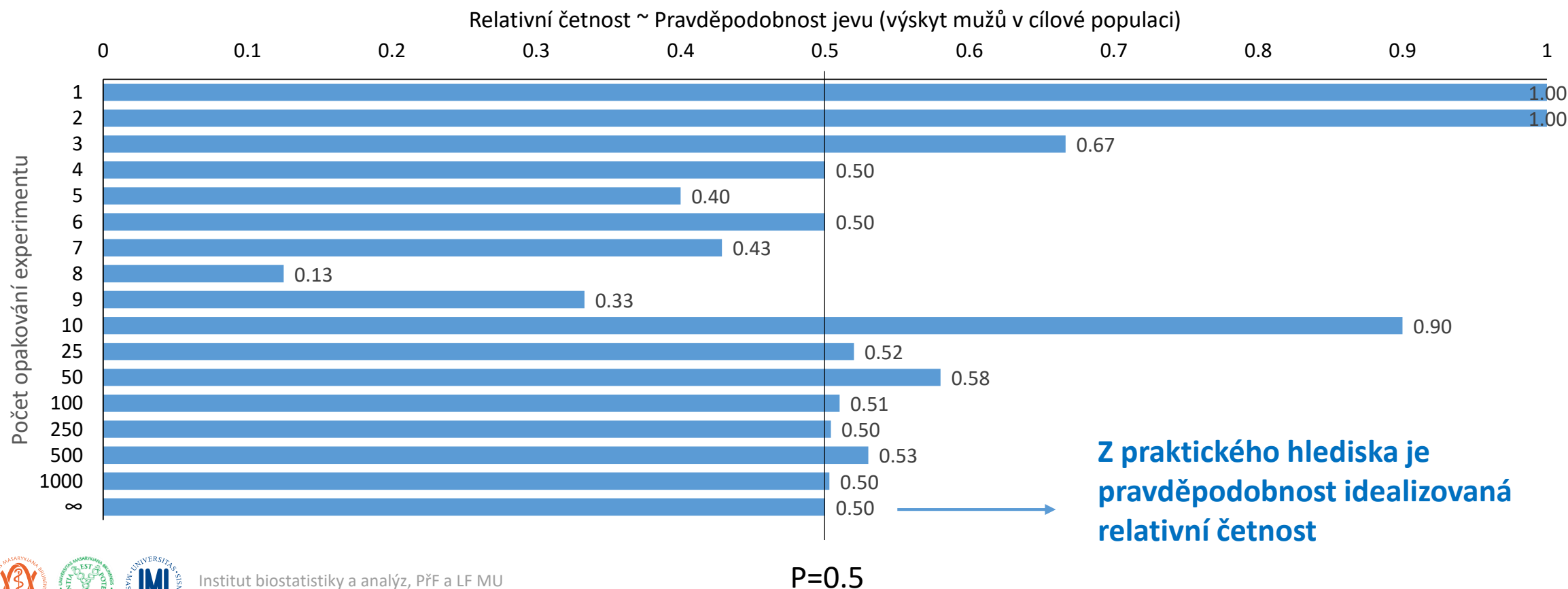
- Při opětovné nezávislé realizaci téhož náhodného experimentu se podíl výskytů sledovaného jevu mezi všemi dosud provedenými realizacemi zpravidla ustaluje kolem konstanty.
- Pravděpodobnost je libovolná reálná funkce definovaná na jevovém poli A (např. hody kostkou), která každému jevu A (např. strany 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
- $P(A \cap B) = P(A) \cdot P(B)$ nezávislé jevy
- $P(A \cap B) = P(A) \cdot P(B/A)$ závislé jevy
- $P(A / B) = P(A \cap B) / P(B)$ 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



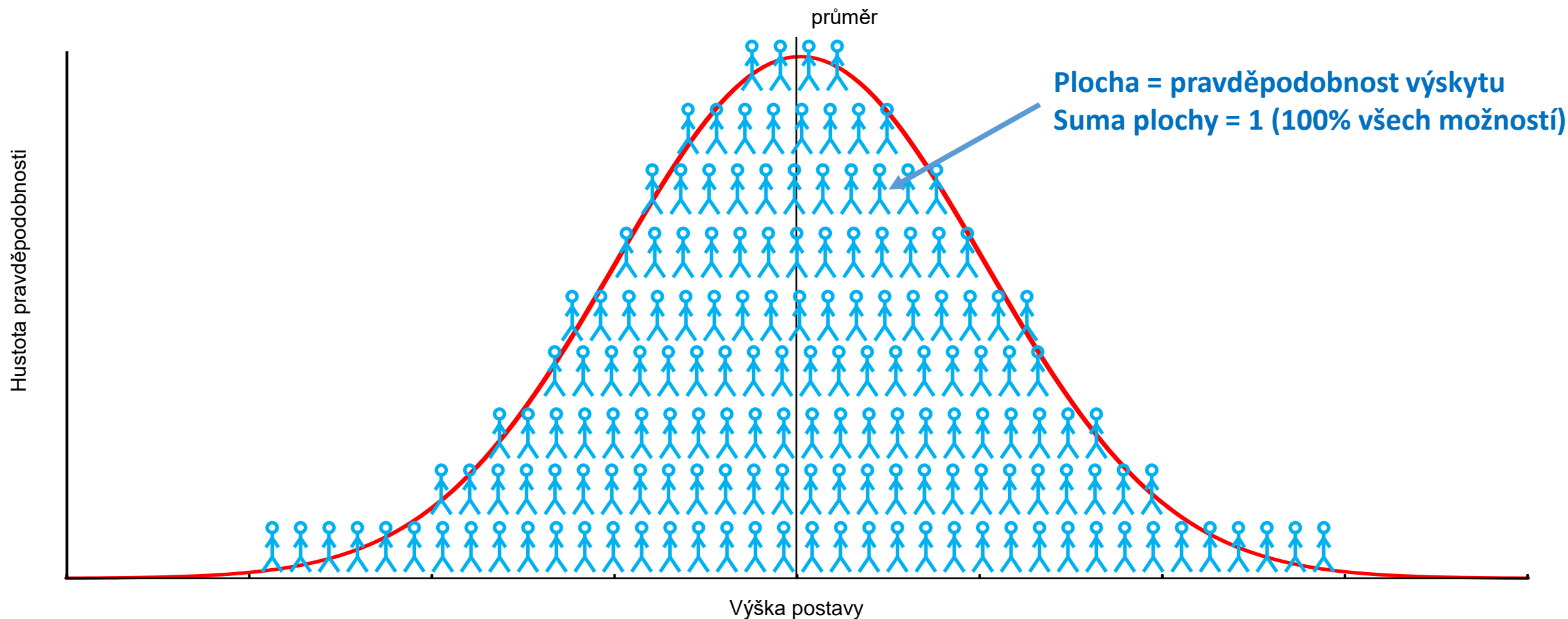
Pravděpodobnost výskytu jevu – rozložení kategoriálních dat

- 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

Spojité 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...

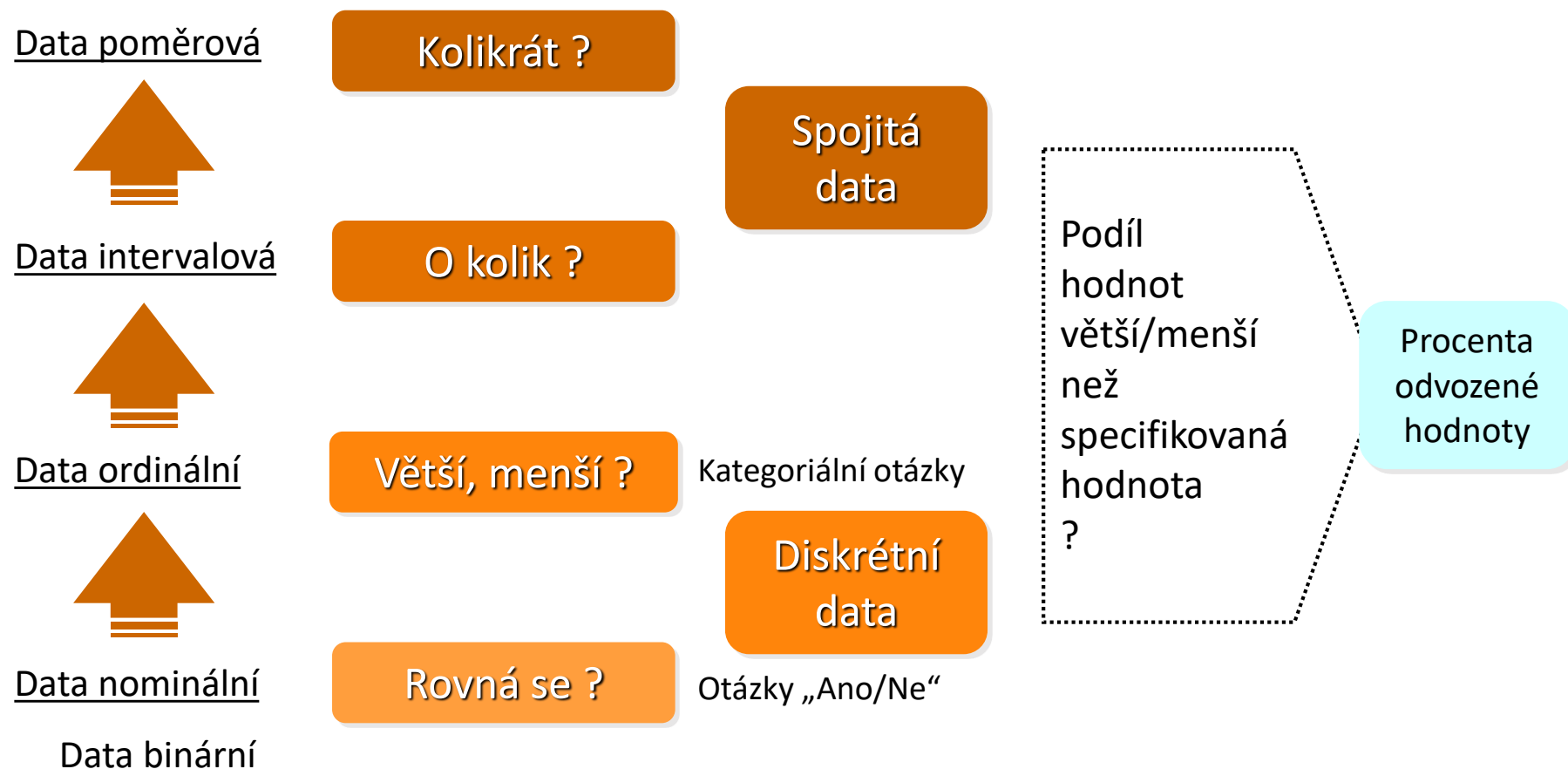
... kterou chceme dále studovat → smysluplnost?

(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 hladu



Cílová populace

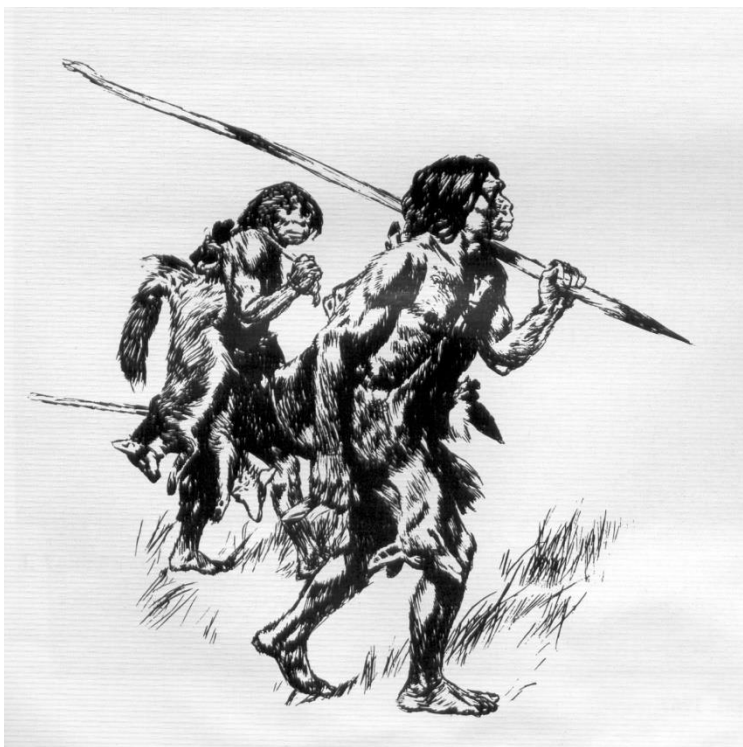
- Vzorkujeme 3 kategorie sledované proměnné kořist

Kořist

Veverka

Jelen

Mamut



Binární data – chytili jsme něco?

- Informačně nejméně obsáhlá jsou data binární

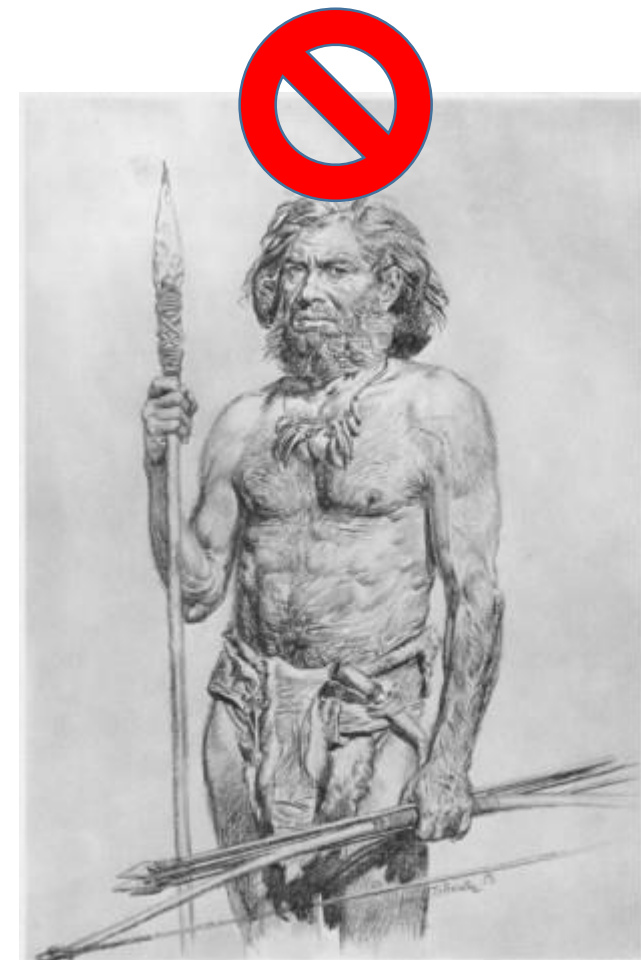


Hodnotíme dva možné stavy:

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:

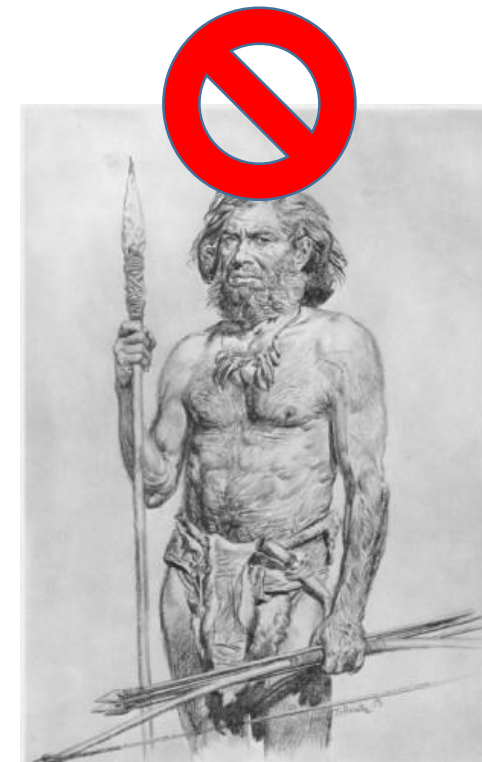
Celkový počet lovů (**báze hodnocení**)



Počet úlovků (**absolutní četnost**)



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



Jsou binární data dostatečná za všech okolností?

Kategoriální data – co jsme chytili?

- Více informací získáme z dat kategoriálních

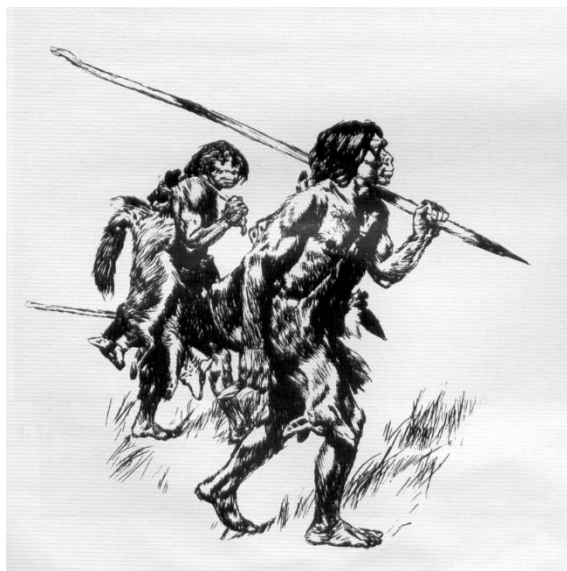
Hodnotíme několik možných stavů:

Jak můžeme popsat:

Celkový počet lovů (báze hodnocení)

Počet různých kategorií úlovků
(absolutní četnost)

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



N = 4 (40%)



N = 1 (10%)



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 seřaditelných 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?



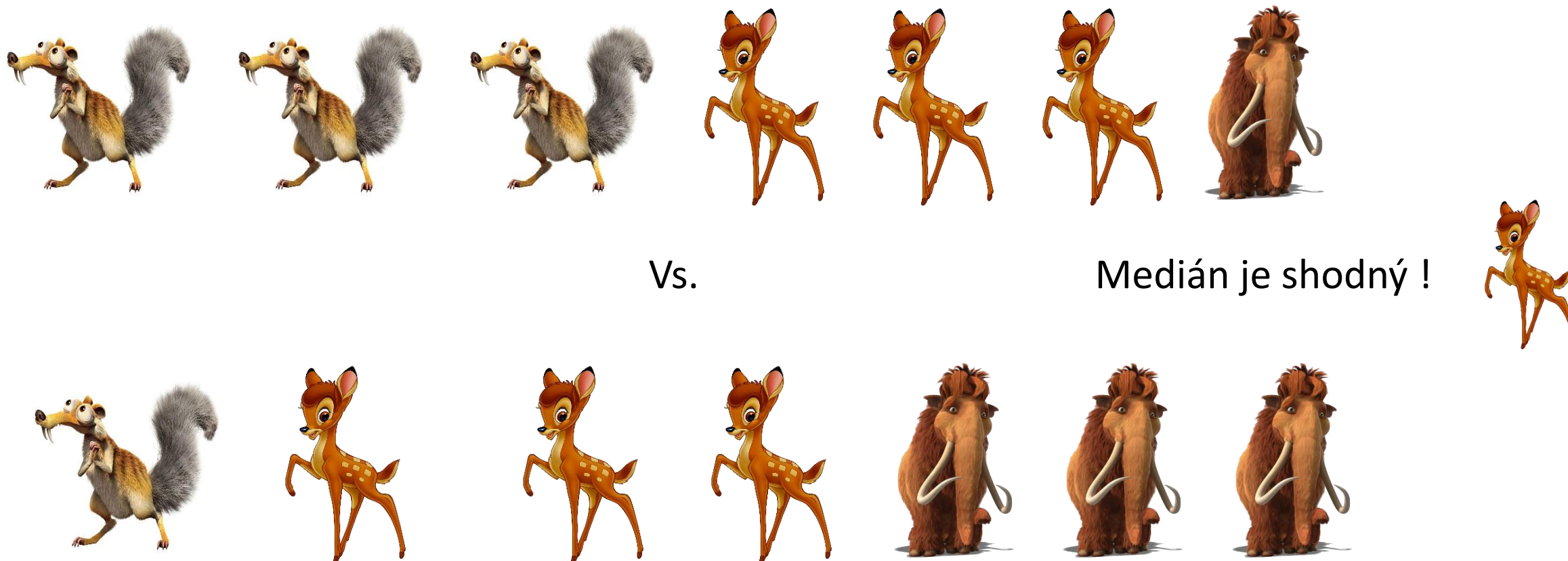
Medián?

Vs.



Medián?

Pozor na medián u ordinálních dat



- 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.



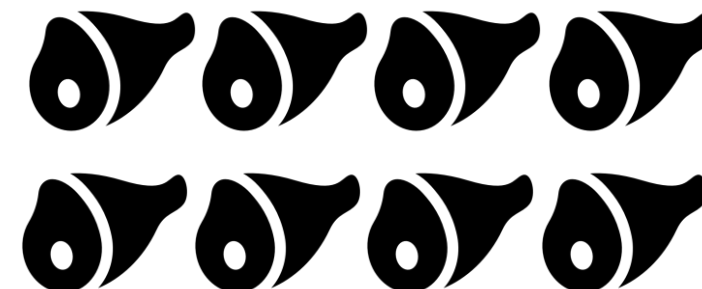
=



=



=



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



PRŮMĚR

Spojité data

Data intervalová



MEDIÁN

Data ordinální



MODUS

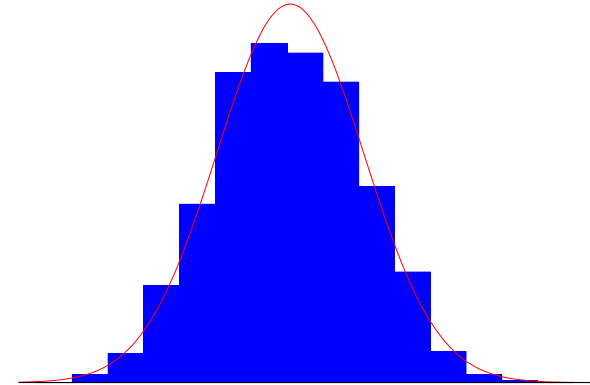
Data nominální

Data binární

Absolutní a
relativní četnosti

Diskrétní data

- Kvantitativní data - četnost hodnot rozložení v jednotlivých intervalech.

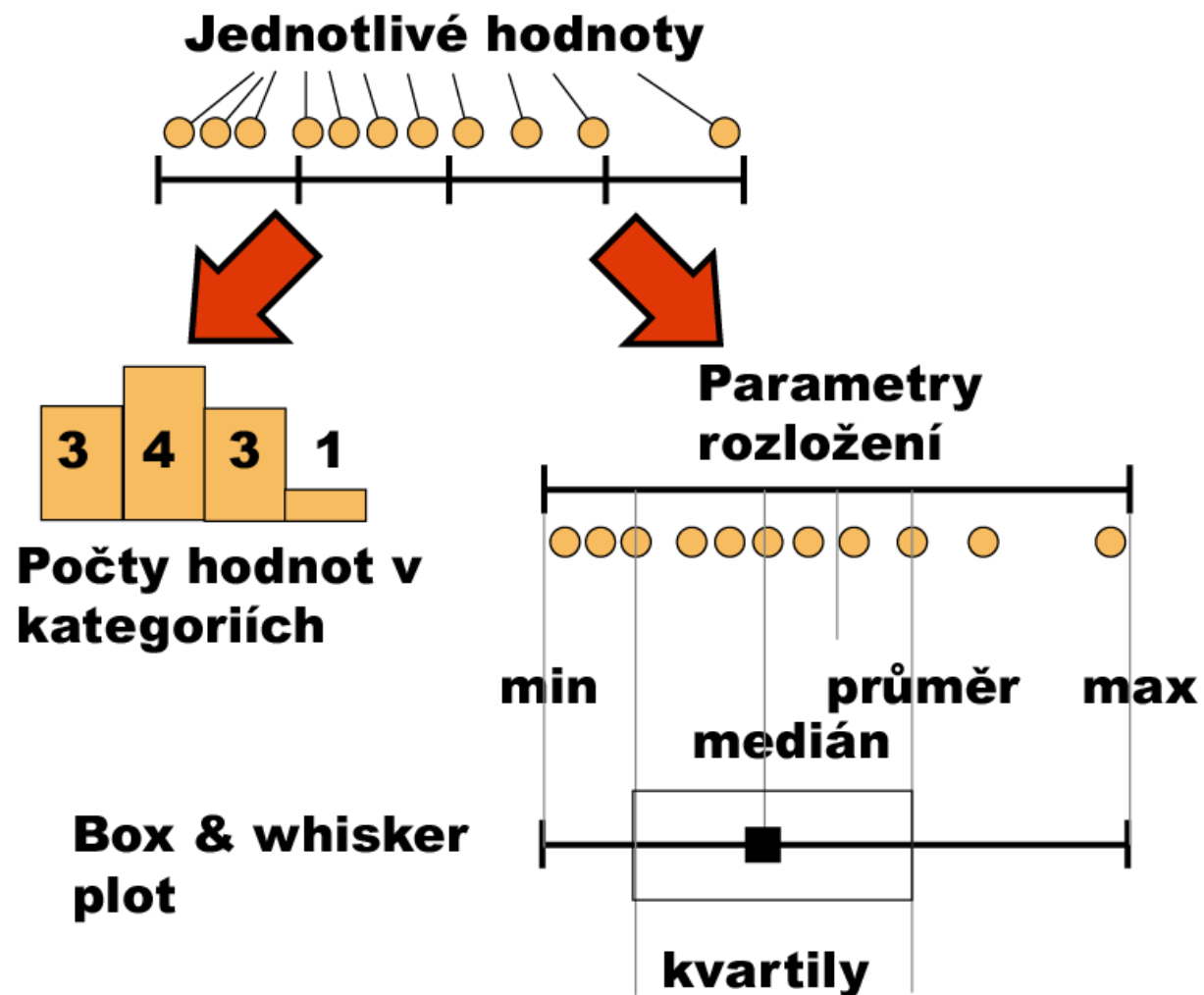


- Kvalitativní data - tabulka s četností jednotlivých kategorií.

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



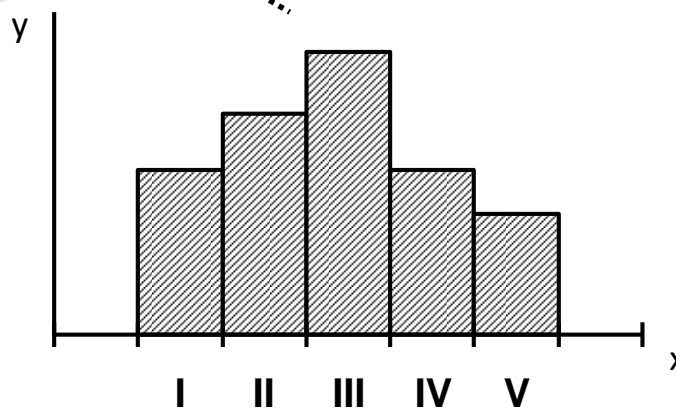
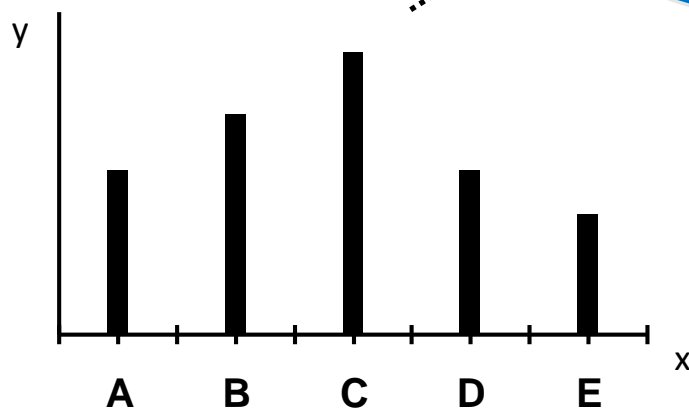
Odvozená data: pozor na odvozené indexy

- 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) \longrightarrow + / - 3,8 %
 - Y: 1,8 : (1,75 - 1,84) \longrightarrow + / - 2,5 %
 - $\frac{X}{Y} = 0,667 : \left(\frac{1,15}{1,84} - \frac{1,24}{1,75} \right) \longrightarrow$ + / - 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

Spojité data

Frekvenční sumarizace - základní nástroj popisu dat: kvalitativní data

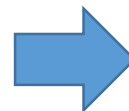
- Cílem sumarizace je zjednodušení dat do přehledné formy
- N = 100 pacientů s hemofilií
- Hodnocenou proměnnou je počet krvácivých epizod za měsíc
- Nejjednodušší sumarizací je frekvenční tabulka

*Untitled2 [DataSet1] - IBM SPSS Statistics

File Edit View Data Transform

1:

	epizody
1	1
2	0
3	1
4	2
5	2
6	1
7	1
8	3
9	2
10	1
11	3
12	1
13	2
14	2
15	2
16	0
17	0
18	3
19	1
20	1
21	1
22	0
23	1
24	2
25	1
26	3
27	1
28	2
29	1
30	2
31	0
32	1
33	1
34	2

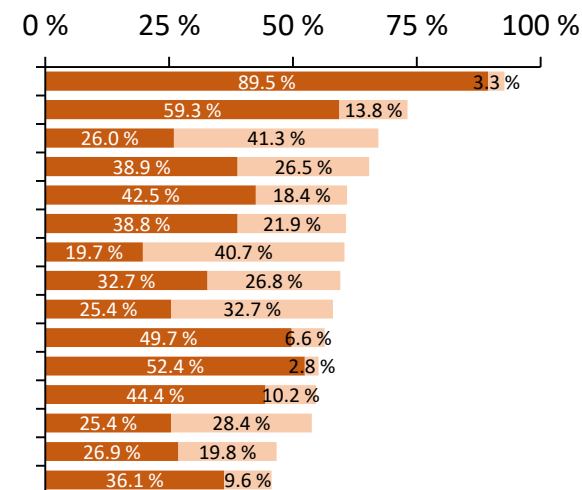
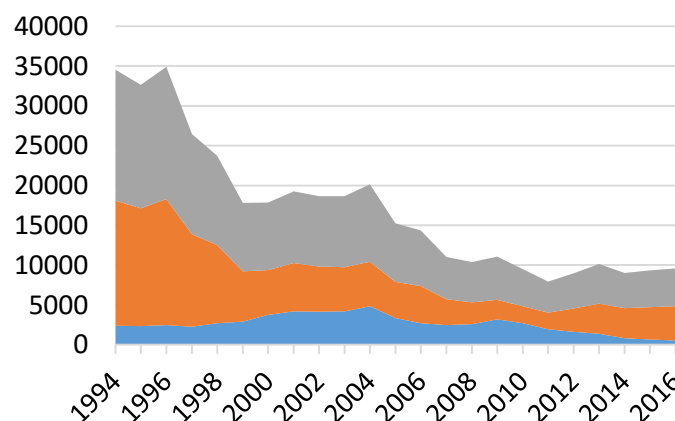
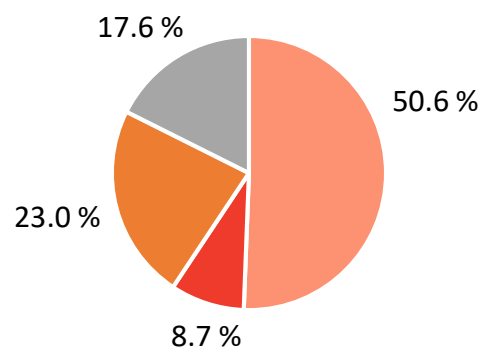
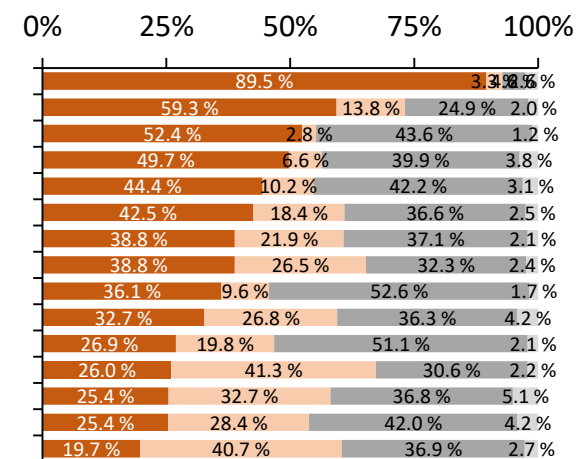
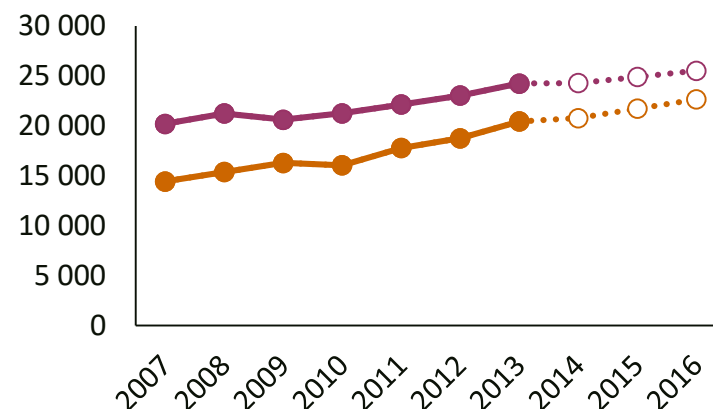
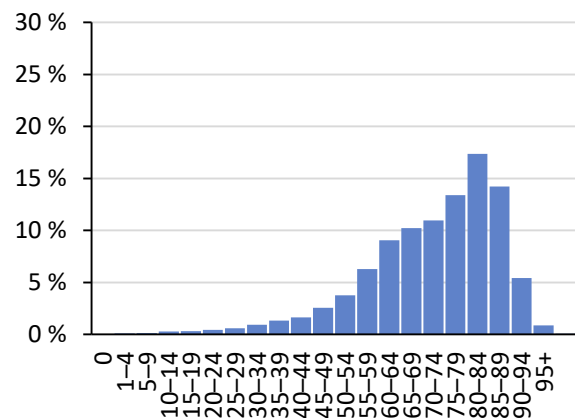


		epizody			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	22	22,0	22,0	22,0
	1	27	27,0	27,0	49,0
	2	29	29,0	29,0	78,0
	3	22	22,0	22,0	100,0
	Total	100	100,0	100,0	

- 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)

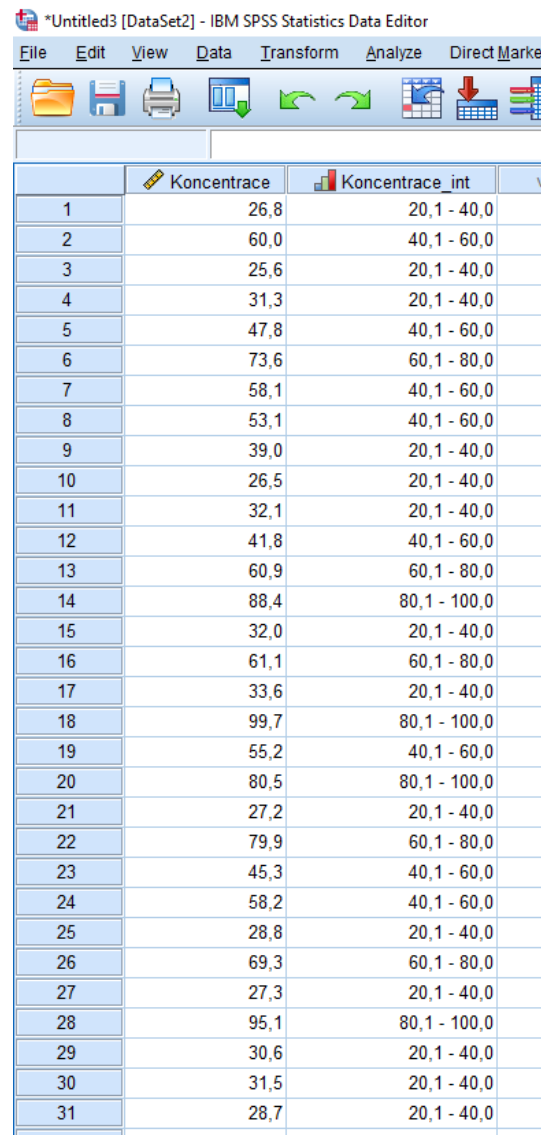
Vizualizace frekvenční tabulky kvalitativních dat

- Libovolné grafy umožňující vizualizaci počtů a procent (koláčový, páskový, sloupkový, 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.)



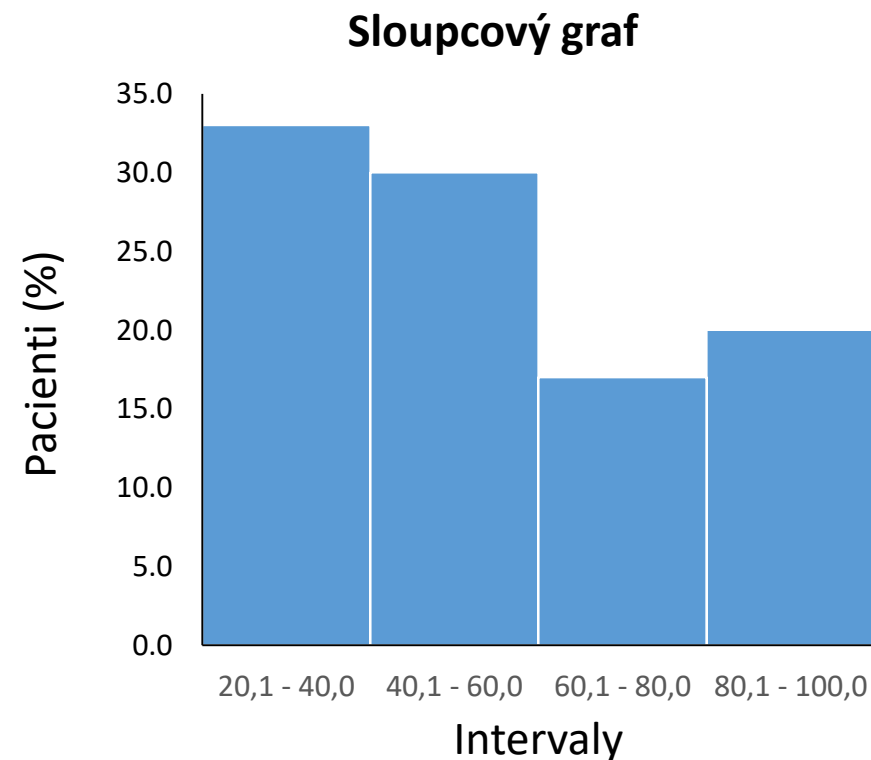
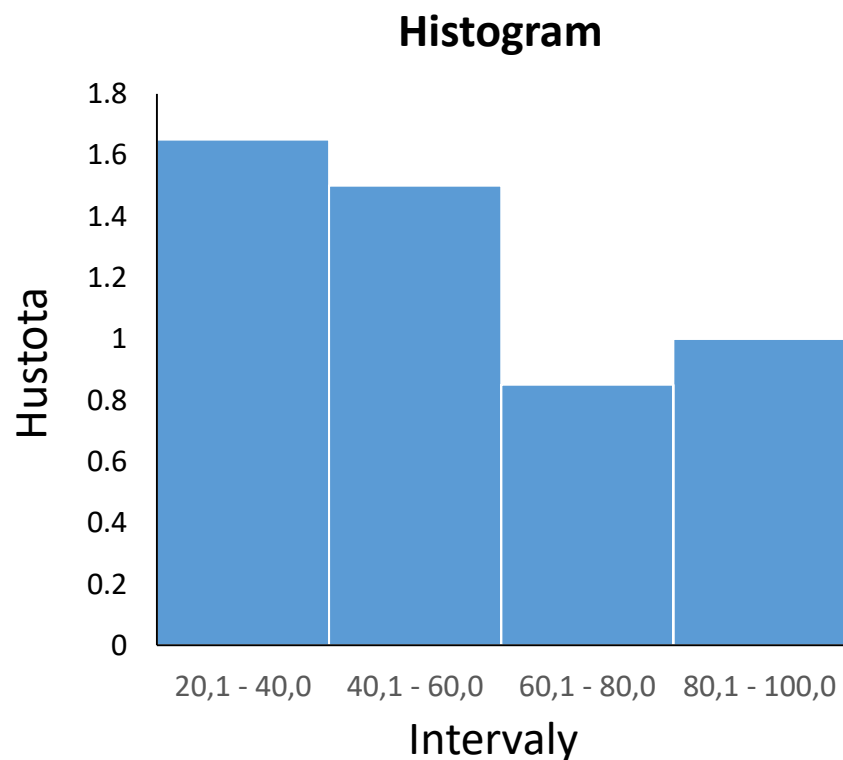
	Koncentrace	Koncentrace_int
1	26,8	20,1 - 40,0
2	60,0	40,1 - 60,0
3	25,6	20,1 - 40,0
4	31,3	20,1 - 40,0
5	47,8	40,1 - 60,0
6	73,6	60,1 - 80,0
7	58,1	40,1 - 60,0
8	53,1	40,1 - 60,0
9	39,0	20,1 - 40,0
10	26,5	20,1 - 40,0
11	32,1	20,1 - 40,0
12	41,8	40,1 - 60,0
13	60,9	60,1 - 80,0
14	88,4	80,1 - 100,0
15	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

		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)

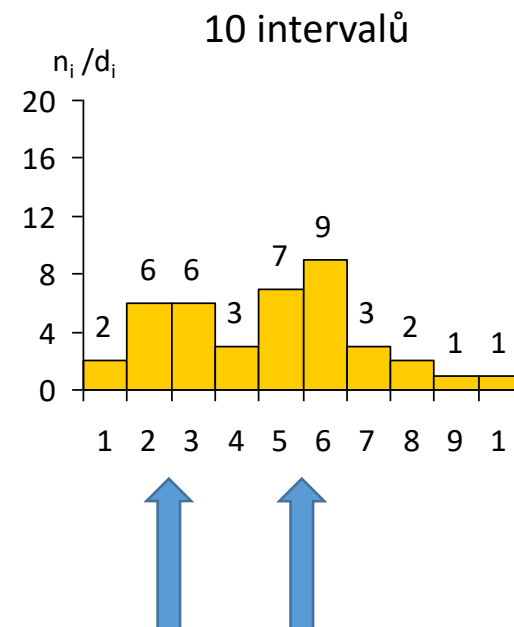
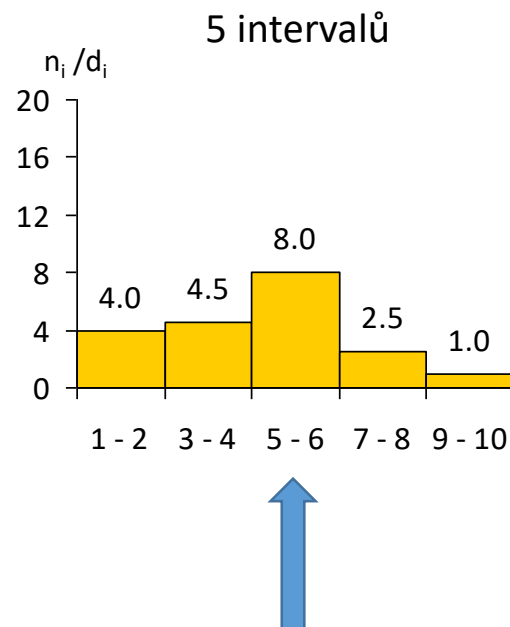
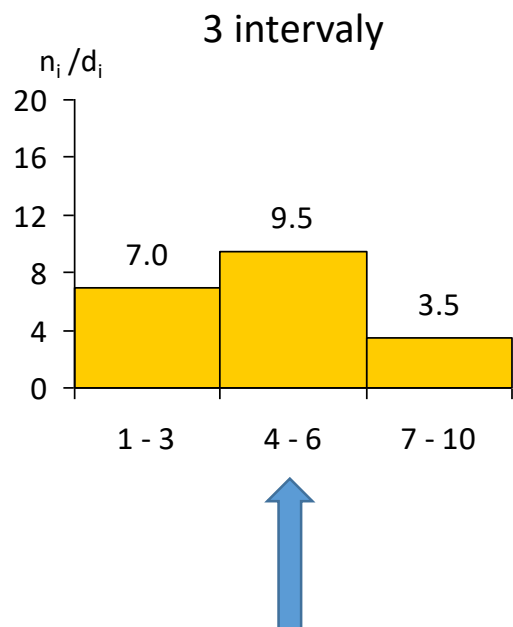
Vizualizace frekvenční tabulky kvantitativních dat

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



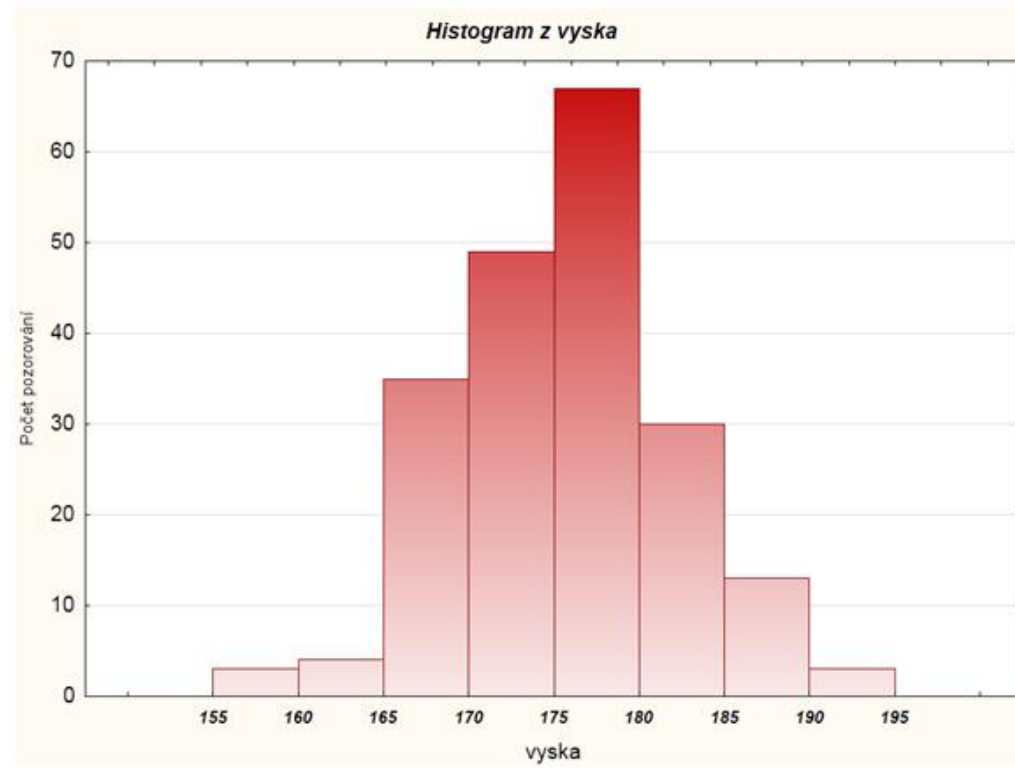
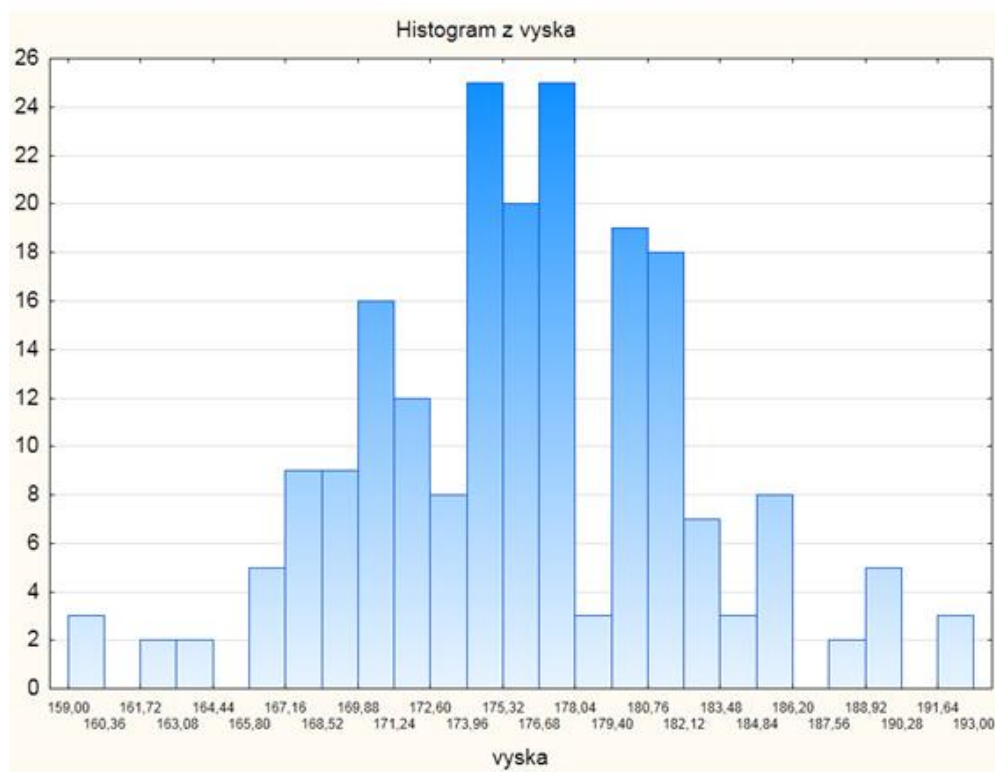
Histogram: vliv kategorizace dat

- 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 roztržštěná.



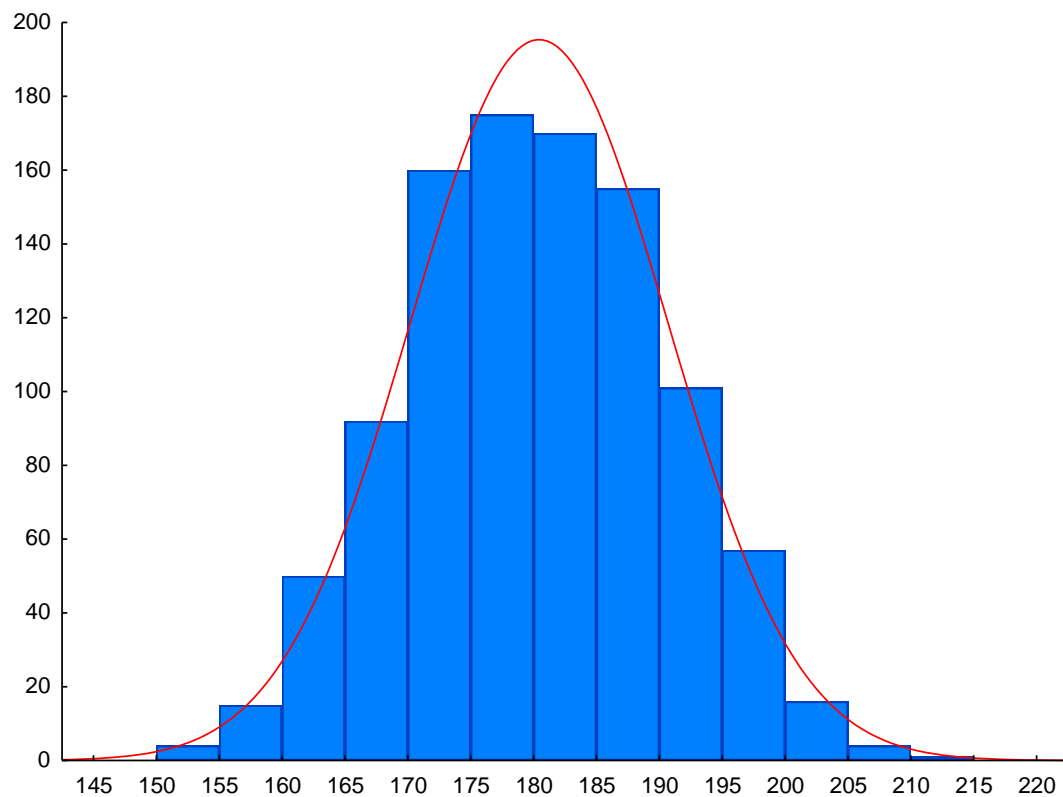
Histogram: vliv kategorizace dat

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

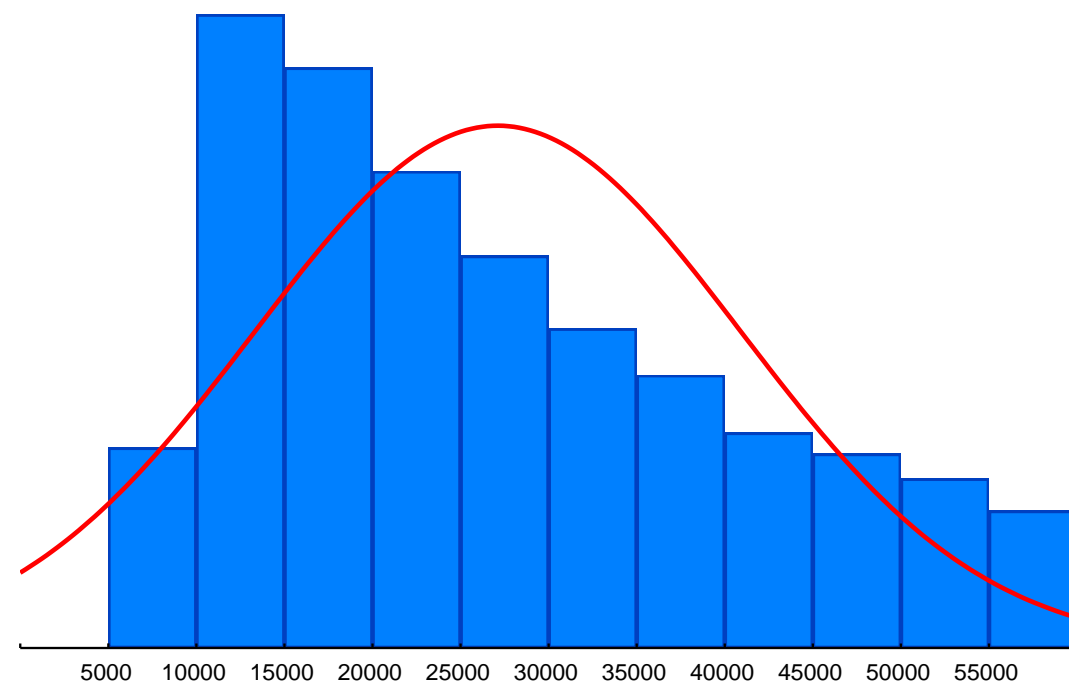


Histogram: nástroj posouzení rozložení dat

- Histogram reálných dat má vazbu na modelové rozdělení



?



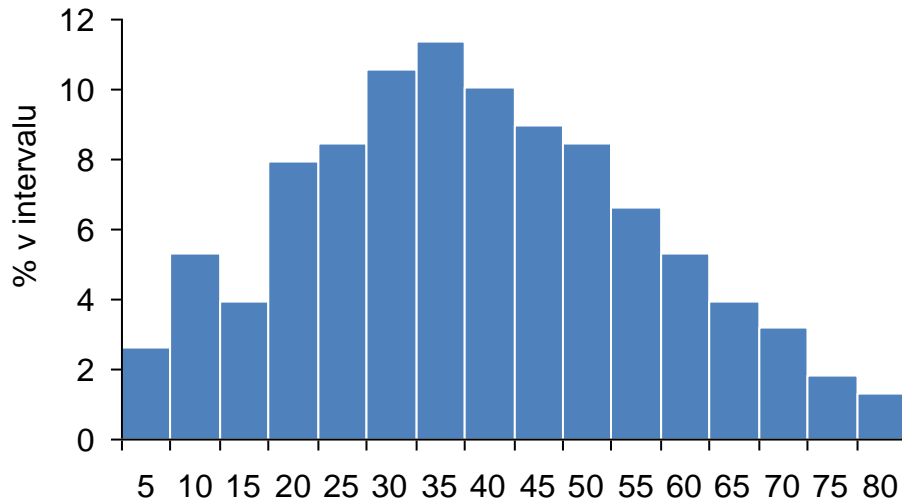
Proč je důležité vědět co je to skutečný histogram I

- 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)

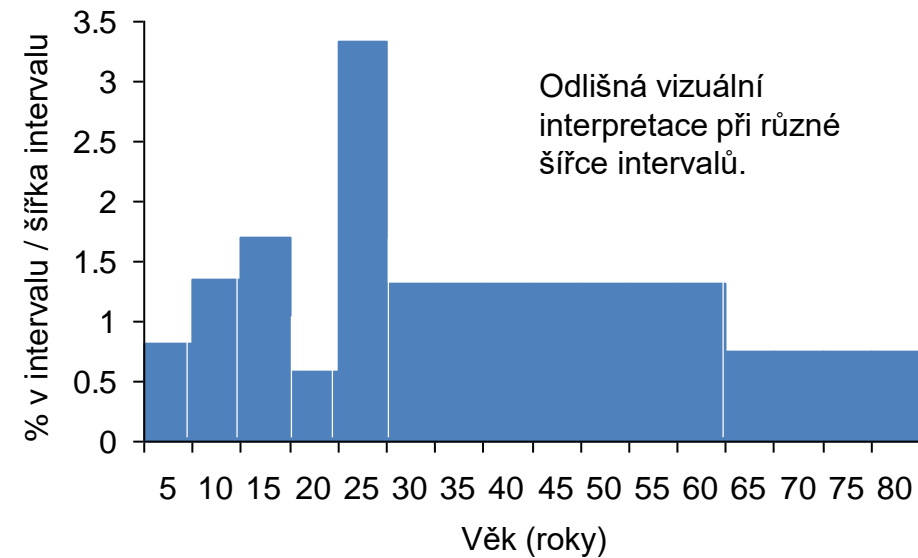
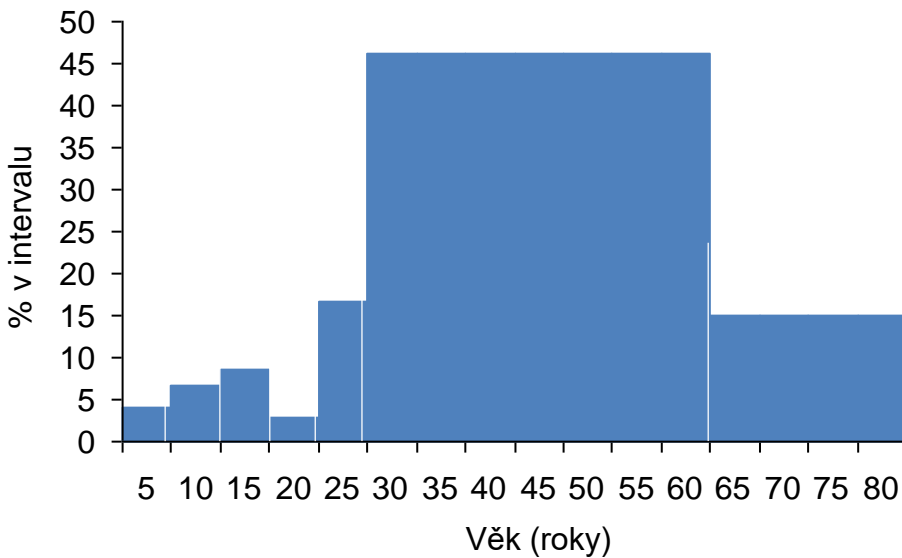
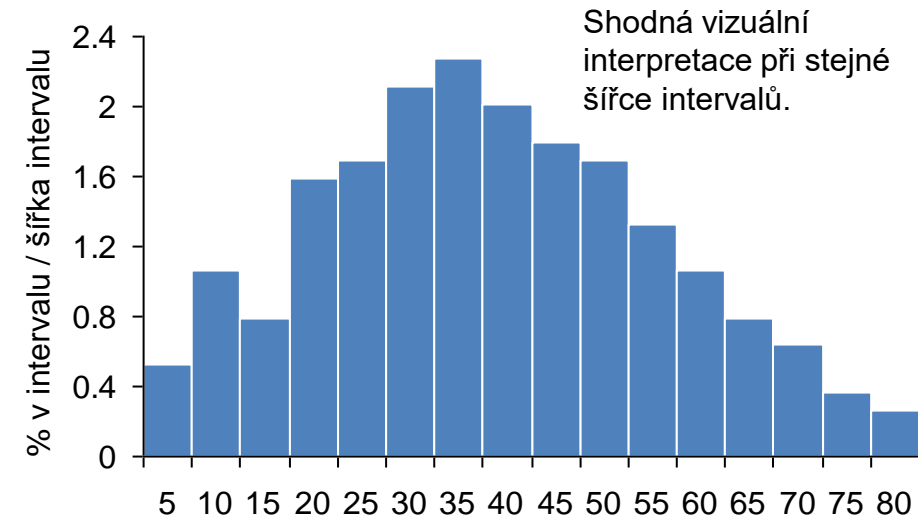


Histogram a sloupcový graf

Sloupcový graf

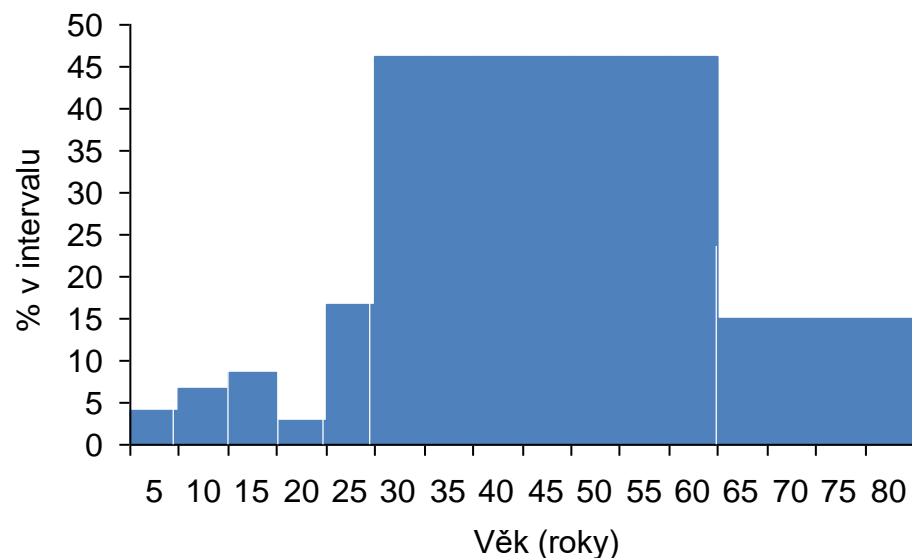


Histogram

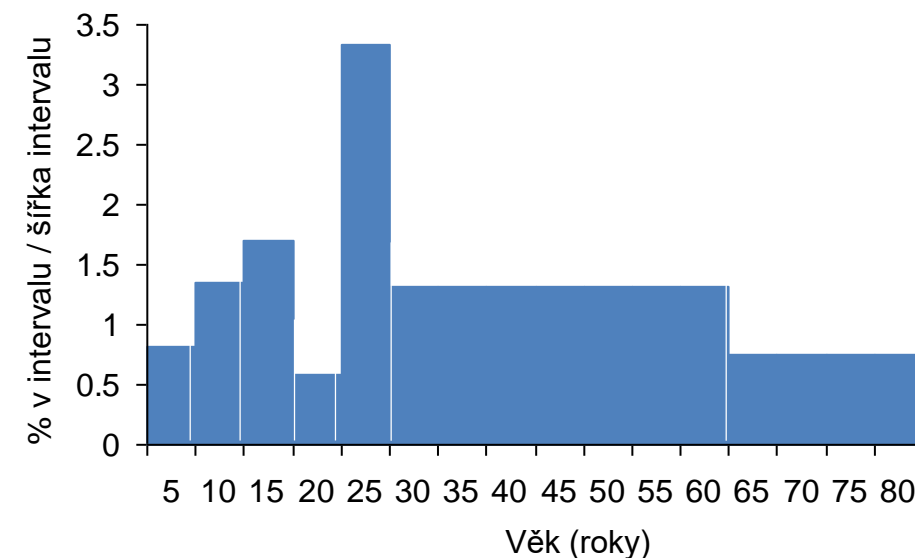


Příklad: věk účastníků vážných dopravních nehod

- 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?



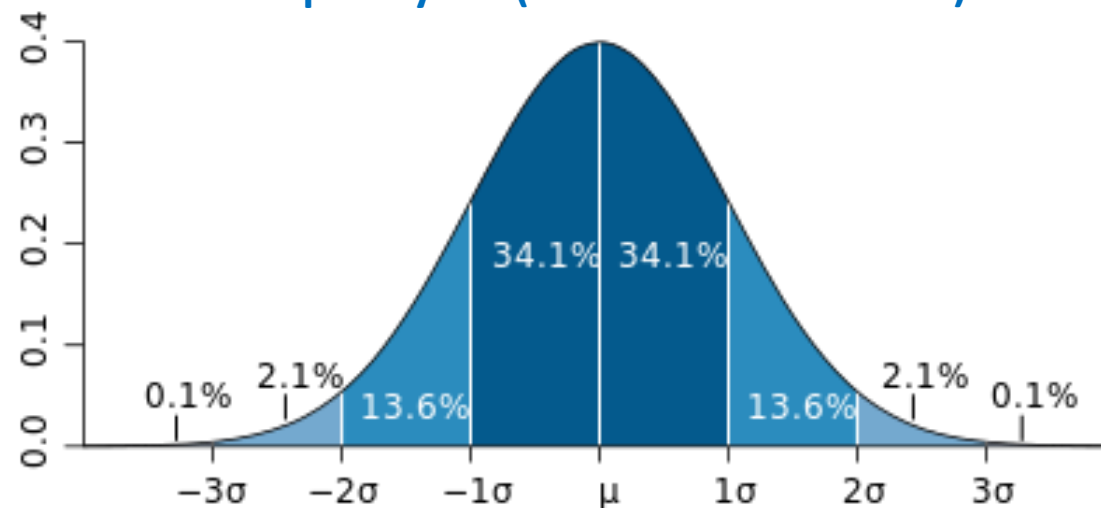
Věk	N	%
0 - 4	28	4,1%
5 - 9	46	6,7%
10-15	58	8,5%
16 - 19	20	2,9%
20 - 24	114	16,6%
25 - 59	316	46,1%
> 60	103	15,0%



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 ???

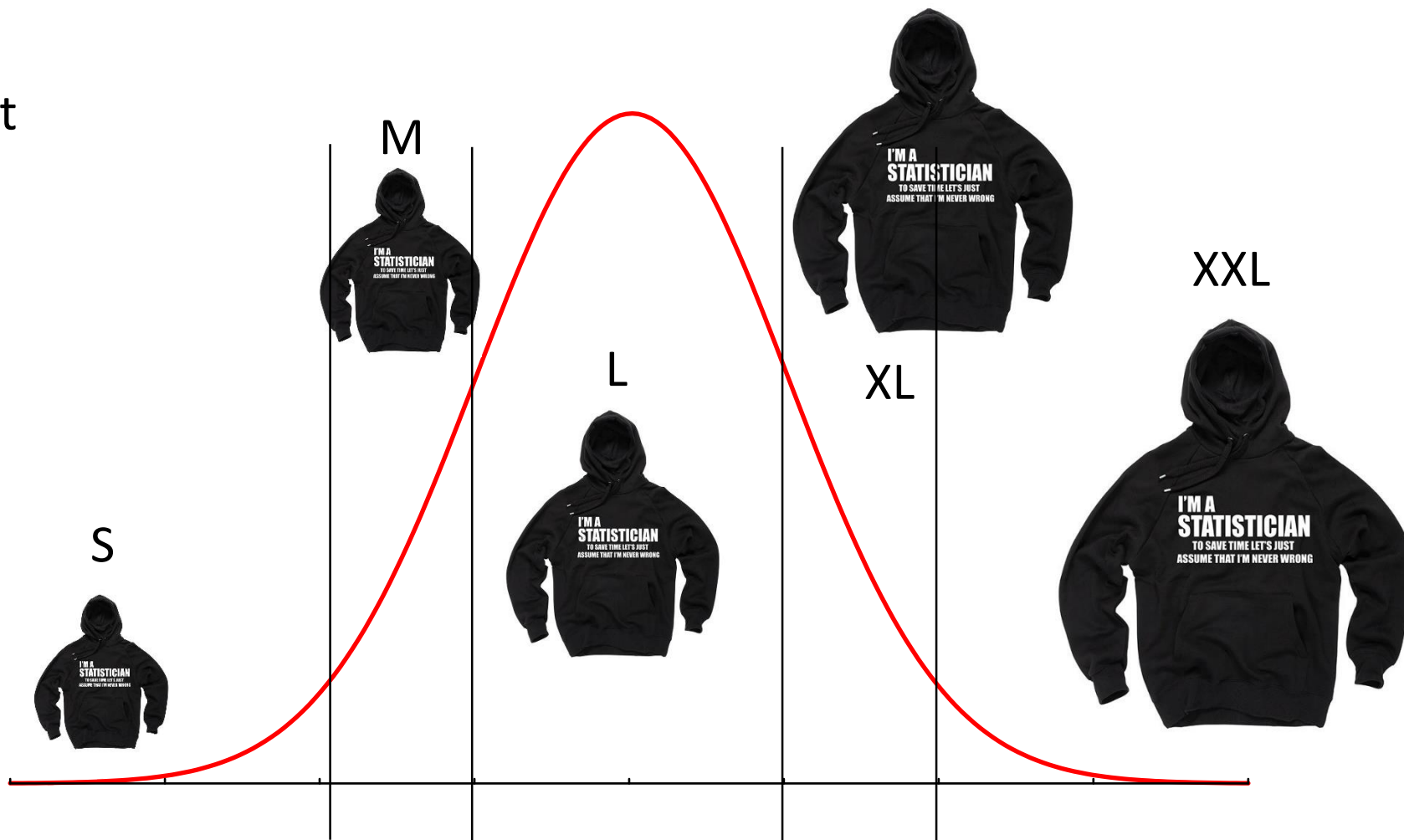


Příklad: optimalizace skladových zásob oblečení

- 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?

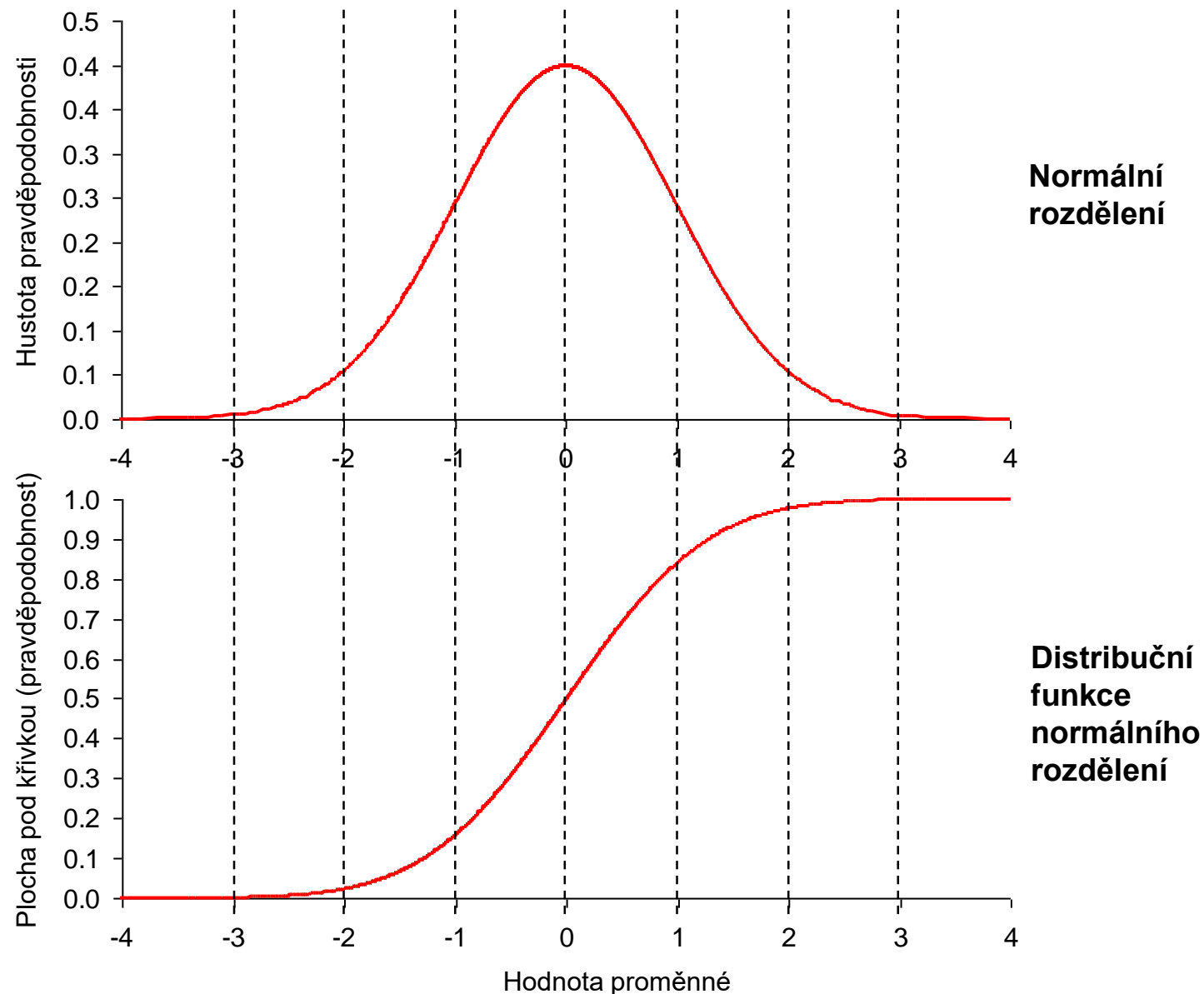
• Lze jednodušeji?



Velikost člověka relevantní k velikosti oblečení

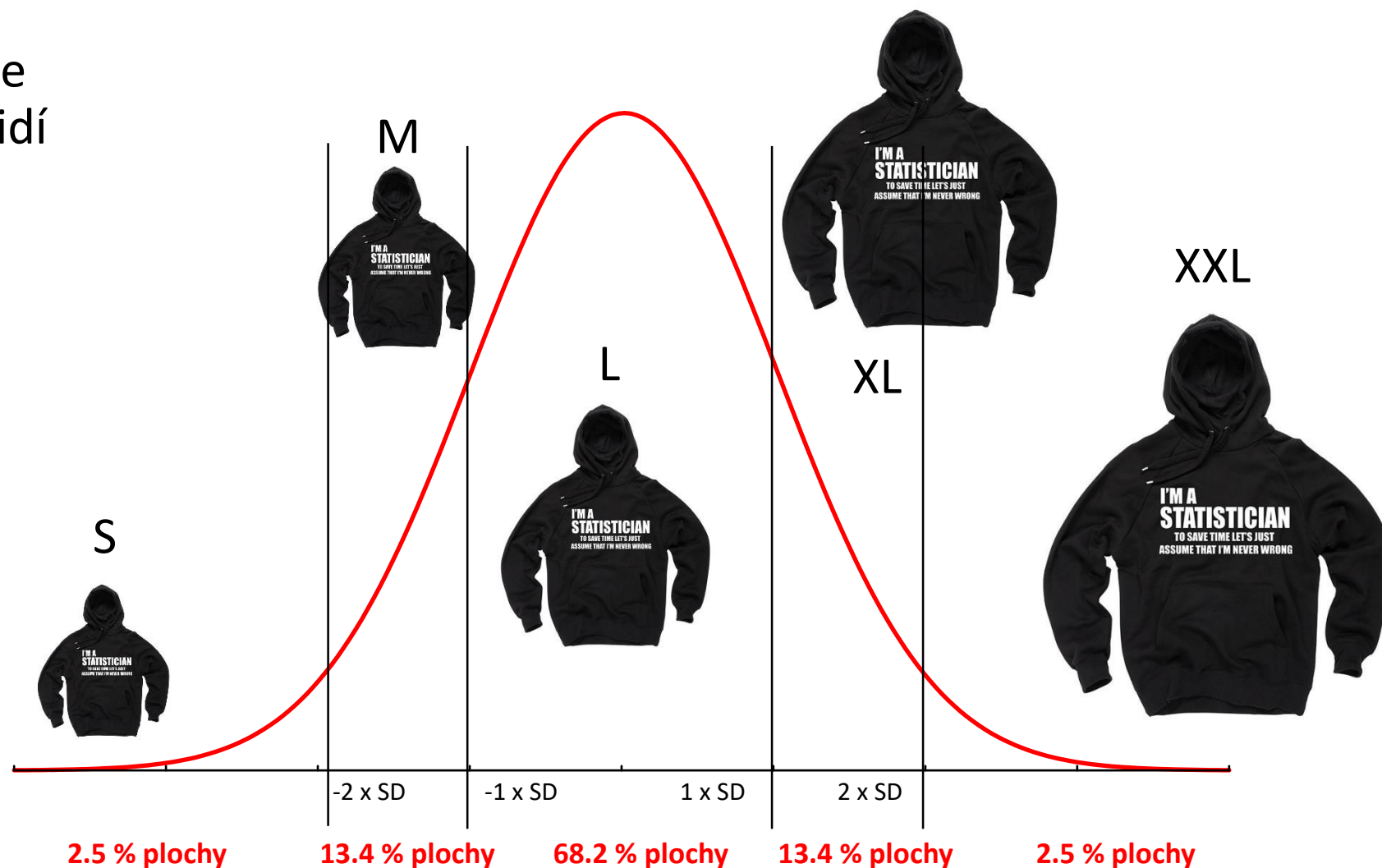
Normální rozdělení a jeho distribuční funkce

- 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 již odpovídá daná plocha pod křivkou rozdělení (např. 95% kvantil je hodnota proměnné pod níž leží 95% všech hodnot)



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

Přednáška 4

Modelová rozložení

Normální rozložení jako statistický model

Aplikace modelových rozložení

Přehled modelových rozložení

Anotace

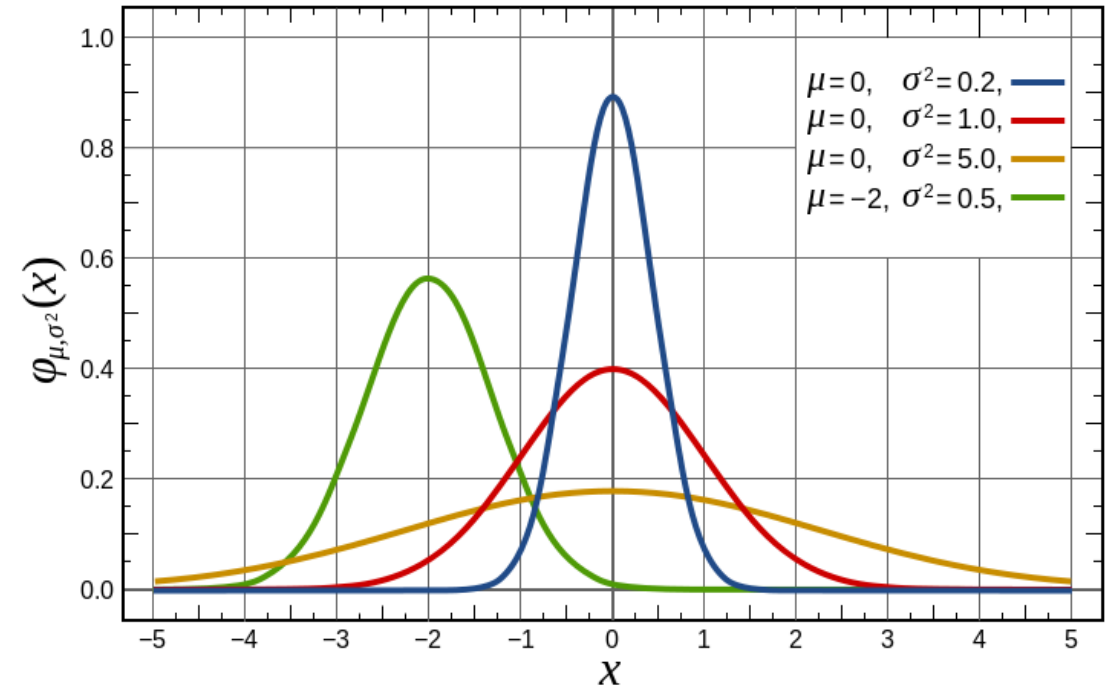
- Klasickým postupem statistické analýzy je na základě vzorku cílové populace identifikovat typ a charakteristiky modelového rozložení dat, využít jeho matematického modelu k popisu reality a získané výsledky zobecnit na hodnocenou cílovou populaci.
- Využití tohoto přístupu je možné pouze v případě shody reálných dat s modelovým rozložením, v opačném případě hrozí získání zavádějících výsledků.
- Nejklasičtějším modelovým rozložením, od něhož je odvozena celá řada statistických analýz je tzv. normální rozložení, známé též jako Gaussova křivka.

All models are wrong but some are useful.

George Box, 1978

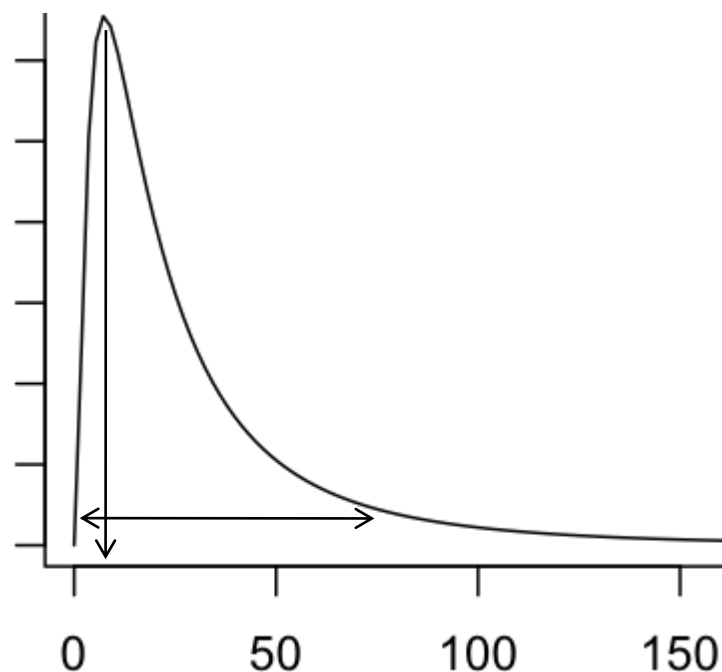
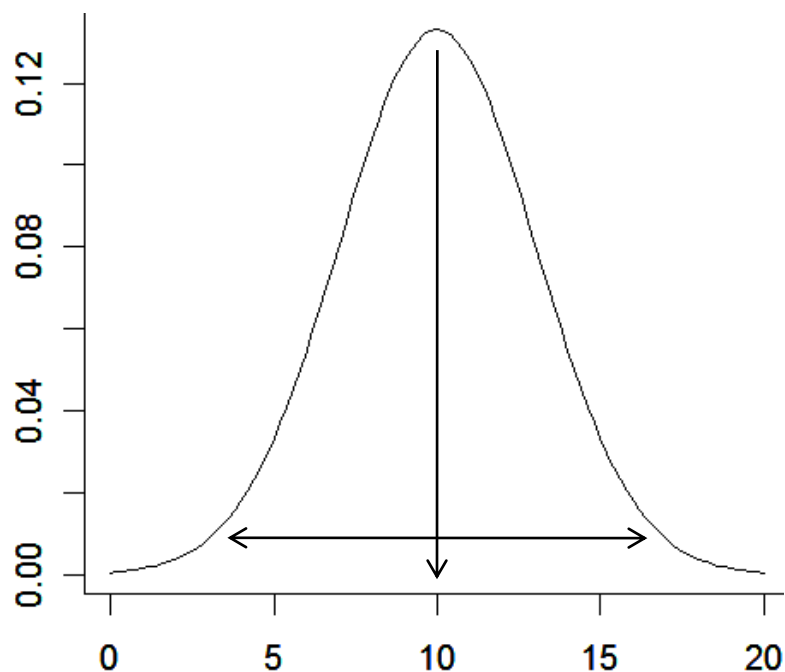
Normální rozdělení

- Nejklasičtějším modelovým rozložením, od něhož je odvozena celá řada statistických analýz je tzv. normální rozložení, známé též jako Gaussova křivka.
- Popisuje rozdělení pravděpodobnosti spojité náhodné veličiny: např. výška v populaci, chyba měření...
- Je kompletně popsáno dvěma parametry:
 - μ – střední hodnota
 - σ^2 – rozptyl
 - Označení: $N(\mu, \sigma^2)$
- Normalita je klíčovým předpokladem řady statistických metod
- Pro ověření normality existuje řada testů a grafických metod



Popis rozdělení kvantitativních dat: co chceme u dat popsat?

- Kvantitativní data – těžiště a rozsah pozorovaných hodnot.



Výpočet charakteristik normálního rozdělení: průměr

- μ – průměr rozdělení (cílová populace)
- \bar{x} – průměr rozložení vzorkovaných dat (odhad průměru cílové populace)
- Průměr lze spočítat z libovolných kvantitativních dat, ale pouze za některých situací jej lze považovat za ukazatel středu dat (symetrické, normální rozdělení dat)
- Odlehlé hodnoty a asymetrie dat výrazně ovlivňují výsledek výpočtu průměru

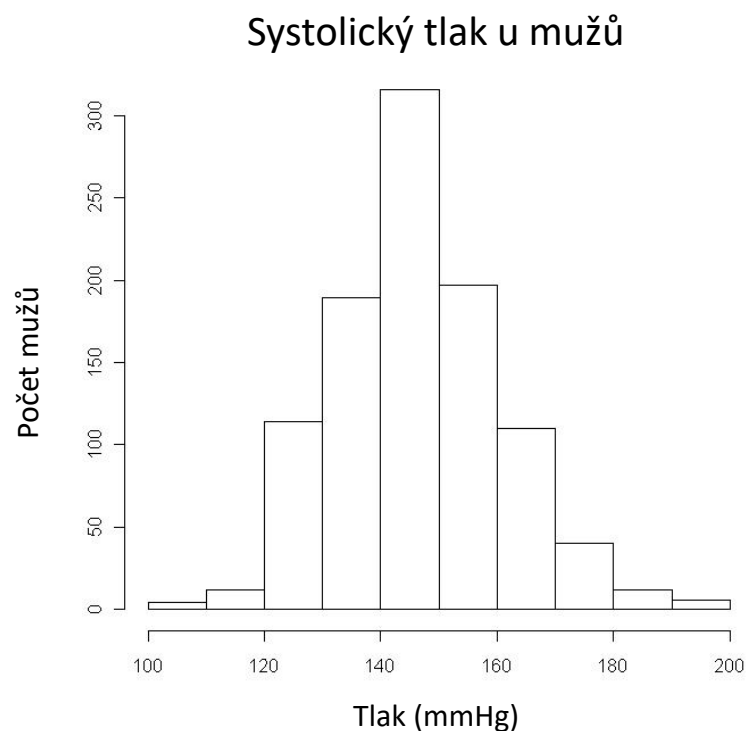
N=5

Objekt	Hodnota
x_1	5
x_2	3
x_3	4
x_4	7
x_5	2

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} = \frac{21}{5} = 4,2$$

Průměr vs. medián

- Máme-li symetrická data, je výsledek výpočtu průměru i mediánu podobný.
- Vše je OK.



Průměr = 149,9 mmHg



Medián = 150,0 mmHg

Průměr vs. medián

- Nemáme-li symetrická data, je výsledek výpočtu průměru i mediánu rozdílný.
- Není to OK. Výpočet průměru je v tuto chvíli nevhodný!

- **Příklad 1: známkování ve škole**

- Student A: 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 5

Průměr = 1,35

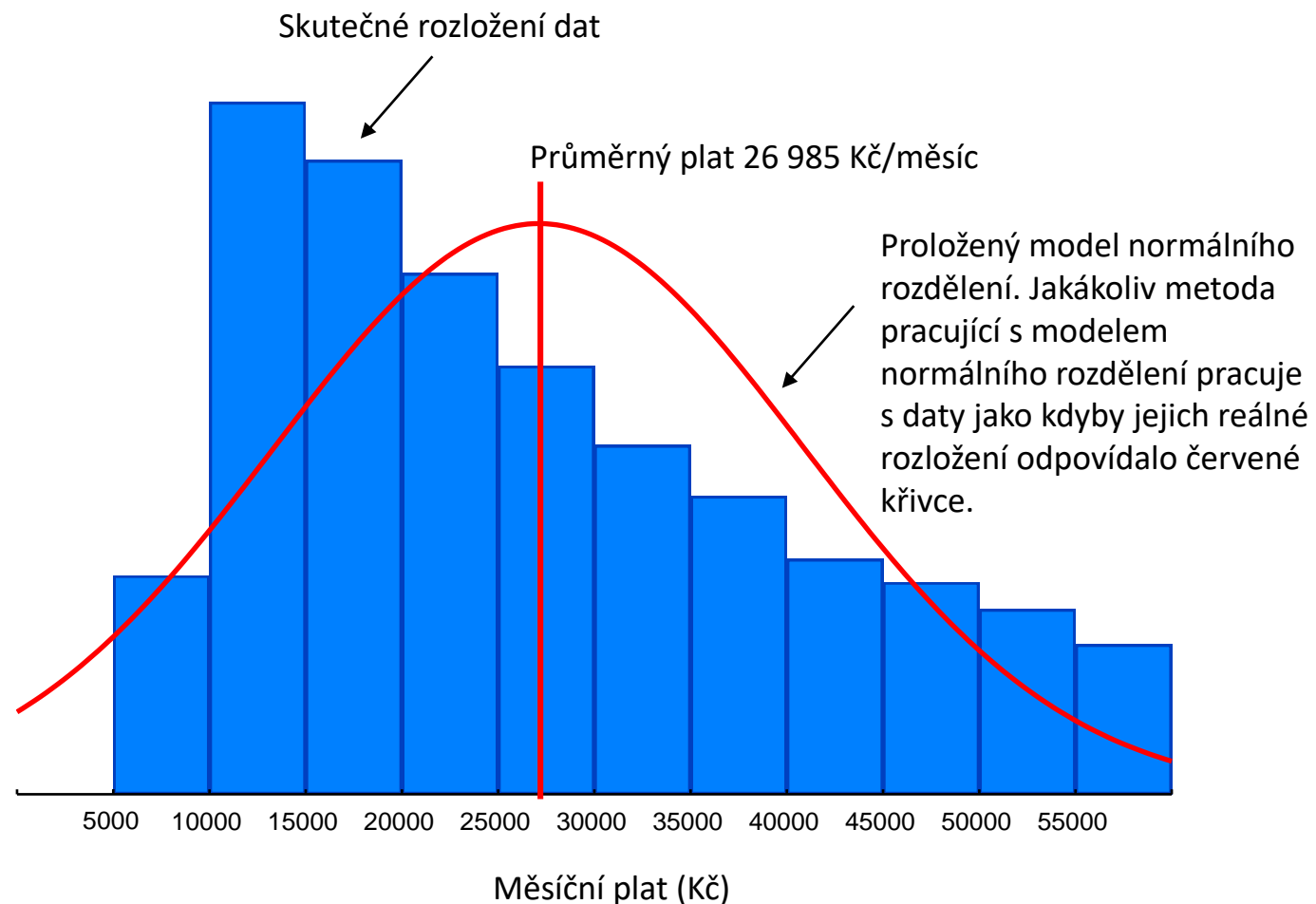
Medián = 1,00

- Student B: 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 2

Průměr = 1,13

Medián = 1,00

- **Příklad 2: plat v ČR**



Popis „těžiště“ – míry polohy

- Mějme pozorované hodnoty: x_1, x_2, \dots, x_n
- Seřadíme je podle velikosti: $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$

- **Minimum a maximum** – nejmenší a největší pozorovaná hodnota nám dávají obraz o tom, kde se na ose x pohybujeme.

$$x_{\min} = x_{(1)}$$

$$x_{\max} = x_{(n)}$$

- **Průměr** – charakterizuje hodnotu, kolem které kolísají ostatní pozorované hodnoty. Je to fyzikální obraz těžiště stejně hmotných bodů ose x .

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

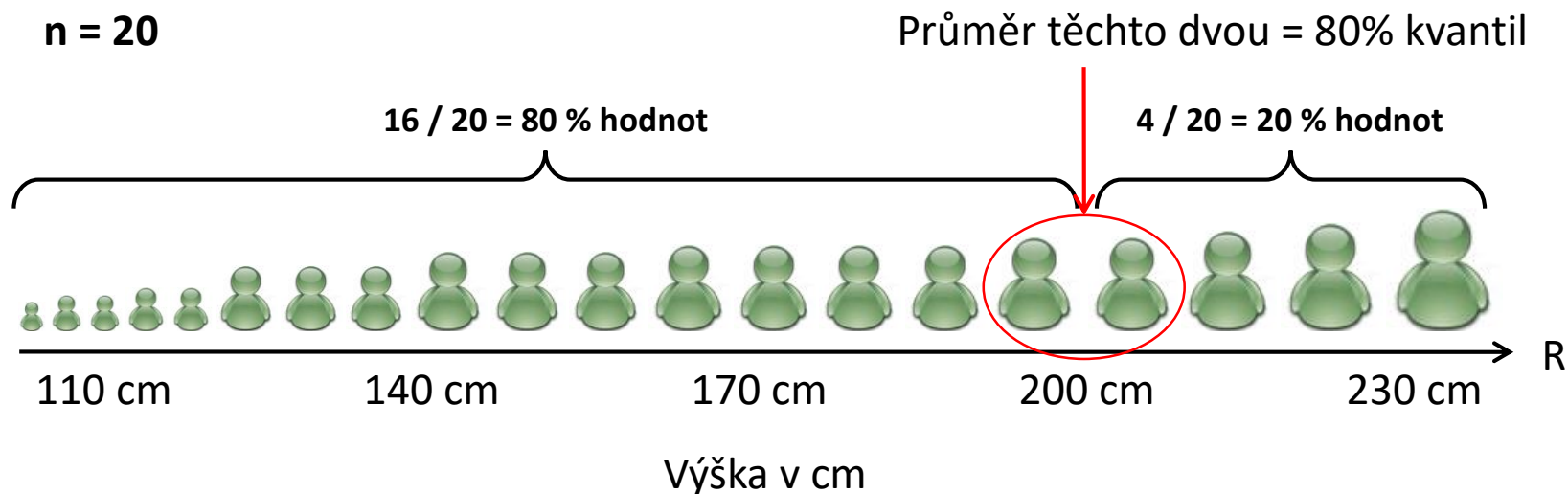
- **Medián** – je to prostřední pozorovaná hodnota. Dělí pozorované hodnoty na dvě půlky, půlka hodnot je menší a půlka hodnot je větší než medián.

$$\tilde{x} = x_{((n+1)/2)} \quad \text{pro } n \text{ liché}$$

$$\tilde{x} = \frac{1}{2} (x_{(n/2)} + x_{(n/2+1)}) \quad \text{pro } n \text{ sudé}$$

Pojem kvantil

- Laicky lze kvantil definovat jako číslo na reálné ose, které rozděluje pozorovaná data na dvě části: $p\%$ kvantil rozděluje data na $p\%$ hodnot a $(100-p)\%$ hodnot.
- Máme soubor 20 osob, u nichž měříme výšku. Chceme zjistit 80% kvantil souboru pozorovaných dat.

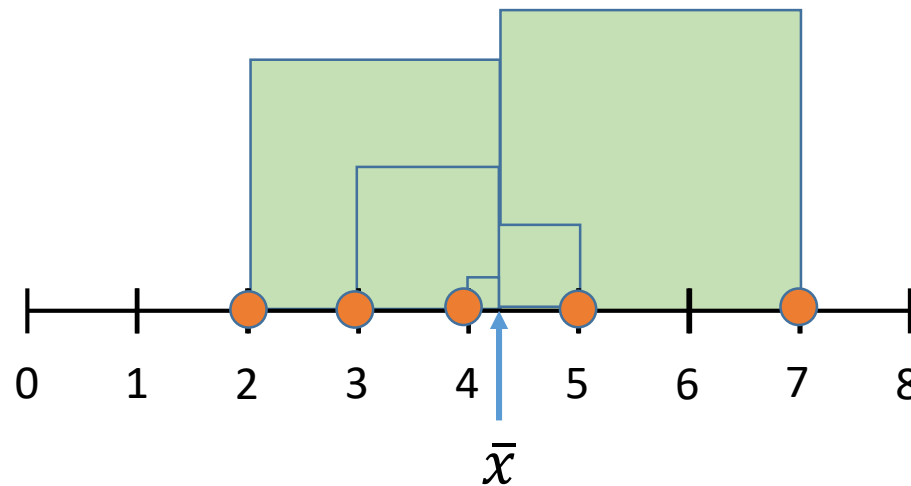


Výpočet charakteristik normálního rozdělení: rozptyl a směrodatná odchylka

- σ^2 – rozptyl rozdělení (cílová populace)
- s^2 – rozptyl rozložení vzorkovaných dat (odhad rozptylu cílové populace)

N=5

Objekt	Hodnota
x_1	5
x_2	3
x_3	4
x_4	7
x_5	2



$$s^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1} = \frac{14,8}{4} = 3,7$$

$$s = \sqrt{s^2} = \sqrt{3,7} = 1,92$$

- Směrodatná odchylka (s, SD=standard deviation) = druhá odmocnina z rozptylu (snazší interpretovatelnost)
- N-1 nebo N ? Dělení N-1 je výpočet rozptylu vzorku, dělení N je pro celou populaci (výjimečně)

Popis „rozsahu“ – míry variability

- Nejjednodušší charakteristikou variability pozorovaných dat je rozsah hodnot (rozpětí) = maximum – minimum. Je snadno ovlivnitelný netypickými (odlehými) hodnotami.
- **Kvantilové rozpětí** je definováno p% kvantilem a (100-p)% kvantilem a je méně ovlivněno odlehými hodnotami. Speciálním případem je kvartilové rozpětí, které pokrývá 50 % pozorovaných hodnot.
- **Rozptyl** – průměrný čtverec odchylky od průměru. Velmi ovlivnitelný odlehými hodnotami.

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)$$

- **Směrodatná odchylka** – odmocnina z rozptylu. Výhodou směrodatné odchylky je, že má stejné jednotky jako pozorovaná data.
- **Koeficient variance** - podíl směrodatné odchylky ku průměru (u normálního rozložení by se 95% hodnot mělo vejít do průměr ± 3 SD), pokud je SD větší než 1/3 průměru jsou teoreticky pravděpodobné záporné hodnoty v rozložení – ukazatel problémů s normalitou dat

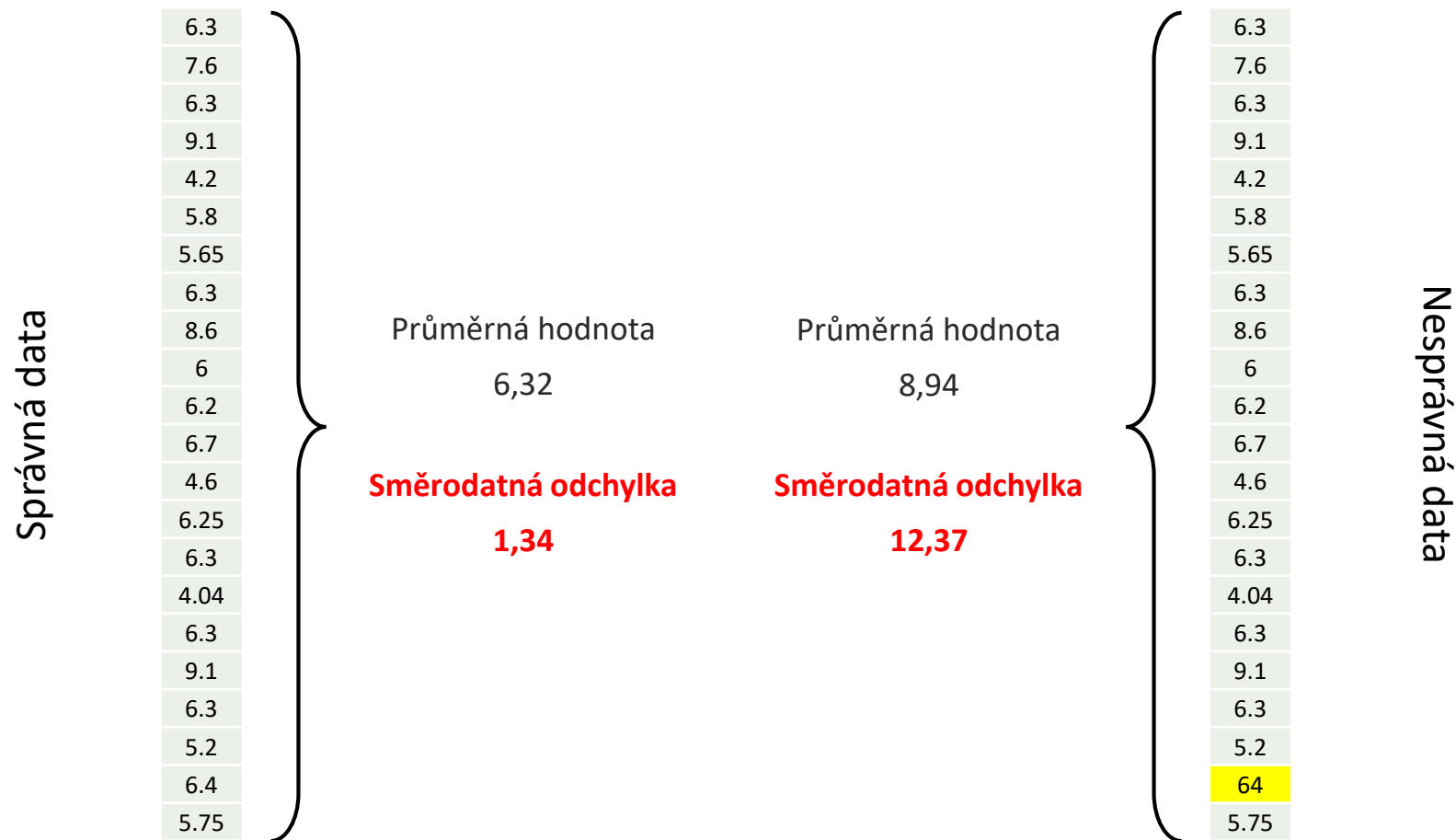
Normální rozdělení: vliv odlehlé hodnoty na popisné statistiky

- Cílem je určit průměrnou hladinu cholesterolu vybrané populace mužů (hodnoty v mmol/l)

Správná data	6.3		Průměrná hodnota 6,32	Průměrná hodnota ?		Nesprávná data	
	7.6						
	6.3						
	9.1						
	4.2						
	5.8						
	5.65						
	6.3						
	8.6						
	6						
	6.2						
	6.7						
	4.6						
	6.25						
	6.3						
	4.04						
	6.3						
	9.1						
	6.3						
	5.2						
	6.4						
	5.75						
		Směrodatná odchylka 1,34		Směrodatná odchylka ?			
		Která charakteristika se zvýší výrazněji? Průměr nebo směrodatná odchylka?					

Normální rozdělení: vliv odlehlé hodnoty na popisné statistiky

- Cílem je určit průměrnou hladinu cholesterolu vybrané populace mužů (hodnoty v mmol/l)



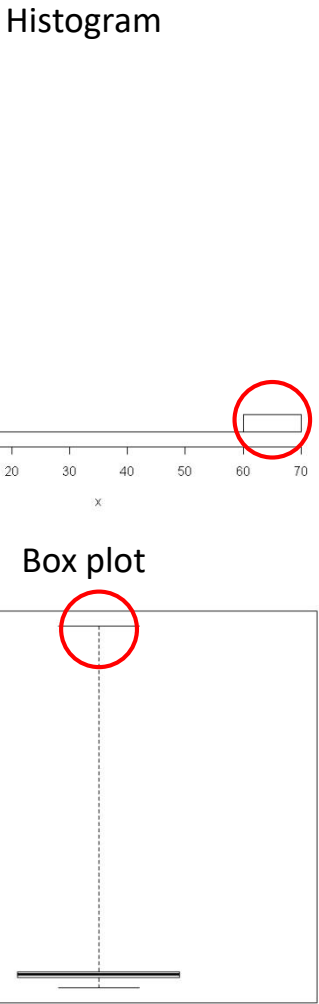
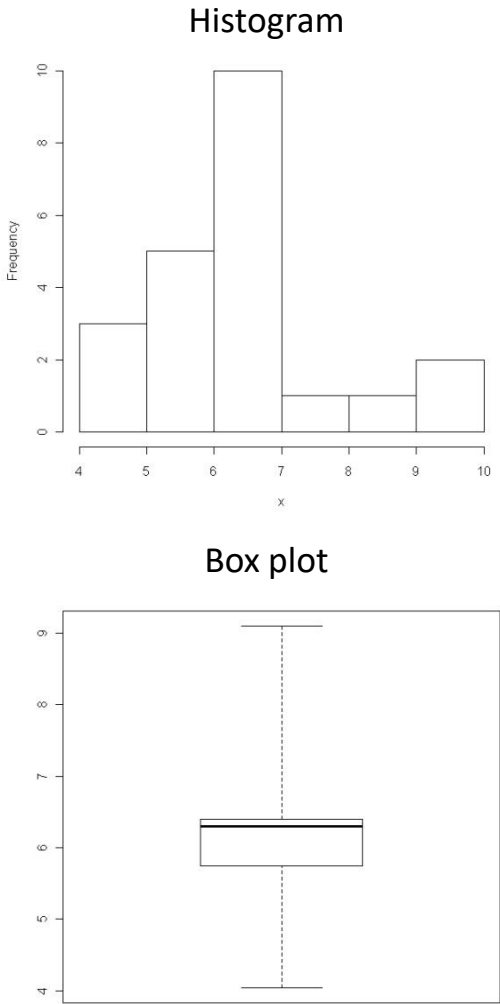
Identifikace odlehlých hodnot

- Na menších souborech stačí vizualizace.
- Na větších datových souborech nelze bez vizualizace a popisných statistik.
- Grafická identifikace: pomocí histogramu a box plotu.
- Identifikace pomocí popisných statistik: srovnání mediánu a průměru.

Identifikace odlehlých hodnot – příklad

Správná data

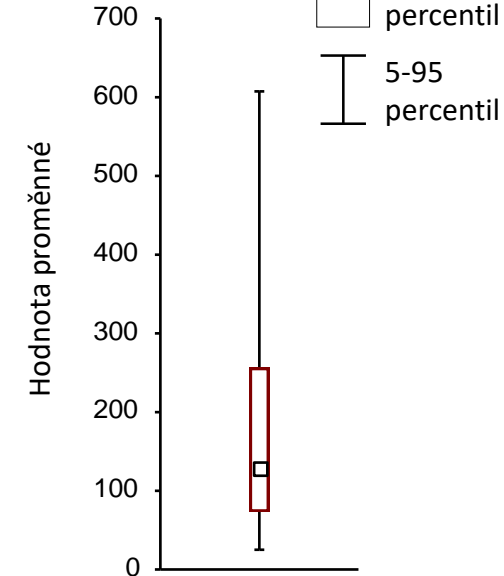
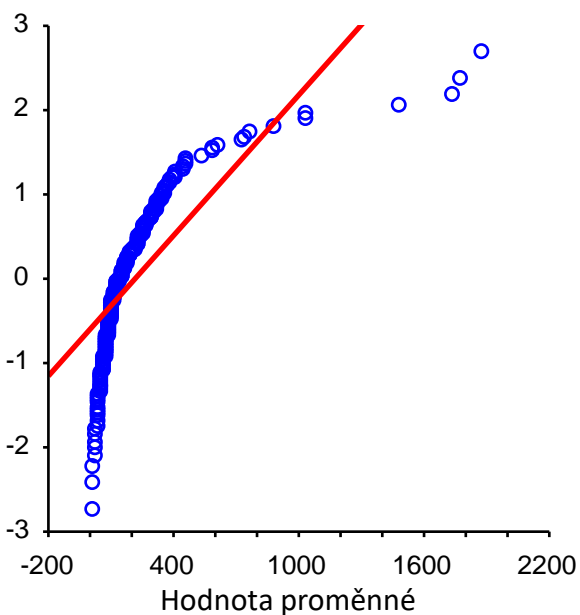
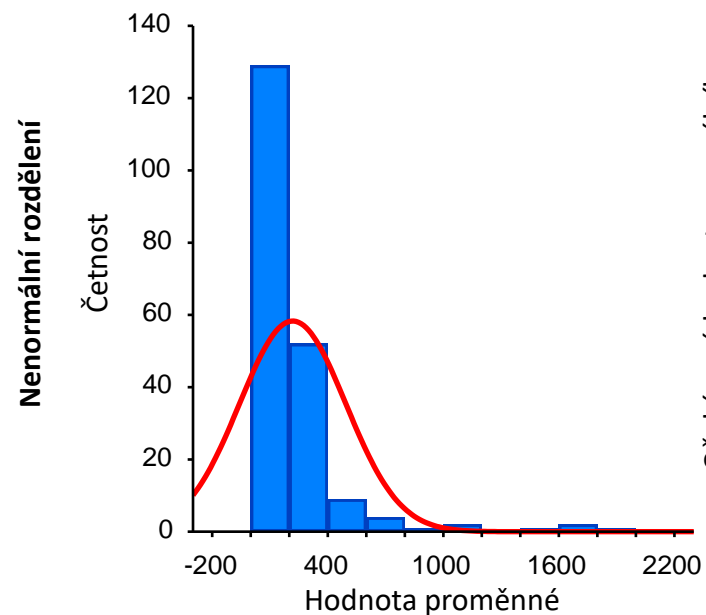
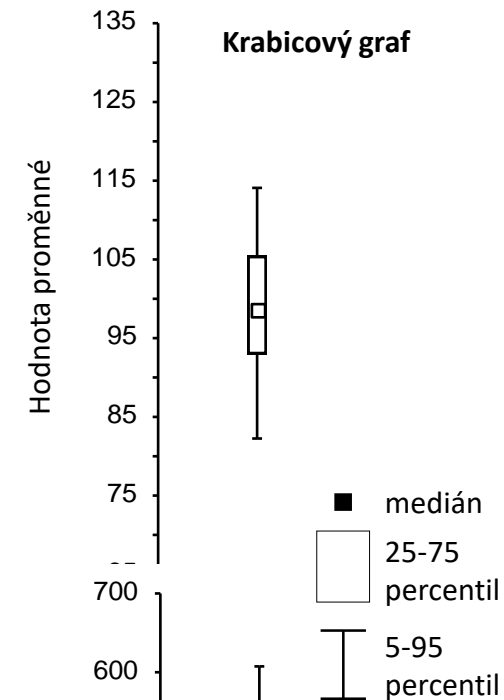
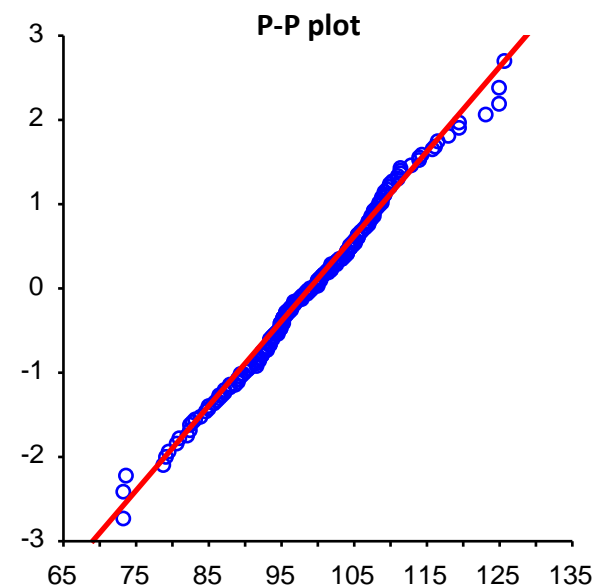
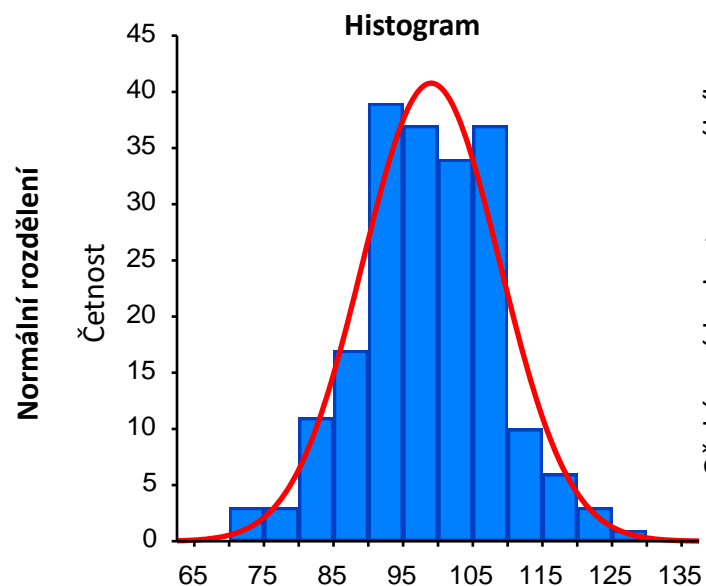
6.3
7.6
6.3
9.1
4.2
5.8
5.65
6.3
8.6
6
6.2
6.7
4.6
6.25
6.3
4.04
6.3
9.1
6.3
5.2
6.4
5.75



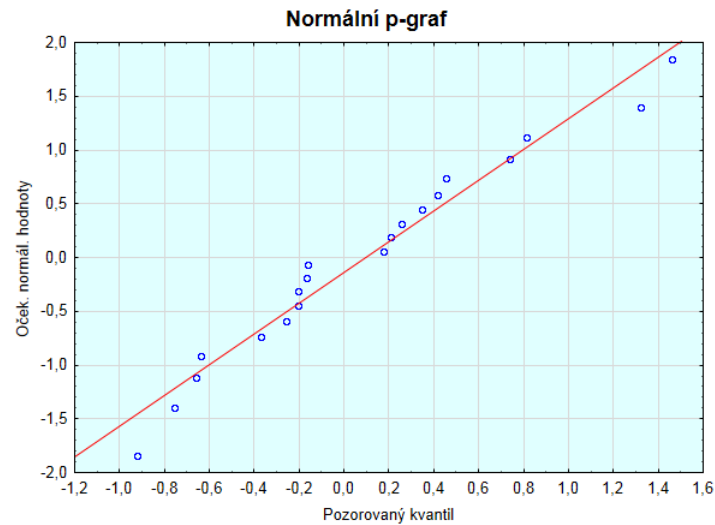
6.3
7.6
6.3
9.1
4.2
5.8
5.65
6.3
8.6
6
6.2
6.7
4.6
6.25
6.3
4.04
6.3
9.1
6.3
5.2
64
5.75

Nesprávná data

Vizuální hodnocení normality

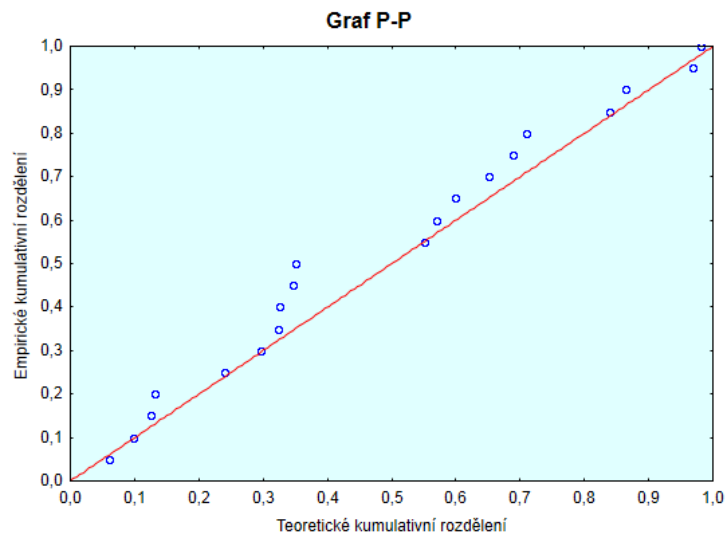
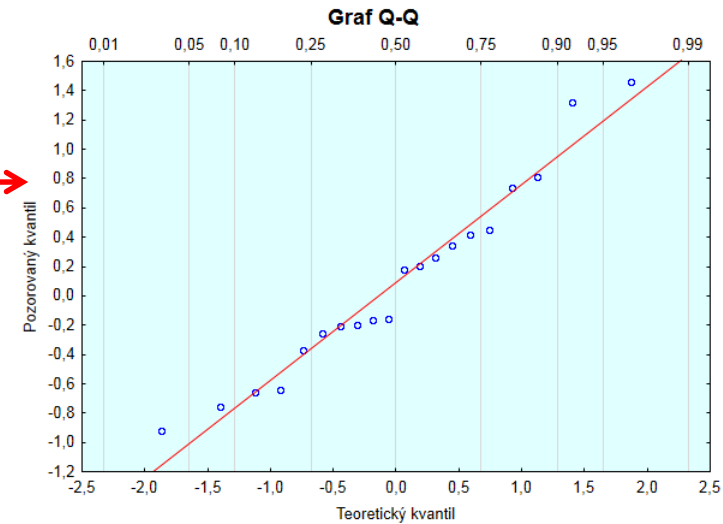


Rozdíl mezi N-P, Q-Q, P-P grafem



???

- Pouze výměna os
- Znázorněn pozorovaný a teoretický kvantil



- Vykresleno kumulativní rozdělení

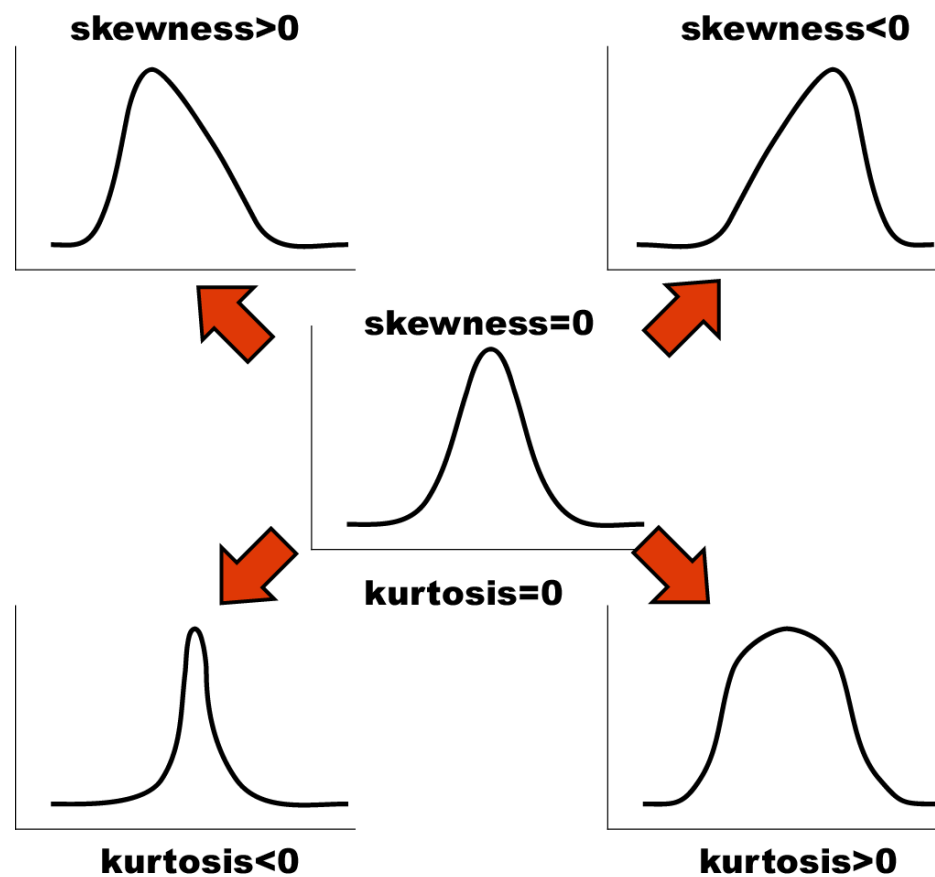
PAMATUJ:

**Pocházejí-li data z normálního rozložení,
pak body budou ležet okolo přímky**

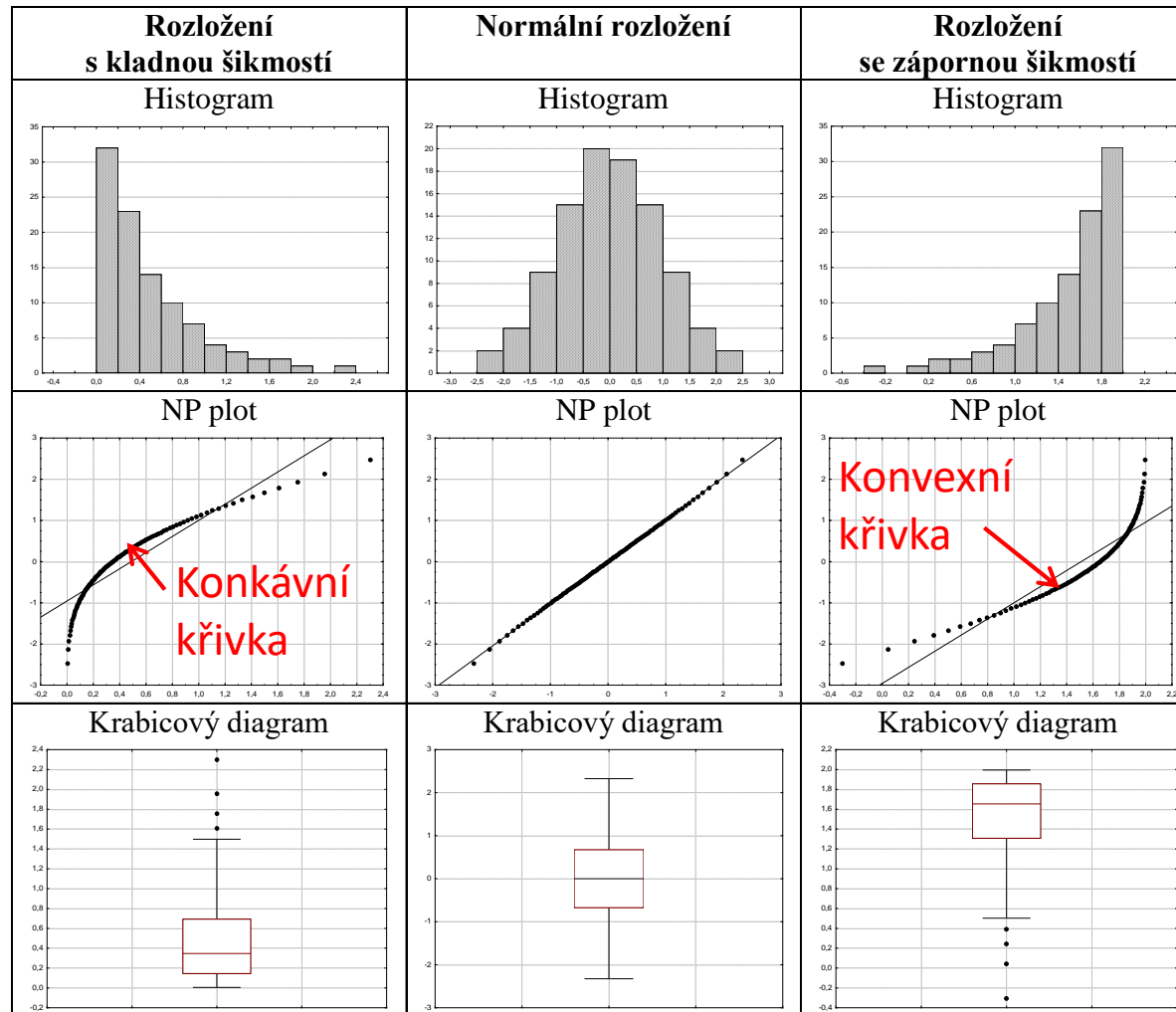


Ukazatele tvaru rozložení

- **Skewness** – ukazatel „šikmosti“ rozložení, asymetrie rozložení
- **Kurtosis** – ukazatel „špičatosti/plochosti“ rozložení



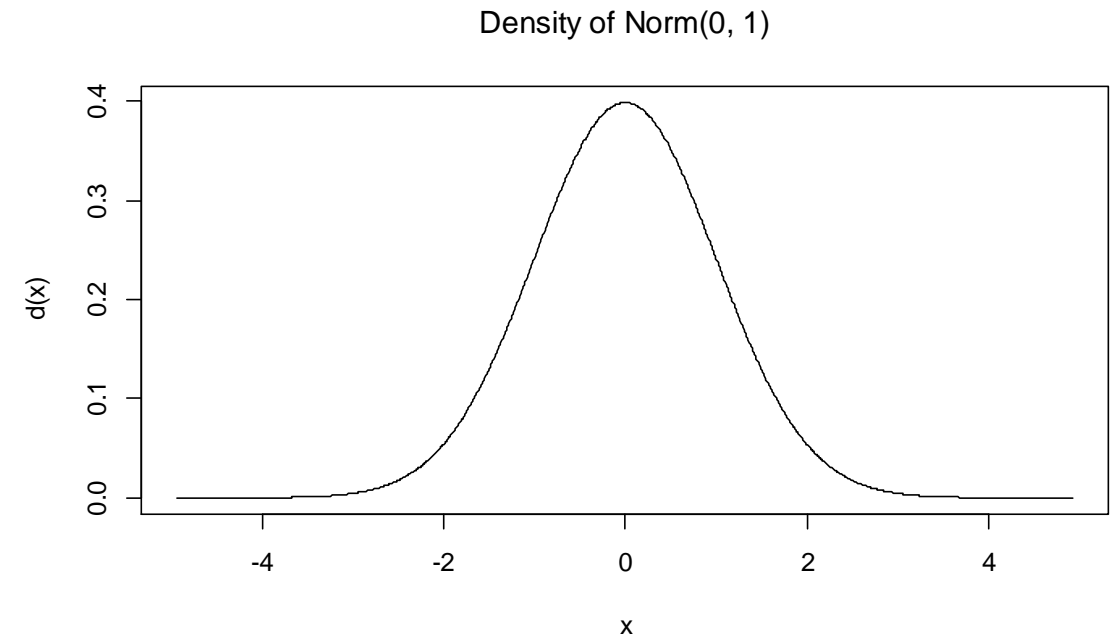
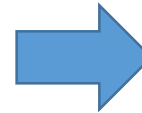
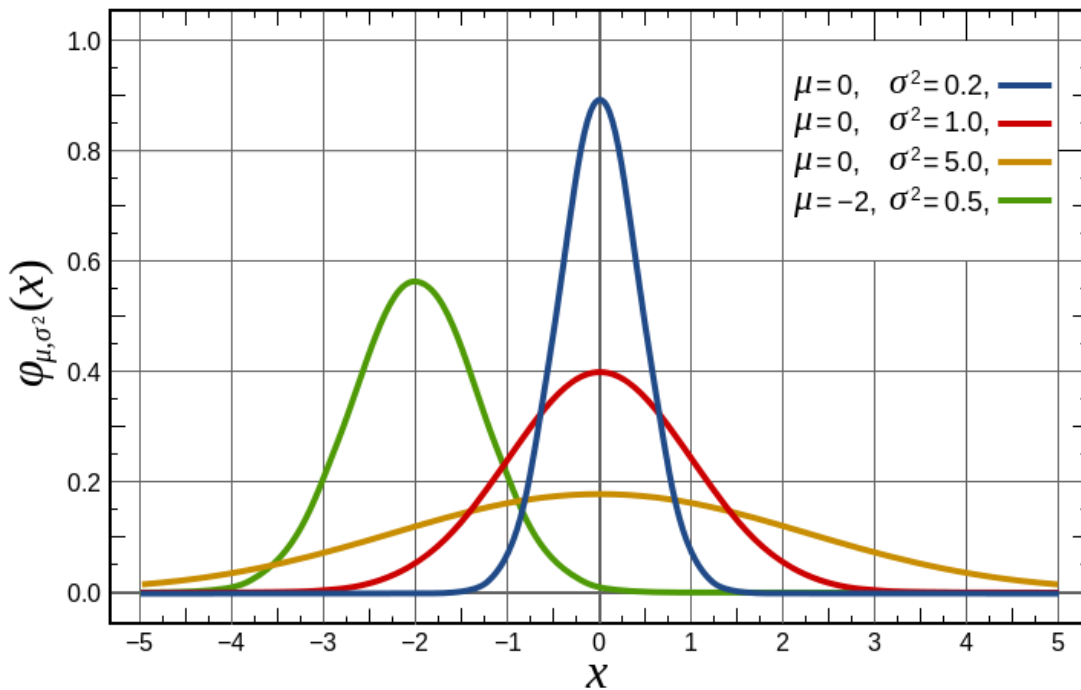
Jak se projeví asymetrie dat v diagnostických grafech?



Výukové materiály: Výpočetní statistika,
RNDr. Marie Budíková, Dr., 2011

Standardní normální rozdělení

- Speciální případ normálního rozdělení s $N(\mu=0, \sigma^2=1)$ - standardizovaná forma využívaná:
 - ve statistických výpočtech
 - pro srovnání extrémnosti / průměrnosti hodnot u proměnných s různými rozsahy nebo jednotkami
 - Jednoduchá interpretace – základní hodnoty vhodné zapamatovat



Přepočet na standardní normální rozdělení

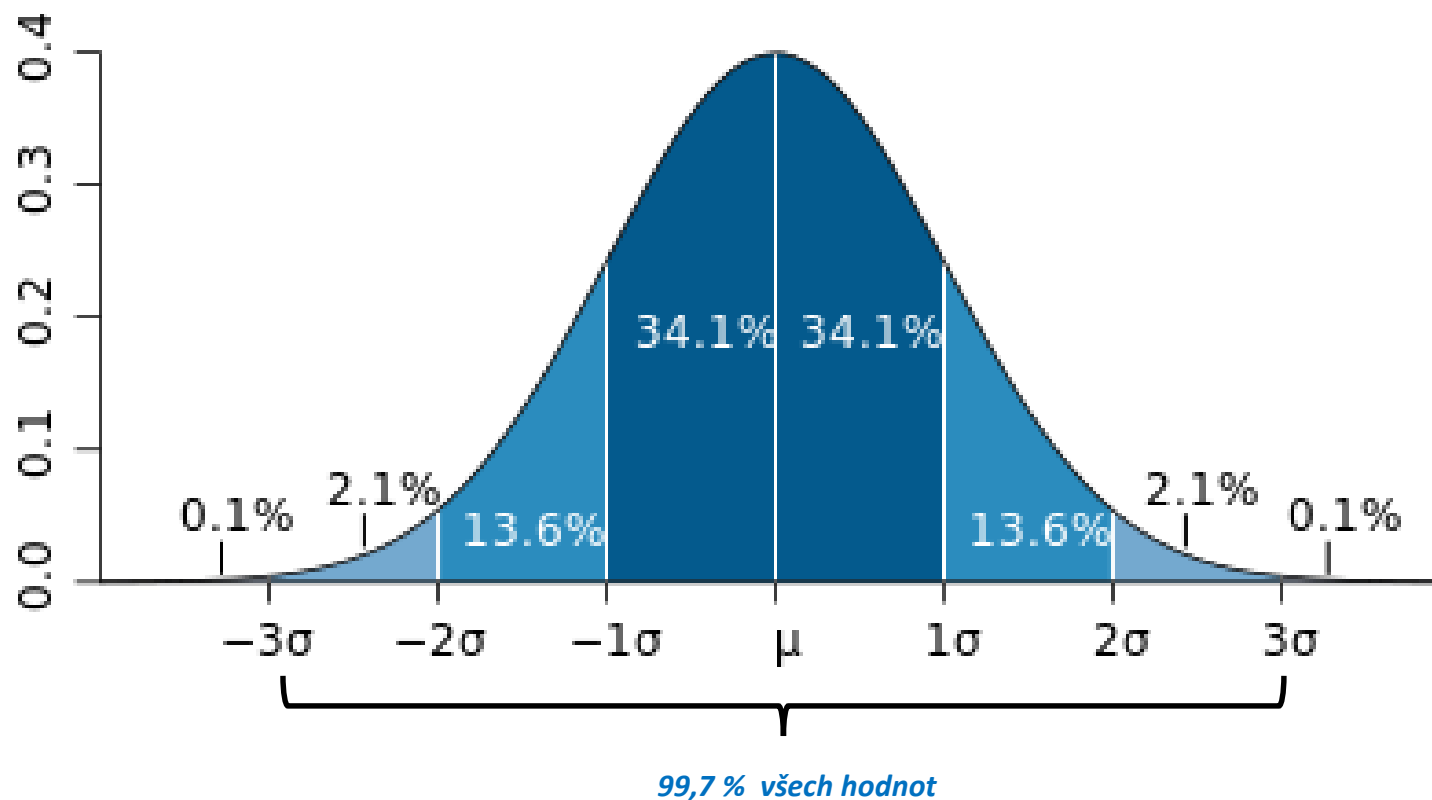
- Tzv. Z skóre – kromě statistických výpočtů využíváno např. v diagnostických skóre (osteoporóza) nebo pro srovnávání extrémnosti / průměrnosti proměnných s různými rozsahy nebo jednotkami (např. měření polutantů)
- Využití při výpočtu standardizovaných charakteristik (např. kovariance -> korelační koeficient)
- Ve vícerozměrné analýze používáno pro dosažení stejné váhy různých proměnných ve výpočtu
- Tabelovaná forma -> využití ve výpočtech

Objekt	Hodnota	Standardizovaná hodnota (z)
x_1	5	0.42
x_2	3	-0.62
x_3	4	-0.10
x_4	7	1.46
x_5	2	-1.14
průměr	4,2	0
s	1,92	1

$$Z_i = \frac{x_i - \mu}{\sigma}$$

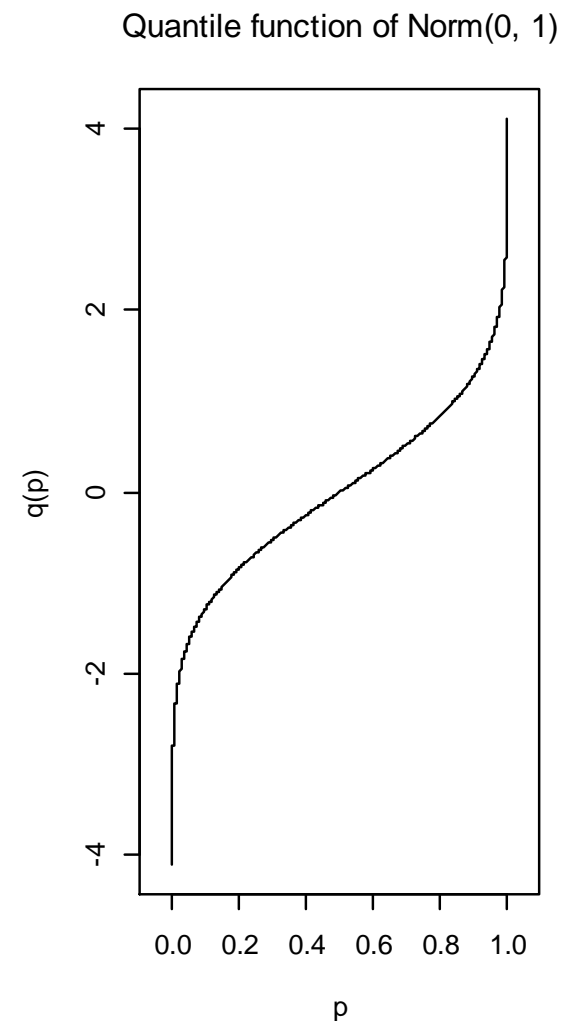
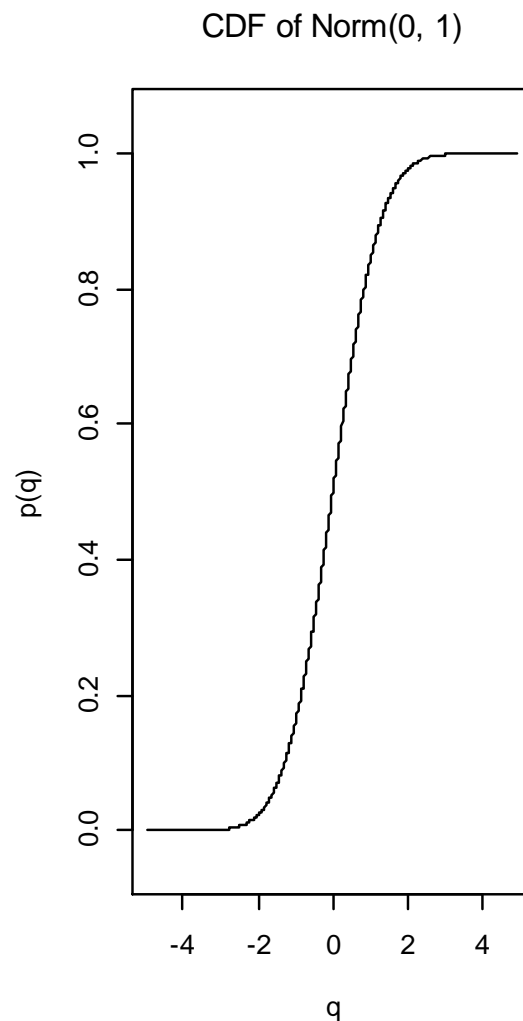
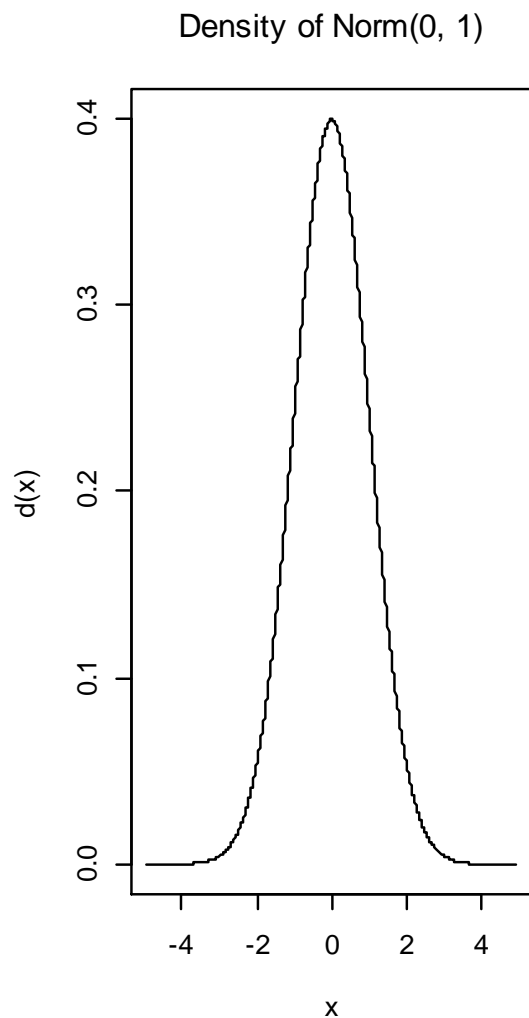
Pravidlo 3 sigma

- V rozmezí $\mu \pm 3\sigma$ by se mělo vyskytovat 99,7 % všech hodnot
- Vhodné znát pro orientační posouzení rozsahu dat
- U proměnných, které nemohou být záporné využití pro orientační posouzení normality



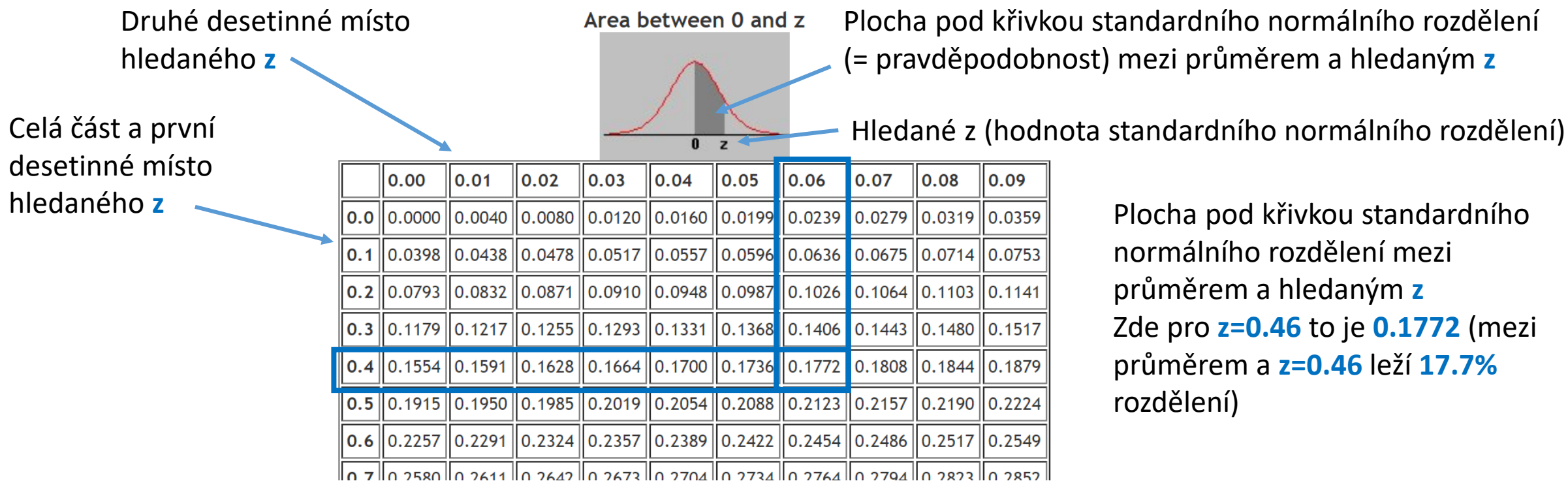
99,7 % všech hodnot

Standardizované normální rozdělení a jeho charakteristiky



Statistické tabulky

- Přehledné vyjádření distribuční funkce pro modelová rozdělení
- V předpočítačovém období základní pomůcka, nyní hlavně výukový význam
- <http://www.statsoft.com/Textbook/Distribution-Tables> (potřebné i pro zkoušku)

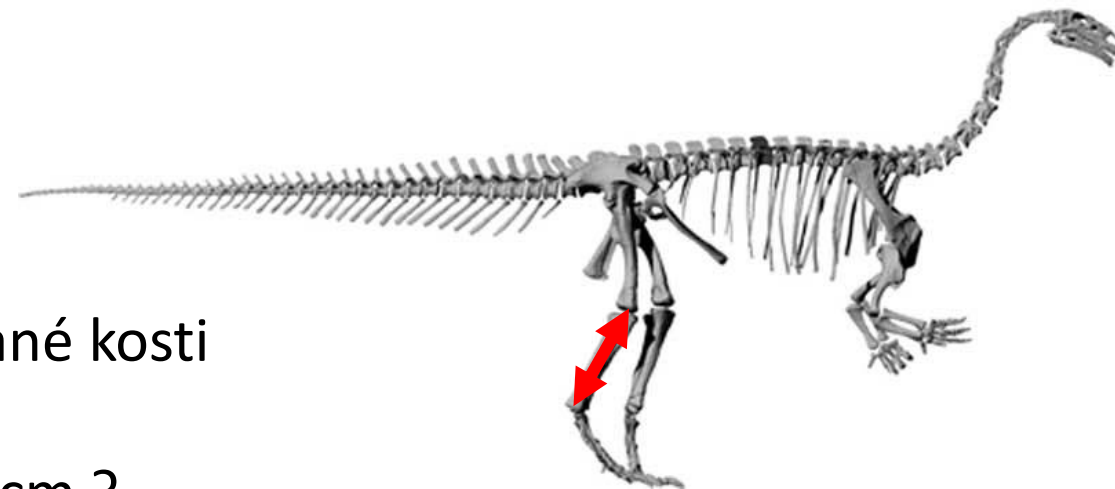


Využití statistických modelů

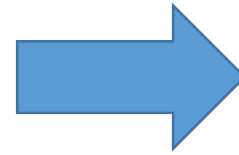
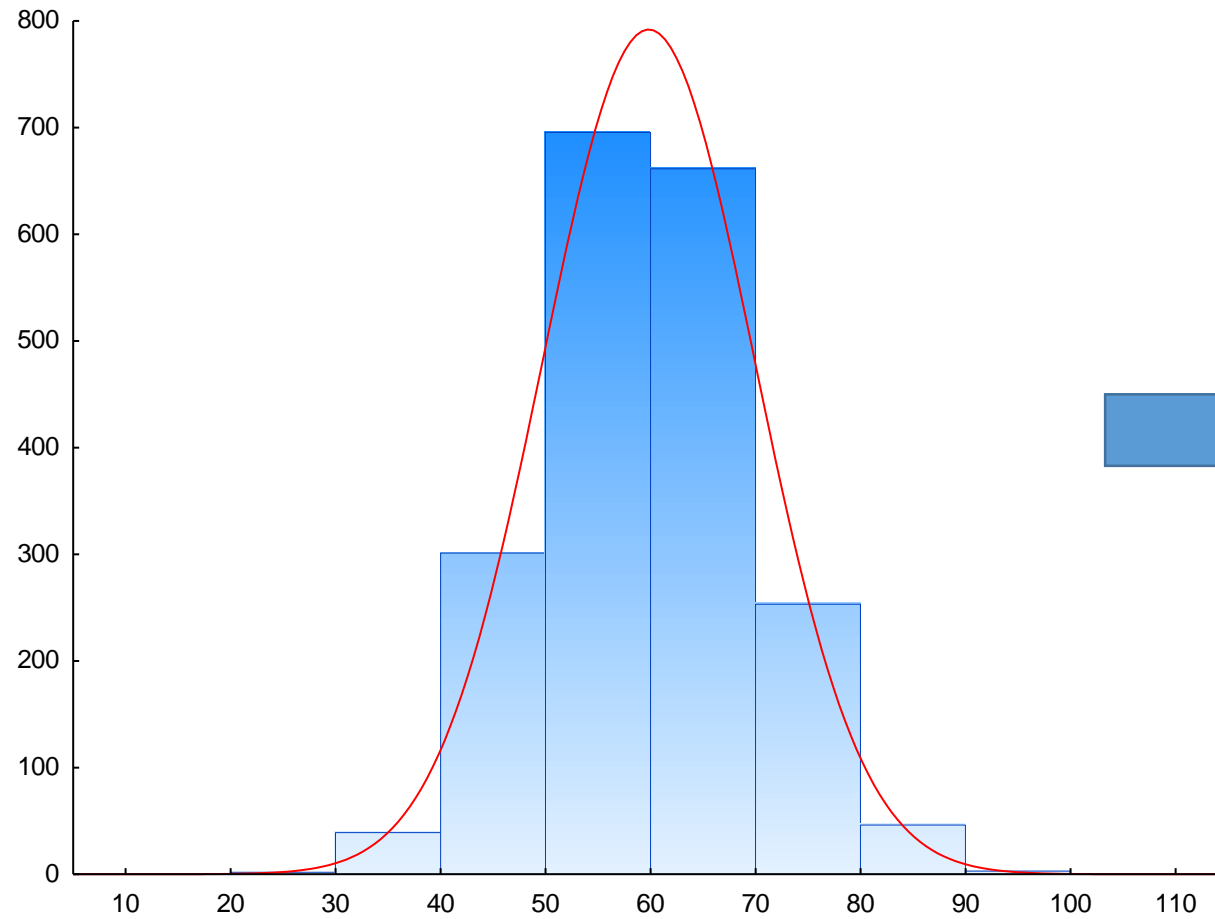
1. Máme nějaký znak v populaci, který chceme pro účely analýz nahradit statistickým modelem (de facto to děláme při každém výpočtu průměru, který považujeme za ukazatel středu)
2. Ověříme předpoklad, že je znak rozložen podle daného modelu = **Platí vybraný model?** Např. vizuální posouzení normality nebo její testování.
3. Spočítáme charakteristiky modelu (průměr a směrodatná odchylka v případě normálního rozdělení)
4. Převedeme na standardní formu modelu (standardní normální rozdělení v případě normálního rozdělení)
5. Využijeme známé vlastnosti rozdělení pro odpověď na položené otázky (distribuční funkce, její hodnoty ve statistických tabulkách)

Příklad aplikace modelu normálního rozdělení

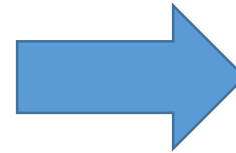
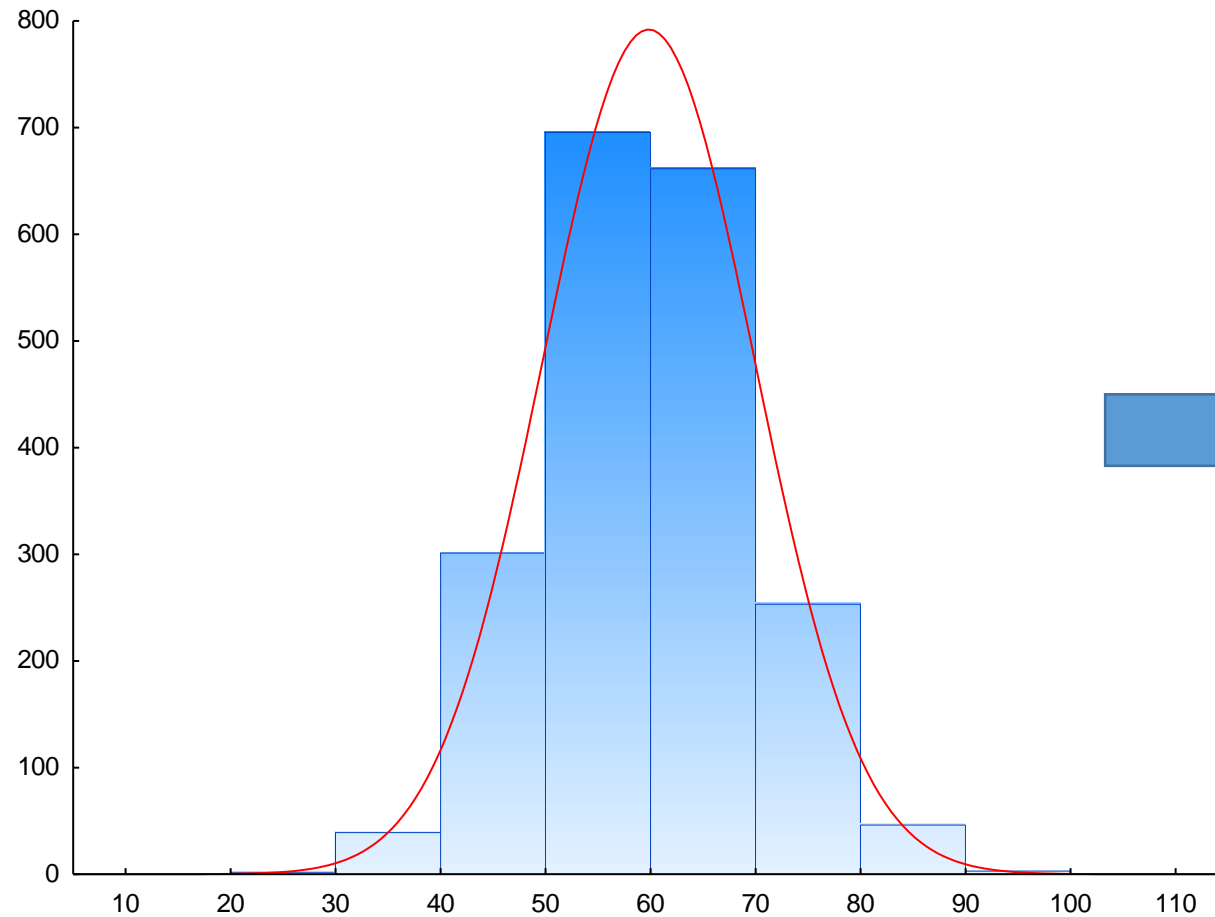
- Máme data z průzkumu kostí prehistorického zvířete
 - $N=2\ 000$
 - Průměrná délka = 60 cm
 - Směrodatná odchylka = 10 cm
- Výzkumné otázky:
 - Jaká je pravděpodobnost, že by velikost dané kosti překročila velikost 66 cm?
 - Kolik kostí mělo zřejmě délku větší než 66 cm ?
 - Jaký podíl kostí ležel svou délkou v rozsahu od 60 cm do 66 cm ?



Ověření rozložení dat a výběr statistického modelu



Ověření rozložení dat a výběr statistického modelu



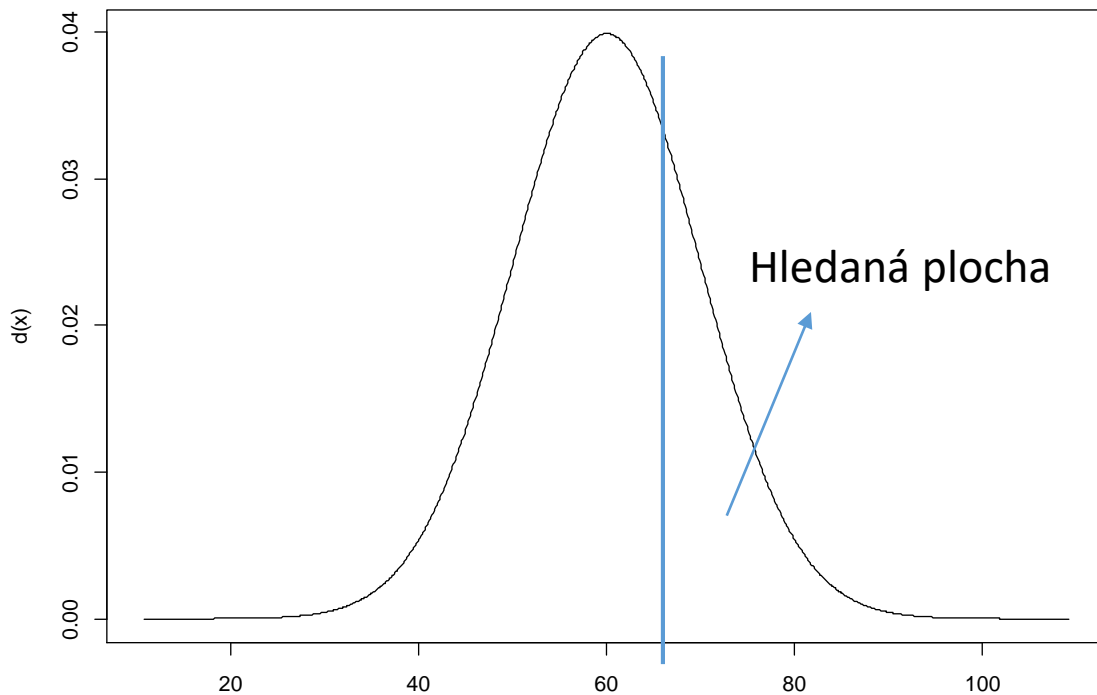
**Předpoklad normálního
rozdělení dat se zdá oprávněný.**

Jaká je pravděpodobnost, že by velikost dané kosti překročila velikost 66 cm?

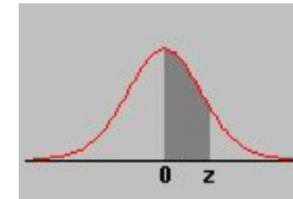
- Přepočet hledané hodnoty na standardizovanou formu normálního rozdělení

$$z = \frac{x - \mu}{\sigma} = \frac{66 - 60}{10} = 0,6$$

Density of Norm(60, 10)



Area between 0 and z



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852

$$P(x > 66) = 1 - P(x \leq 66) = 1 - P\left(\frac{x - m}{s} \leq \frac{66 - 60}{10}\right) = 1 - F(0,6) = 0,27425$$

Aplikace modelu normálního rozdělení

- Kolik kostí mělo zřejmě délku větší než 66 cm ?

$$P(x > 66) * n = 0,27425 * 2000 = 548$$

- Jaký podíl kostí ležel svou délkou v rozsahu x od 60 cm do 66 cm ?

$$P(60 < x < 66) = P\left(\frac{60-60}{10} < Z < \frac{66-60}{10}\right) = F(0,6) - F(0) = 0,22575$$

- 22,6% kostí leží v rozsahu 60-66cm

Stručný přehled modelových rozložení I

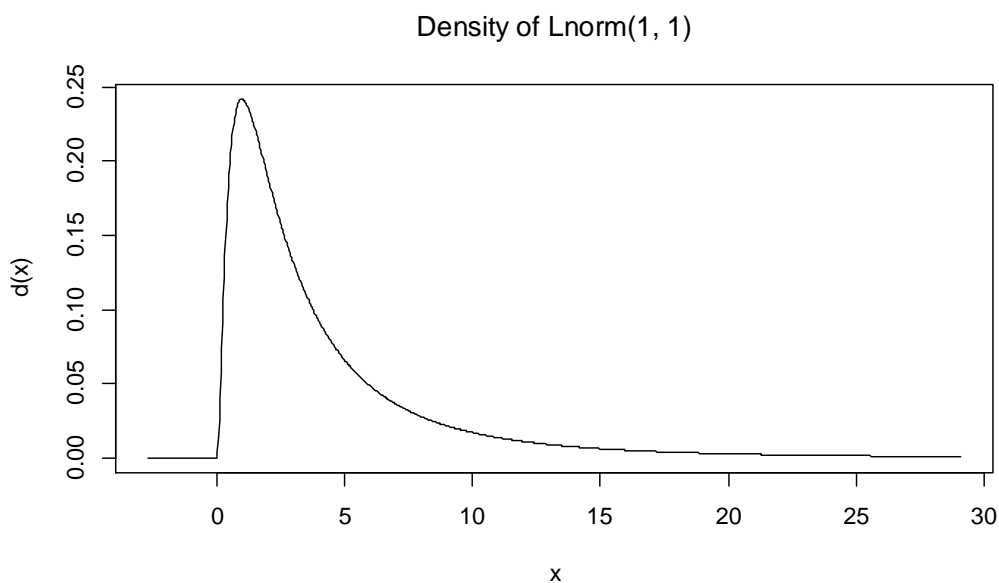
Rozložení	Parametry	Stručný popis
<u>Normální</u>	Průměr (μ) Rozptyl (σ^2)	Symetrická funkce popisující intervalovou hustotu četnosti; nejpravděpodobnější jsou průměrné hodnoty znaku v populaci.
<u>Log-normální</u>	Medián Geometrický průměr Rozptyl (σ^2)	Funkce intervalové hustoty četnosti, která po logaritmické transformaci nabude tvaru normálního rozložení.
Weibullovo	α - parametr tvaru β - parametr rozsahu hodnot	Změnou parametru α lze modelovat distribuci doby přežití, např. stresovaného organismu. Rozložení využívané i jako model k odhadu LC_{50} nebo EC_{50} u testů toxicity.
Rovnoměrné	Medián Geometrický průměr Rozptyl (σ^2)	Funkce intervalové hustoty četnosti, která po logaritmické transformaci nabude tvaru normálního rozložení.
Triangulární	$f(x) = [b - \text{ABS}(x - a)] / b^2$ $a - b < x < a + b$	Pravděpodobnostní funkce pro typ rozložení, kdy jsou střední hodnoty výrazně pravděpodobnější než hodnoty okrajové.
Gamma	Parametry distribuční funkce: α - parametr tvaru β - parametr rozsahu hodnot	Umožňuje flexibilně modelování distribučních funkcí nejrůznějších tvarů. Např. χ^2 rozložení je rozložení typu Gamma. Gamma rozložení s $\alpha = 1$ je známo jako exponenciální rozložení.

Stručný přehled modelových rozložení

Rozložení	Parametry	Stručný popis
Beta	Parametry distribuční funkce: α - parametr tvaru β - parametr rozsahu hodnot	Pravděpodobnostní funkce pro proměnnou omezenou rozsahem do intervalu [0; 1]. Je matematicky komplikovanější, ale velmi flexibilní při popisu změn hodnot proměnné v ohraničeném intervalu.
<u>Studentovo</u>	Stupně volnosti - uvažuje velikost vzorku Průměr Rozptyl	Simuluje normální rozložení pro menší vzorky čísel. Pro větší soubory ($n > 100$) se limitně blíží k normálnímu rozložení.
<u>Pearsonovo</u>	Stupně volnosti - uvažuje velikost vzorku	Slouží především k porovnání četností jevů ve dvou a více kategoriích. Používá se k modelování rozložení odhadu rozptylu normálně rozložených dat.
<u>Fisher-Snedecorovo</u>	Dvojí stupně volnosti - uvažuje velikost dvou vzorků	Používá se k testování hodnot průměrů - F test pro porovnání dvou výběrových rozptylů; F test, ANOVA atd.

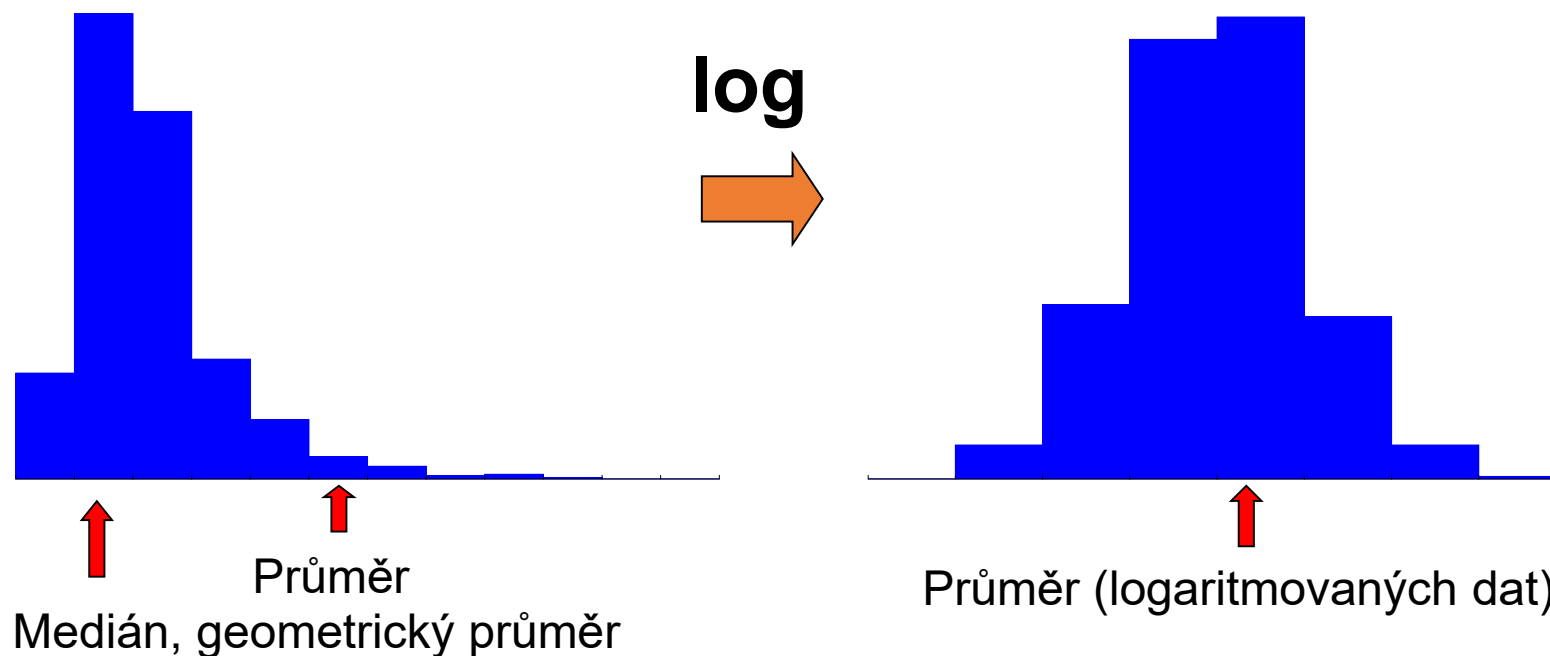
Lognormální rozdělení

- Asymetricky rozložená data – velmi častá v biologii (ale i jinde, např. platy)
- Spolu s normálním rozdělením nejčastější model
- S rozdělením je spjat geometrický průměr jako ukazatel středu

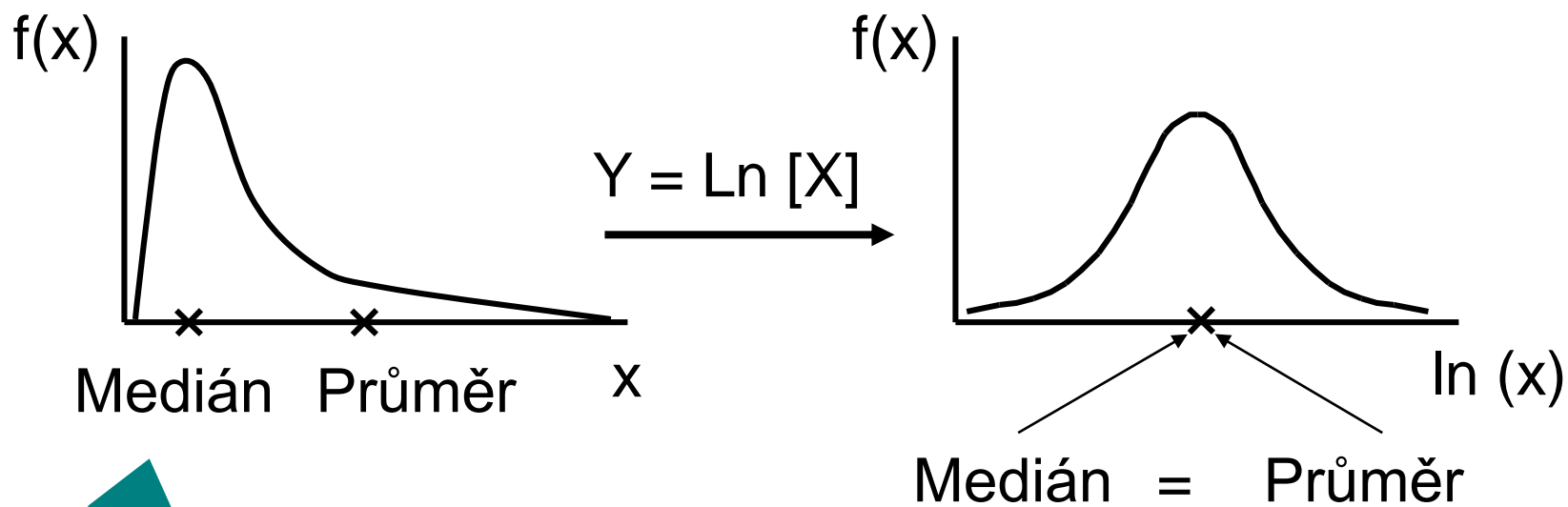


Logaritmická transformace

- Geometrický průměr – antilogaritmus průměru logaritmovaných dat, je vhodný pro doleva asymetrická data (lognormální rozložení), která jsou v biologii velmi častá, jeho hodnota v podstatě odpovídá mediánu
- Takto asymetrická data je možné převést logaritmickou transformací na normální rozložení



Geometrický průměr



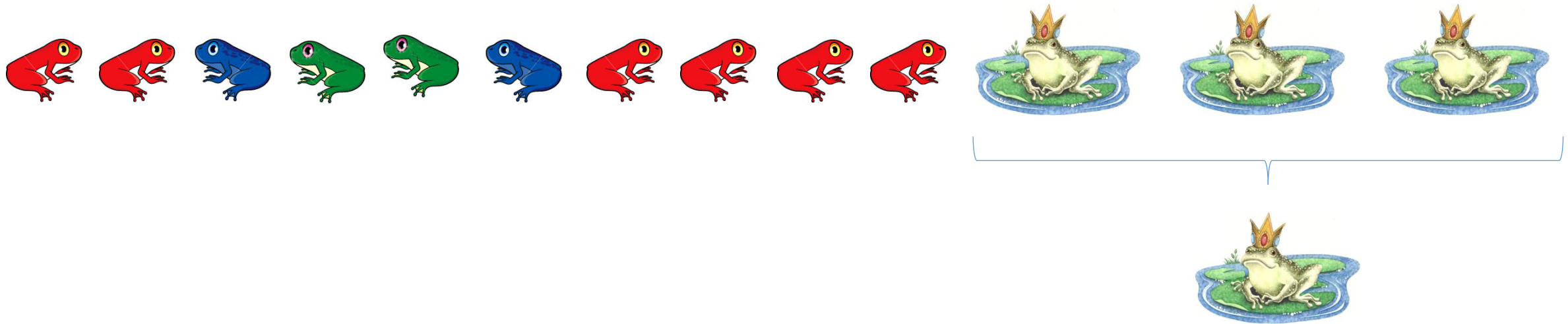
EXP (Y) = Geometrický průměr X

$$\bar{Y} = \sum_{i=1}^n \frac{Y_i}{n}$$

$\bar{Y} \pm \text{Standardní chyba}$

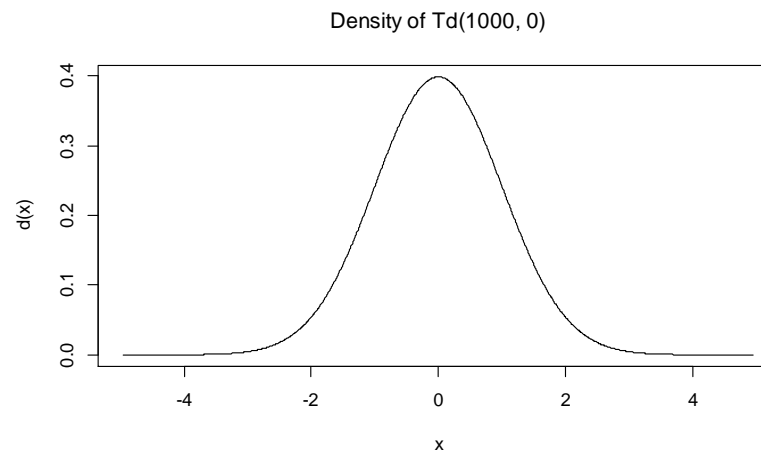
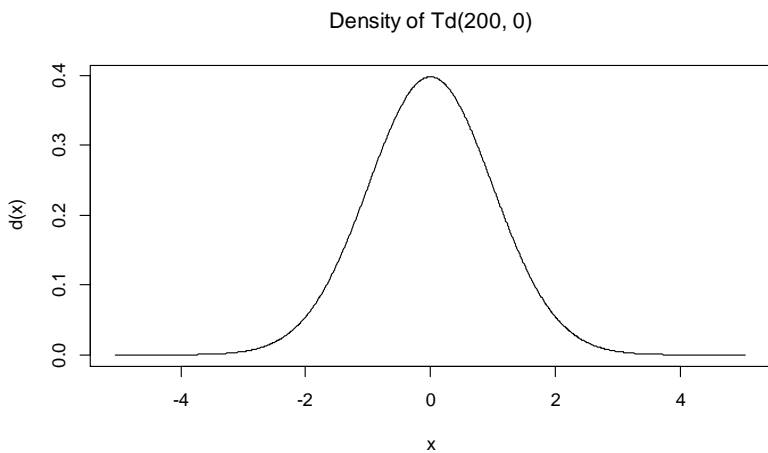
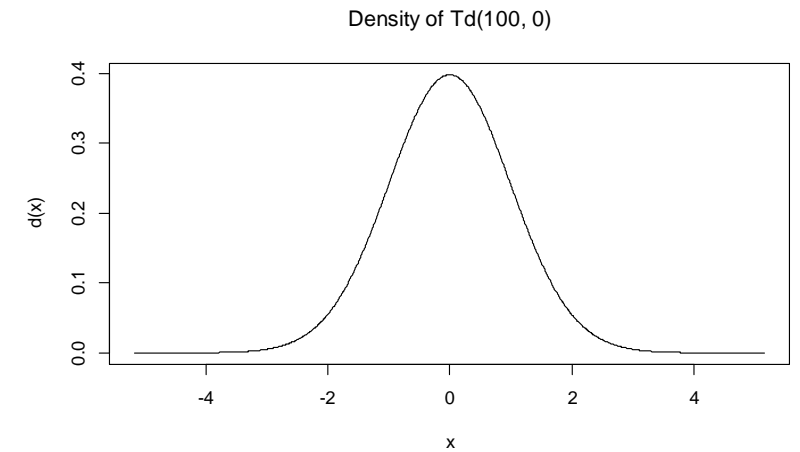
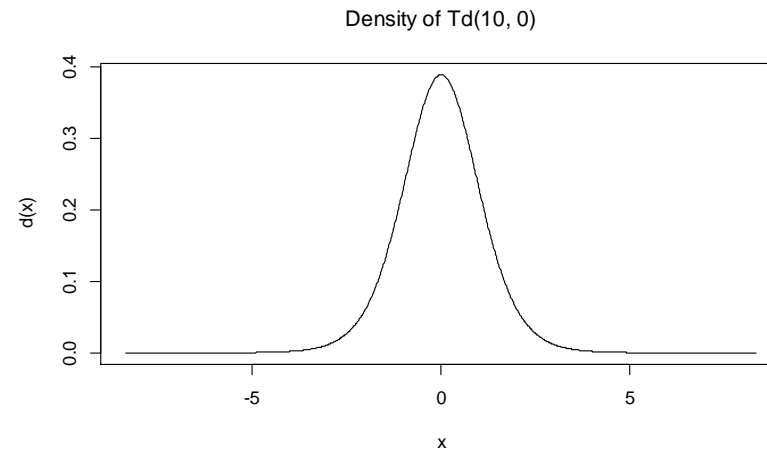
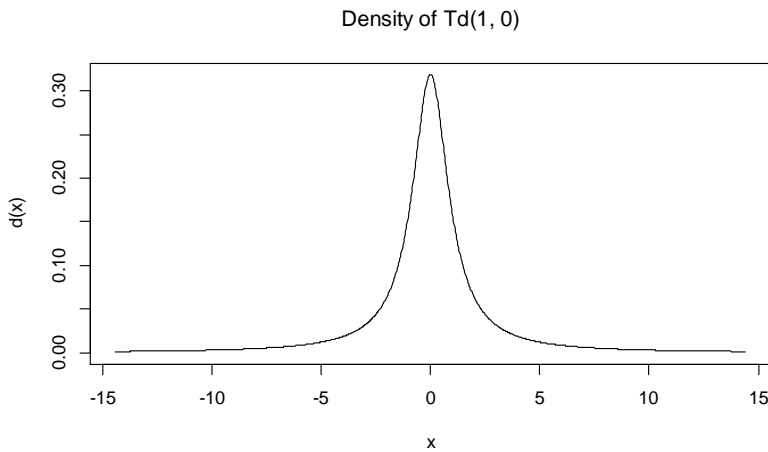
Stupně volnosti

- Nezávislé jednotky informace
- Spjaty s počtem objektů, popřípadě skupin v datech
- Klesají s výpočtem každé souhrnné statistiky (=odečítáme od celkového počtu vzniklé závislé statistiky)



Studentovo rozdělení

- Pro reálnější popis reality než umožňuje normální rozdělení
- Stupně volnosti – ve vazbě na velikost vzorku



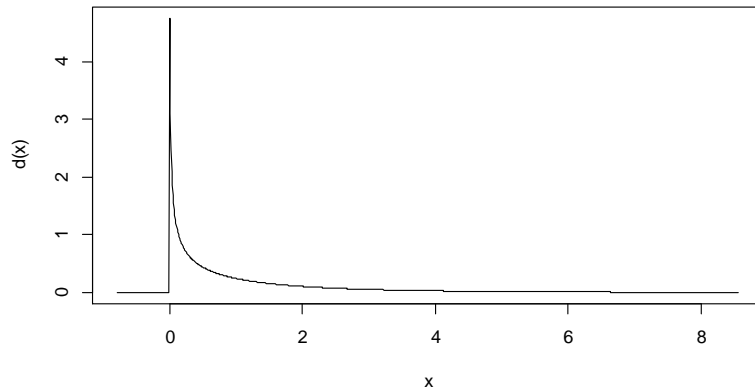
William Sealy Gosset

Publikace pod pseudonymem Student
t rozdělení na základě experimentů s kvasinkami

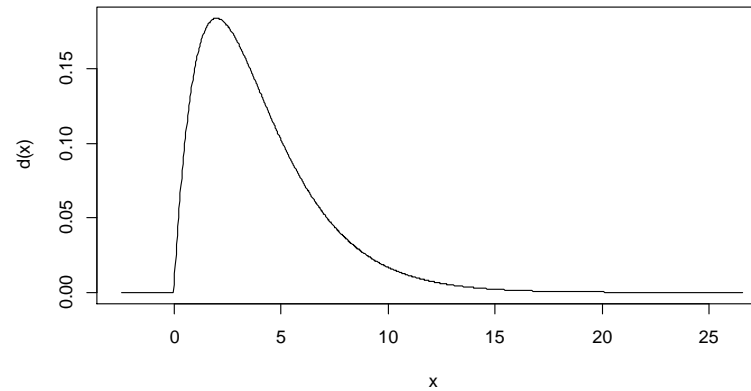
Pearsonovo (Chi-kvadrát) rozdělení

- Pro data, která nemohou být principiálně nikdy záporná
- Tvar ovlivněn stupni volnosti
- Očekávané a pozorované počty, rozptyly
- Často v genetice

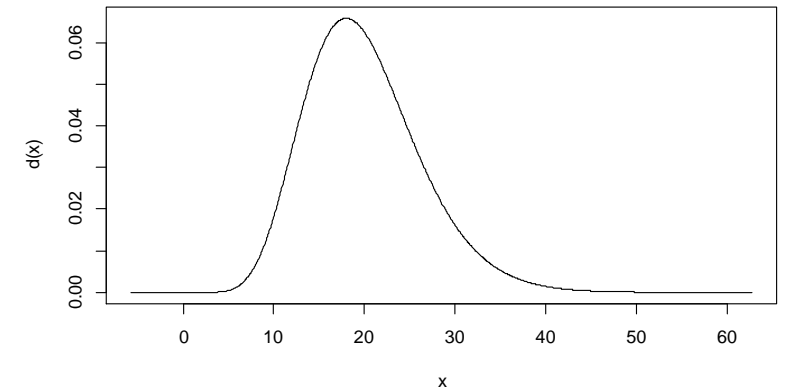
Density of Chisq(1, 0)



Density of Chisq(4, 0)



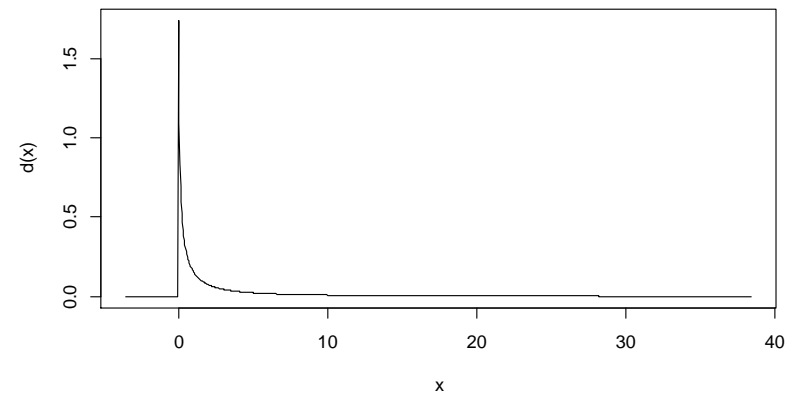
Density of Chisq(20, 0)



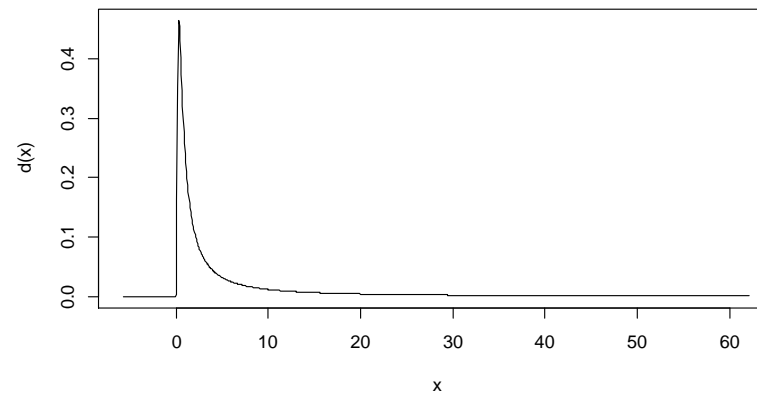
Fisher-Snedecorovo rozdělení

- Pro data, která nemohou být principiálně nikdy záporná
- Typicky poměr dvou rozptylů – využití v řadě, zejména pokročilejších statistických testů
- Dva různé stupně volnosti

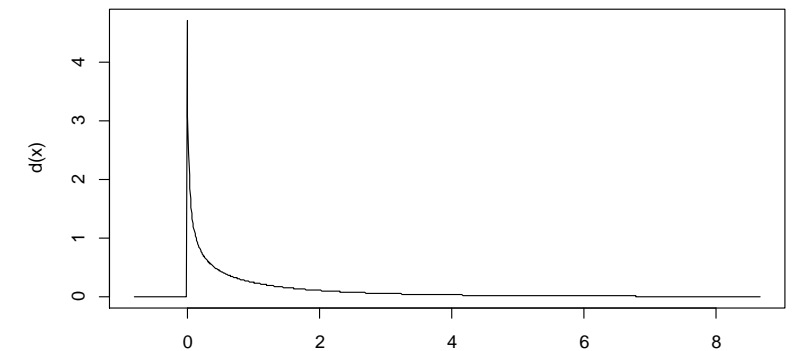
Density of Fd(1, 1, 0)



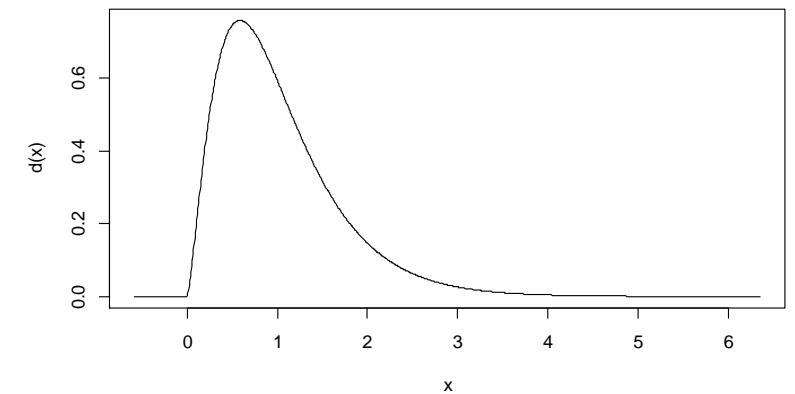
Density of Fd(100, 1, 0)



Density of Fd(1, 100, 0)



Density of Fd(5, 100, 0)



Transformace dat - legitimní úprava rozložení

- **Základní typy transformací vedou k normalitě rozložení nebo k homogenitě rozptylu**
- **Logaritmická transformace**
- Logaritmická transformace je velmi vhodná pro data s odlehlými hodnotami na horní hranici rozsahu. Při porovnání průměrů u více souborů dat je pro tuto transformaci indikující situace, kdy se s rostoucím průměrem mění proporcionálně i směrodatná odchylka, a tedy jednotlivé proměnné mají stejný koeficient variance, ačkoli mají různý průměr.
- Za takovéto situace přináší logaritmická transformace nejen zeslabení asymetrie původního rozložení, ale také vyšší homogenitu rozptylu proměnných. Pro transformaci se nejčastěji používá přirozený logaritmus a pokud jsou v původním souboru dat nulové hodnoty, je vhodné použít operaci $Y = \ln(X+1)$.
- Je-li průměr logaritmovaných dat (tedy průměrný logaritmus) zpětně transformován do původních hodnot, výsledkem není aritmetický, ale geometrický průměr původních dat.

Transformace dat - legitimní úprava rozložení

- **Základní typy transformací vedou k normalitě rozložení nebo k homogenitě rozptylu**
- **Odmocninová transformace**
- Transformace je vhodná pro proměnné mající Poissonovo rozložení, tedy proměnné vyjadřující celkový počet nastání určitého jevu (spíše vzácného) v n nezávisle opakovaných pokusech. Obecněji lze tento typ transformace doporučit v případě normalizace dat typu počtu jedinců (buněk, apod.). Jde o transformaci:
 - $Y = \sqrt{x}$ nebo $Y = \sqrt{x+1}$ nebo $Y = \sqrt{x} + \sqrt{x+1}$
- Transformace s přičtenou hodnotou 1 jsou efektivní, pokud X nabývá velmi malých nebo nulových hodnot. Situace indikující vhodnost odmocninové transformace je také proporcionalita výběrového rozptylu a průměru, tedy obecně jestliže $s^2x = k$ (výběrový průměr).

Transformace dat - legitimní úprava rozložení

- **Základní typy transformací vedou k normalitě rozložení nebo k homogenitě rozptylu**
- **Arcsin transformace**
- Tzv. úhlová transformace - velmi vhodná pro data typu podílů výskytu určitého jevu (znaku) mezi n hodnocenými jedinci - tedy pro data mající binomické rozložení. Pokud se určitý znak vyskytuje r -krát mezi n možnostmi (jedinci, opakováními), pak lze vyjádřit relativní četnost jeho výskytu jako $p = r/n$ s variabilitou $p \cdot (1-p)/n$. Arcsin transformace odstraní ze souborů dat podíly blízké 0 nebo 1, a tak efektivně sníží variabilitu odhadů středu. Transformace však není schopná odstranit variabilitu vyvolanou rozdílným počtem opakování v jednotlivých variantách - v takovém případě lze doporučit provedení vážených transformací dat. Velmi častou formou této transformace je:

$$Y = \arcsin \sqrt{p}$$

- - tedy transformace podílů do hodnot, jejichž sinus je roven druhé odmocnině původních hodnot. Pokud celkový počet jedinců (opakování), mezi kterými je výskyt znaku monitorován, je $n < 50$, pak lze doporučit velmi efektivní empirická opatření pro transformaci podílů blízkých 0 nebo 1. Pro tento případ lze nahrazovat nulové podíly hodnotou $1/4n$ a 100 % podíly hodnotou $(n-1/4)/n$. Pokud se mezi hodnotami vyskytuje větší množství krajních hodnot (menší než 0,2 a větší než 0,8), lze doporučit transformaci:

$$Y = \frac{1}{2} \left[\arcsin \sqrt{\frac{x}{n+1}} + \arcsin \sqrt{\frac{x+1}{n+1}} \right]$$

Přednáška 5

Provádění odhadů

Bodové a intervalové odhady

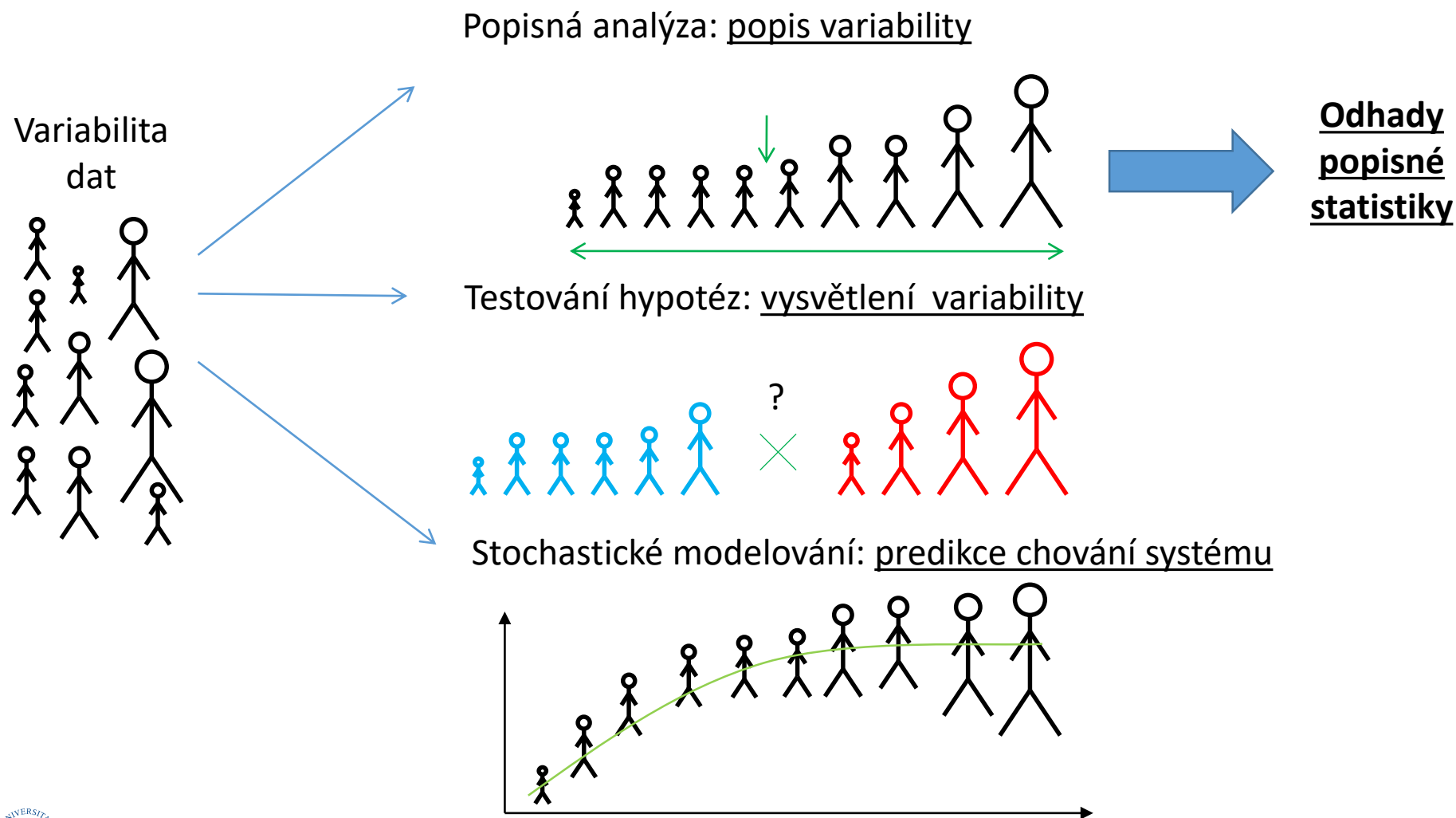
Význam intervalu spolehlivosti

Anotace

- Dva základní přístupy statistického hodnocení jsou popis dat a testování hypotéz.
- Při popisu dat je třeba si uvědomit, že popisné statistiky získané ze vzorku nejsou skutečnou hodnotou v cílové populaci, ale pouze jejím odhadem.
- Přesnost odhadu závisí jednak na variabilitě dat, jednak na velikosti vzorku, při vzorkování celé cílové populace by výsledná popisná statistika již byla přesnou hodnotou, nikoliv odhadem.
- Odhady a s nimi související intervaly spolehlivosti jsou univerzálním statistickým postupem a je možné je dopočítat k libovolné popisné statistice.

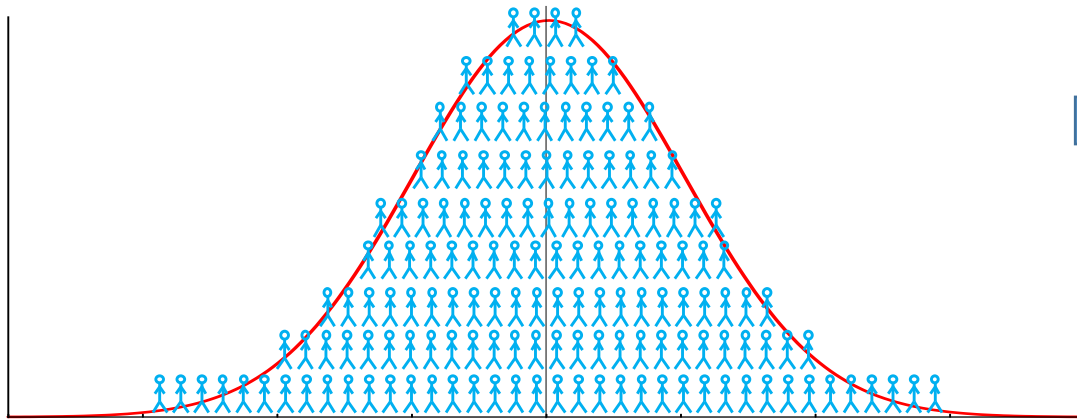
Práce s variabilitou v analýze dat

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



Bodový odhad popisné statistiky

- Výpočtem popisné statistiky vzorku získáme tzv. bodový odhad



Bodový odhad průměru,
směrodatné odchylky



Je to dostatečné?



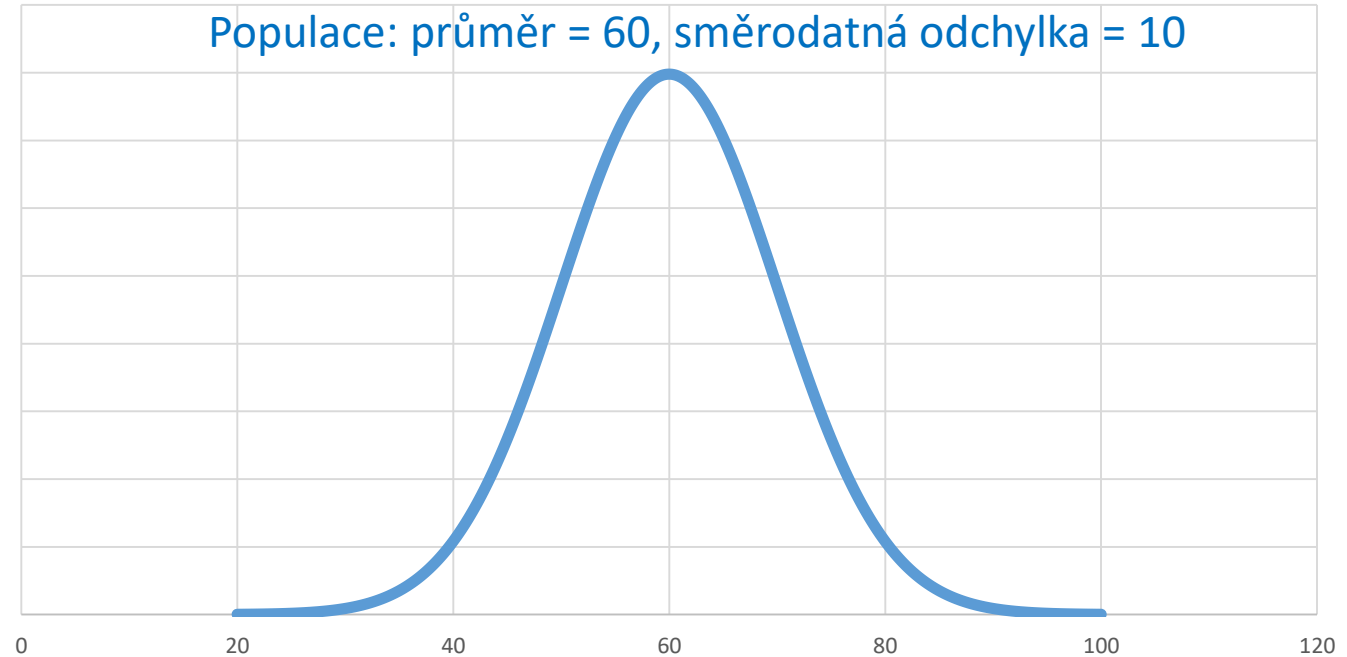
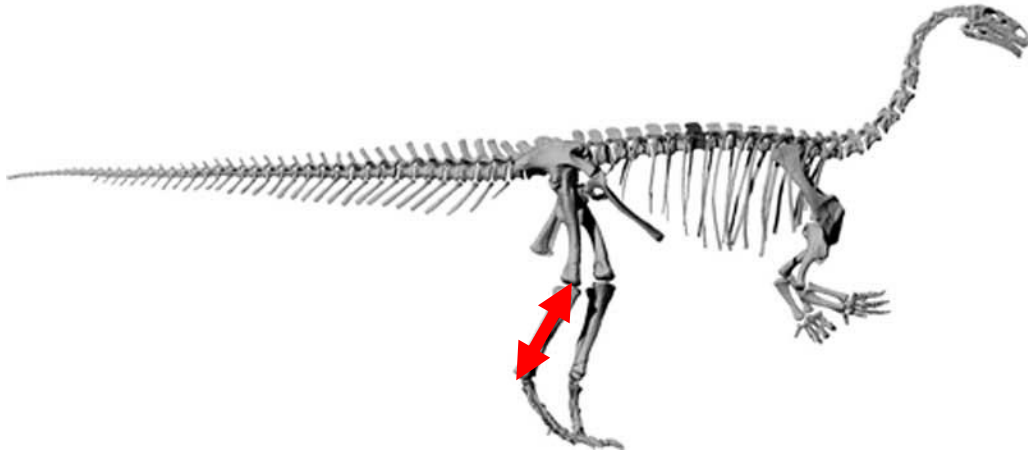
Není, nezohledňujeme vliv náhody,
která se uplatnila při vzorkování !!!

Intervalový odhad

- Bodový odhad je prvním krokem ve statistickém popisu dat.
- Co nám říká jedno číslo? Studie 1 může publikovat číslo x_1 , studie 2 číslo x_2 . Které je správnější, lepší, přesnější?
- **Bodový odhad je sám o sobě nedostatečný pro popis parametru rozdělení pravděpodobnosti náhodné veličiny.**
- Zajímá nás přesnost (spolehlivost) bodového odhadu.

Jaký je význam intervalového odhadu a jeho spolehlivosti?

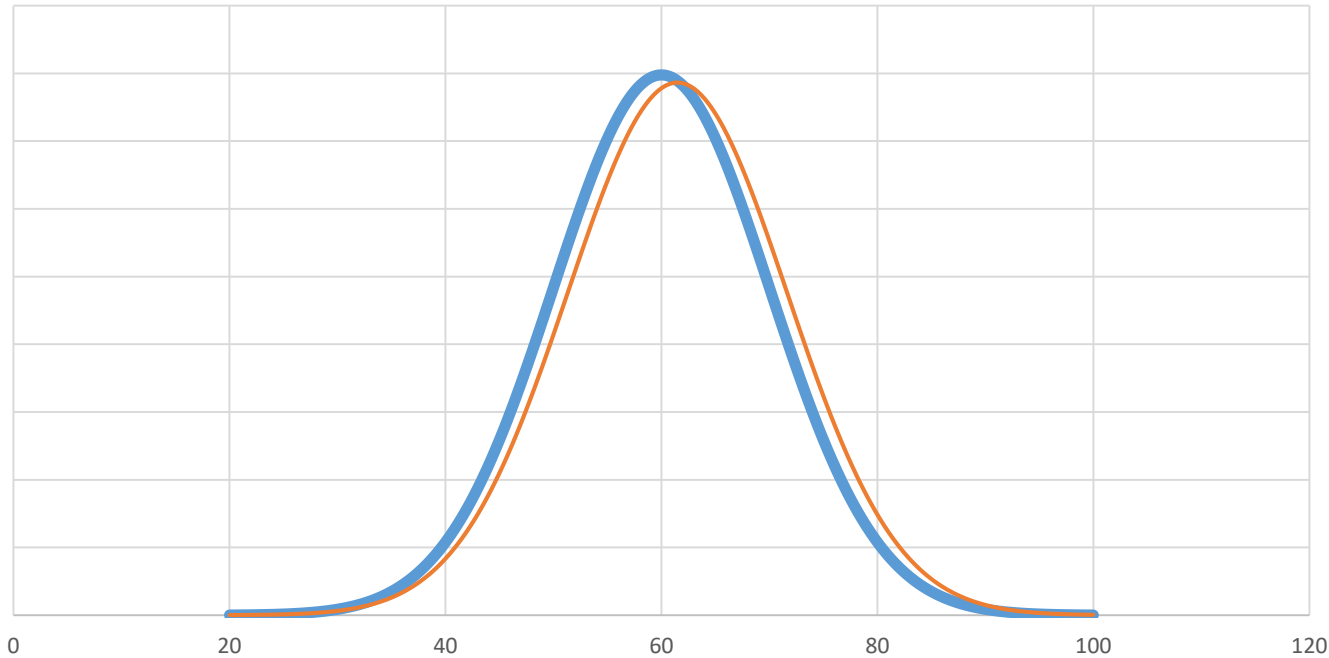
- Provádíme vzorkování populace živočichů a chceme odhadnout průměrnou hodnotu sledované proměnné
- Průměrná délka v populaci = 60, směrodatná odchylka = 10 (tyto hodnoty ve skutečnosti neznáme)



Provedeme vzorkování o velikosti $N = 100$.

Jedno vzorkování

- Je pouze nízká pravděpodobnost, že vzorek zcela přesně odpovídá sledované populaci



Populace: průměr = 60, směrodatná odchylka = 10

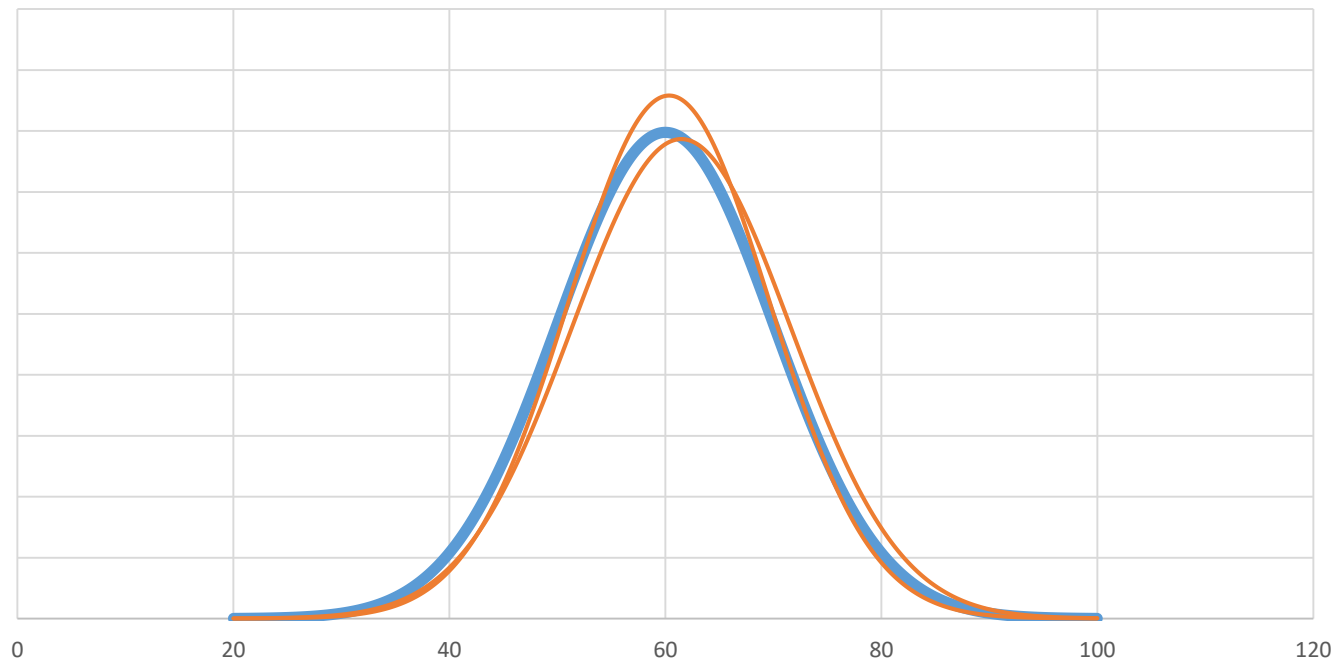
Vzorek 1: průměr = 61.5, směrodatná odchylka = 10.1



Jak by dopadlo další
vzorkování?

Dvě vzorkování

- Je pouze nízká pravděpodobnost, že vzorek zcela přesně odpovídá sledované populaci



Populace: průměr = 60, směrodatná odchylka = 10

Vzorek 1: průměr = 61.5, směrodatná odchylka = 10.1

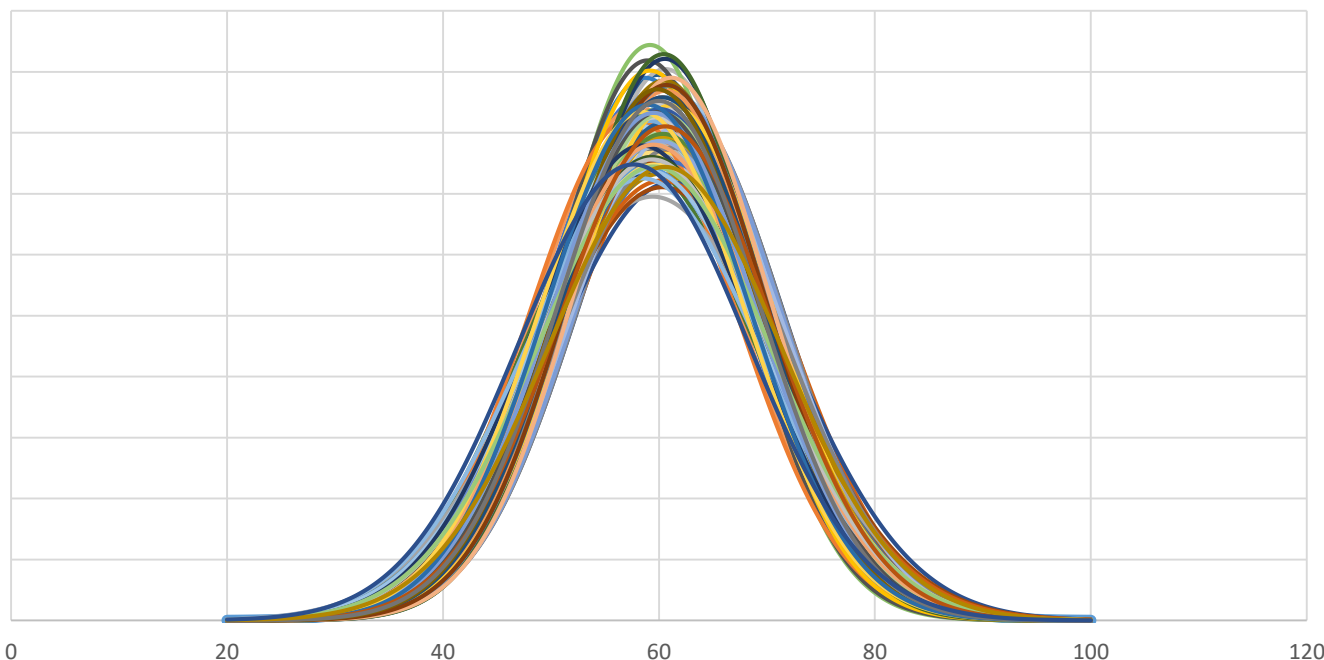
Vzorek 2: průměr = 60.4, směrodatná odchylka = 9.3



Jak by dopadlo další
vzorkování?

Sto vzorkování

- Je pouze nízká pravděpodobnost, že vzorek zcela přesně odpovídá sledované populaci



Populace: průměr = 60, směrodatná odchylka = 10

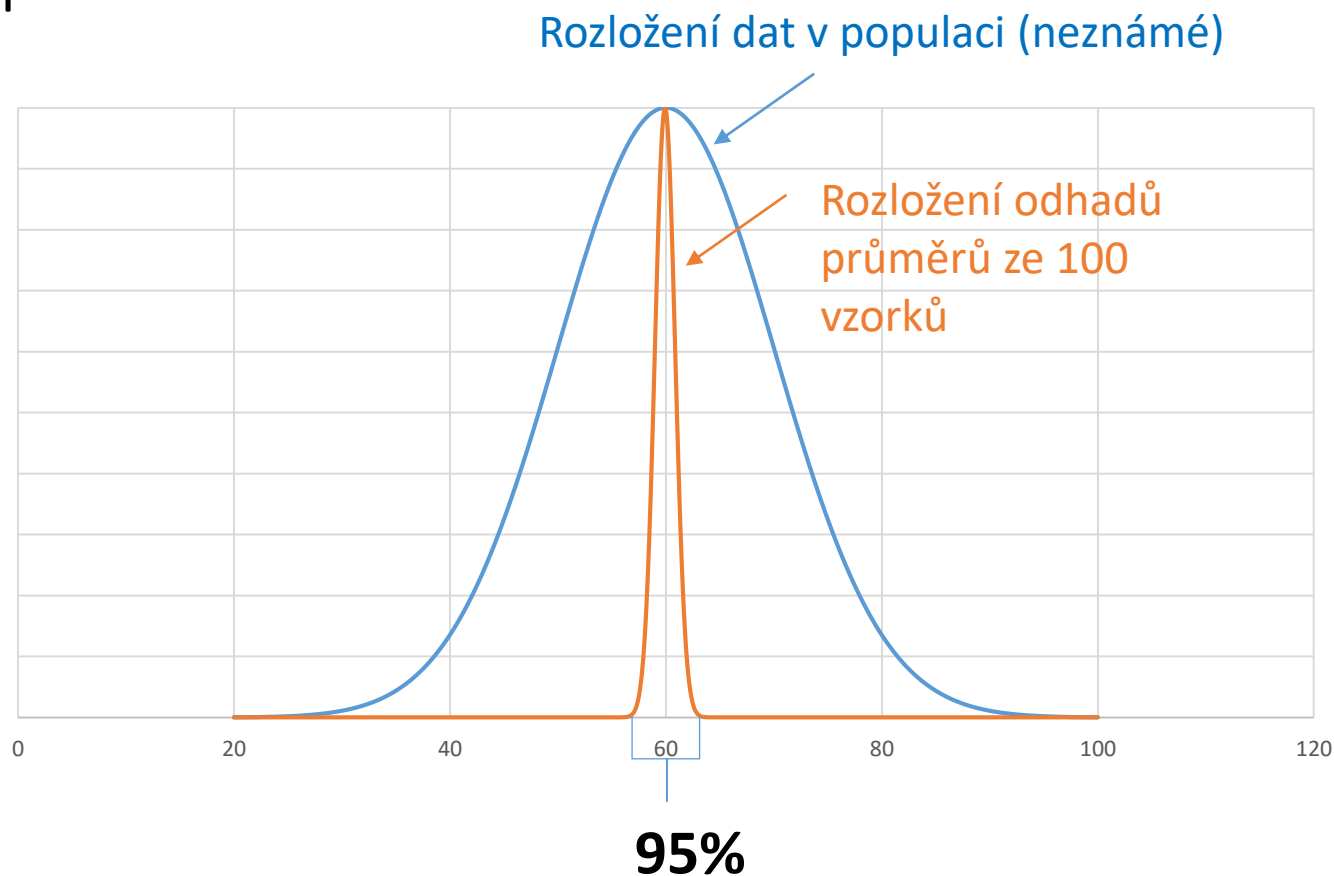
Opakovaným vzorkováním jsme získali různé varianty bodového odhadu simulující jak by při dané velikosti vzorku dopadlo různé vzorkování populace.



Jak by dopadlo další vzorkování?
Jsme schopni jej popsat z pohledu
pravděpodobnosti = **odhad při dalším vzorkování**
skončí s určitou pravděpodobností v určitém
rozsahu hodnot?

Interval spolehlivosti odhadu I

- Odhady průměru z jednotlivých vzorků vytváří rozložení odhadu průměrů
- Pokud známe rozložení jsme snadno určit rozsah, v němž leží zadané procento hodnot = pravděpodobnost s níž při vzorkování narazíme na odhad průměru v tomto rozmezí
- Nejběžněji se používá 95% rozsah = **95% interval spolehlivosti**
- Jak jej spočítat?



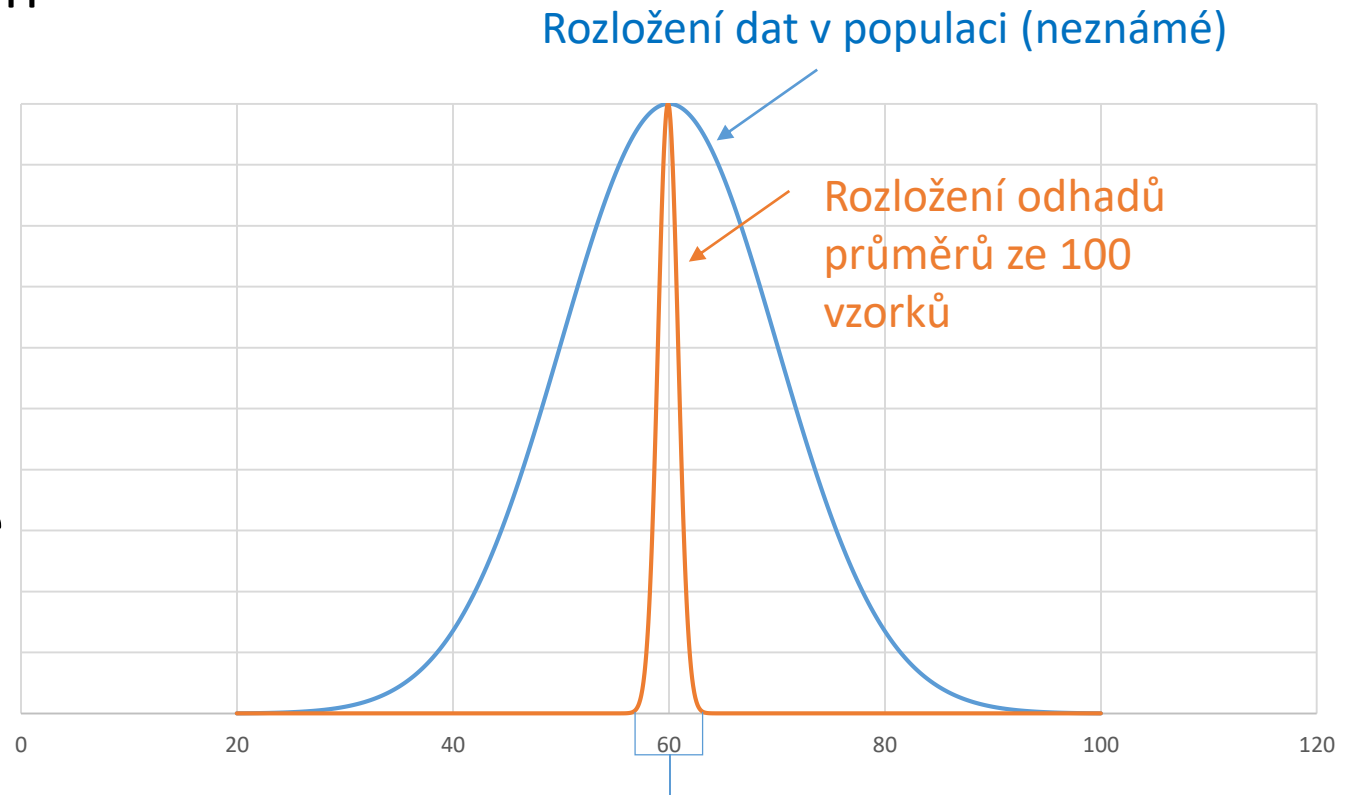
Populace: průměr = 60, směrodatná odchylka = 10

Vzorky (N = 100): průměr = 59.9, směrodatná odchylka odhadů průměru = 0.93

???

Interval spolehlivosti odhadu II

- Jak jej spočítat?
- Empiricky: 2,5% a 97,5% kvantil
- Dle modelového rozdělení:
 - Odhady průměrů mají normální rozdělení
 - Středních 95% hodnot ohraničuje průměr $\pm 1,96 \cdot \text{směrodatná odchylka}$
- Poznámka: popsáný způsob výpočtu intervalu spolehlivosti se používá pouze v počítačových simulacích, ne při reálném vzorkování (zde z výukových důvodů)



95%

Populace: průměr = 60, směrodatná odchylka = 10

Vzorky (N = 100): průměr = 59.9, směrodatná odchylka odhadů průměru = 0.93

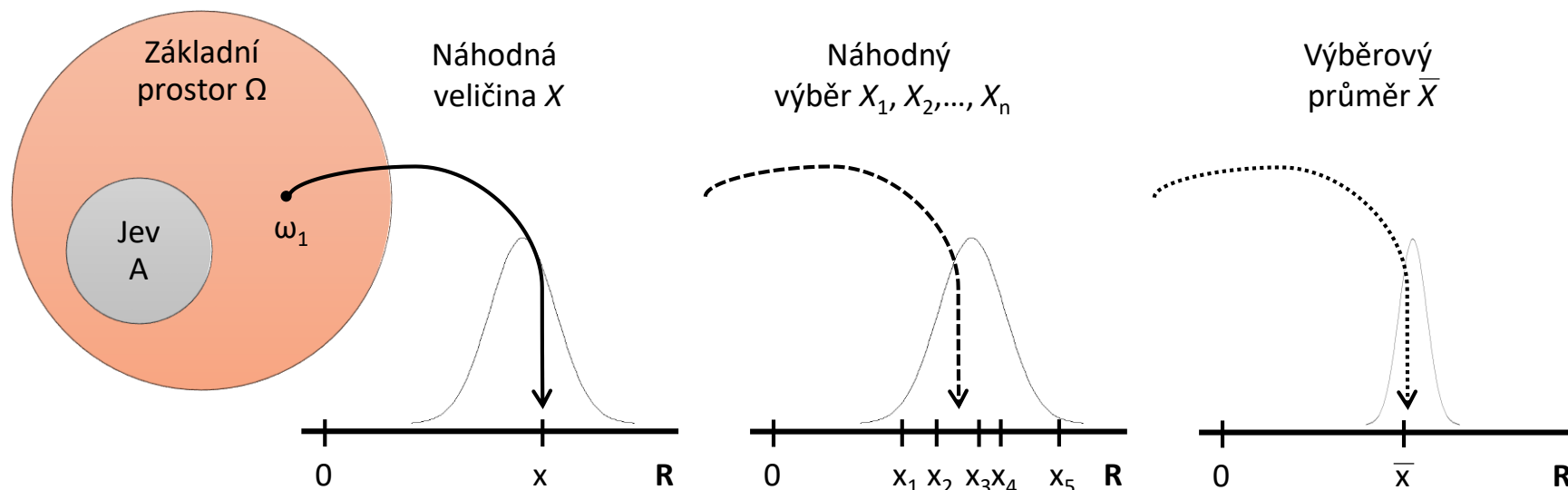
Střední chyba odhadu průměru (standard error, s.e., SE, $s_{\bar{x}}$)

Pravděpodobnostní chování náhodné veličiny

- V klasických statistických výpočtech je interval spolehlivosti odvozen z jednoho vzorku na základě znalosti modelového rozdělení odhadů dané statistiky (např. průměru)
- Dvě charakteristiky odráží vlastnosti rozdělení jedním číslem: střední hodnota a rozptyl. Odmocnina z rozptylu je směrodatná odchylka (SD).
- Platí následující:
 - Jednotlivé realizace náhodné veličiny vykazují variabilitu (dle SD).
 - Jakákoliv statistika (např. průměr) je jako transformace náhodných veličin také náhodnou veličinou. Má tedy i rozdělení pravděpodobnosti.
 - Jednotlivé realizace statistiky nad různými náhodnými výběry také vykazují variabilitu (opět úměrnou SD).
 - S.E. – standard error – střední chyba odhadu

Příklad – výběrový průměr

- V případě průměru jsou jeho odhady popsateľné modelem normálního rozdělení
- Normální rozdělení je popsáno průměrem (vlastní odhad průměru) a směrodatnou odchylkou odhadů (pro odlišení od směrodatné odchylky vzorku se v tomto případě nazývá střední chyba odhadu průměru)



SD a SE

- Směrodatná odchylka (SD) není směrodatná chyba popisné statistiky (SE)!
- Směrodatná odchylka (SD) je odrazem variability náhodné veličiny ve sledované populaci.
- Směrodatná chyba (SE) je odrazem přesnosti popisné statistiky jako odhadu střední hodnoty náhodné veličiny.
- Pozor na rozdíl mezi SD a SE v člancích a knihách – tabulkách a grafech!
- **Na čem závisí velikost SE (a tedy i šířka intervalu spolehlivosti?)**

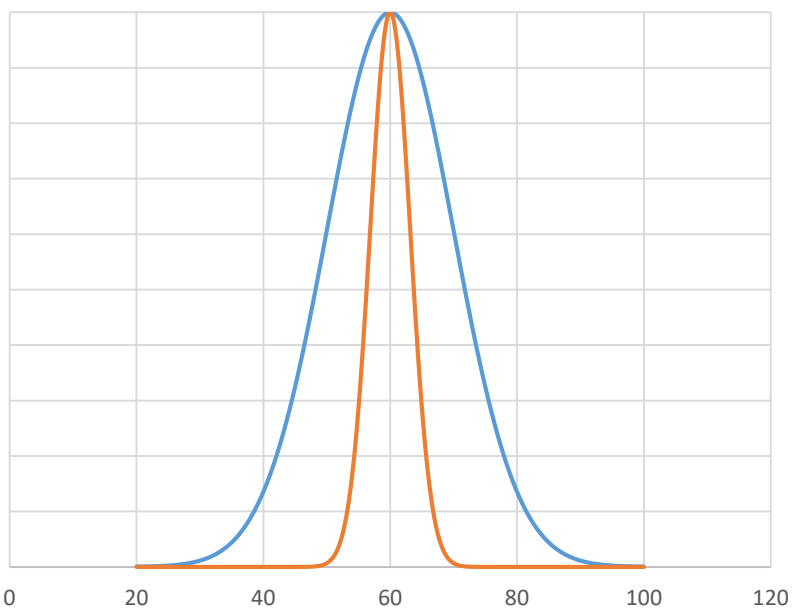
SD a SE

- Směrodatná odchylka (SD) není směrodatná chyba popisné statistiky (SE)!
- Směrodatná odchylka (SD) je odrazem variability náhodné veličiny ve sledované populaci.
- Směrodatná chyba (SE) je odrazem přesnosti popisné statistiky jako odhadu střední hodnoty náhodné veličiny.
- Pozor na rozdíl mezi SD a SE v člancích a knihách – tabulkách a grafech!
- **Na čem závisí velikost SE (a tedy i šířka intervalu spolehlivosti?)**
 - Na velikosti vzorku
 - Variabilitě (směrodatné odchylce) hodnocené proměnné v populaci
- SD populace je daná realitou, ale velikost vzorku je v našich rukou = **změnou velikosti vzorku můžeme měnit šíři intervalu spolehlivosti !!!!**

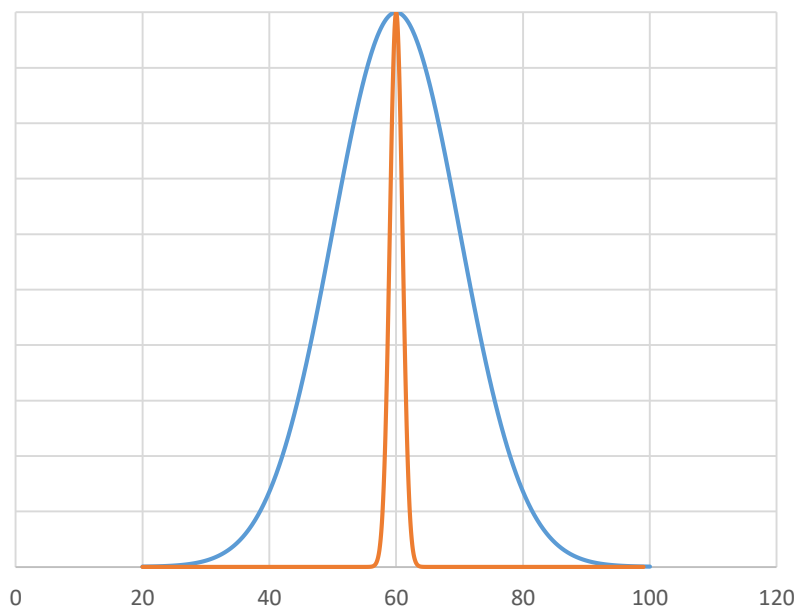
Příklad – interval spolehlivosti při různých velikostech vzorku

- Provádíme vzorkování populace živočichů a chceme odhadnout průměrnou hodnotu sledované proměnné – zkoušíme různé velikosti vzorku
- Průměrná délka v populaci = 60, směrodatná odchylka = 10 (tyto hodnoty ve skutečnosti neznáme)

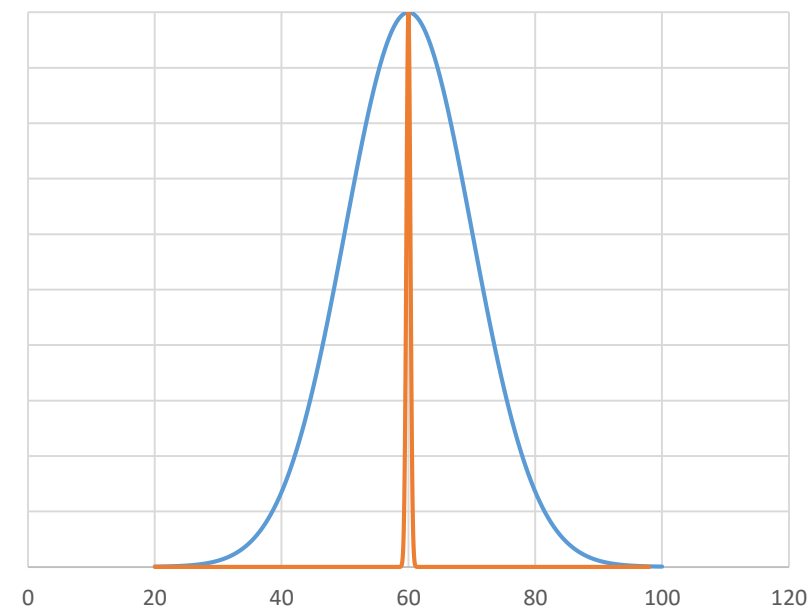
N = 10



N = 100



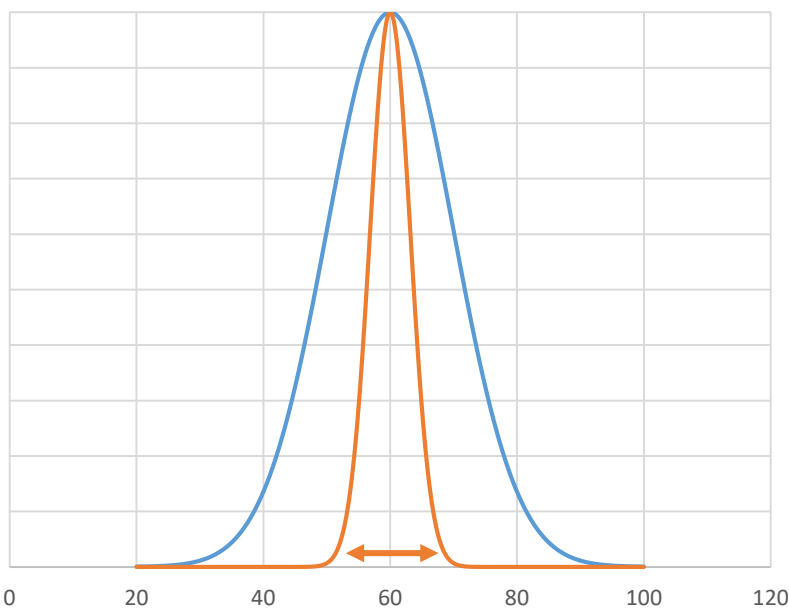
N = 1000



Příklad – interval spolehlivosti při různých velikostech vzorku

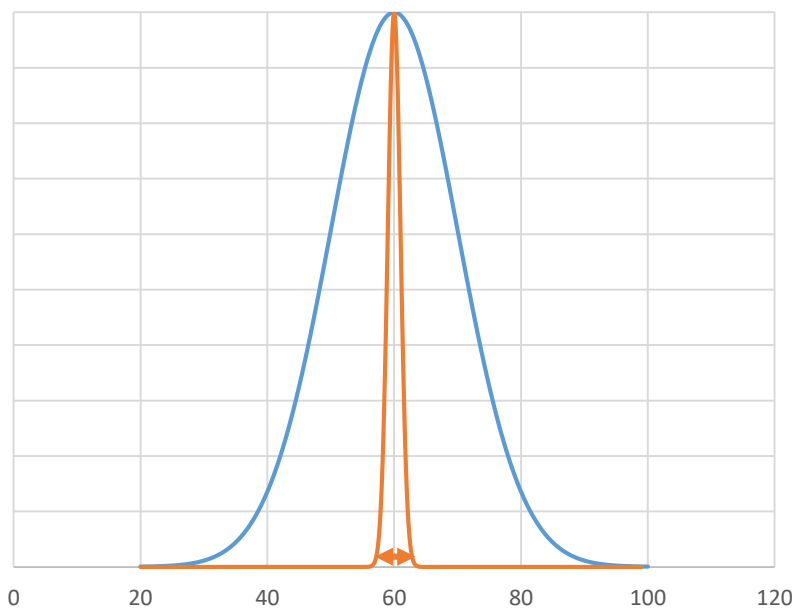
- Provádíme vzorkování populace živočichů a chceme odhadnout průměrnou hodnotu sledované proměnné – zkoušíme různé velikosti vzorku
- Průměrná délka v populaci = 60, směrodatná odchylka = 10 (tyto hodnoty ve skutečnosti neznáme)

N = 10



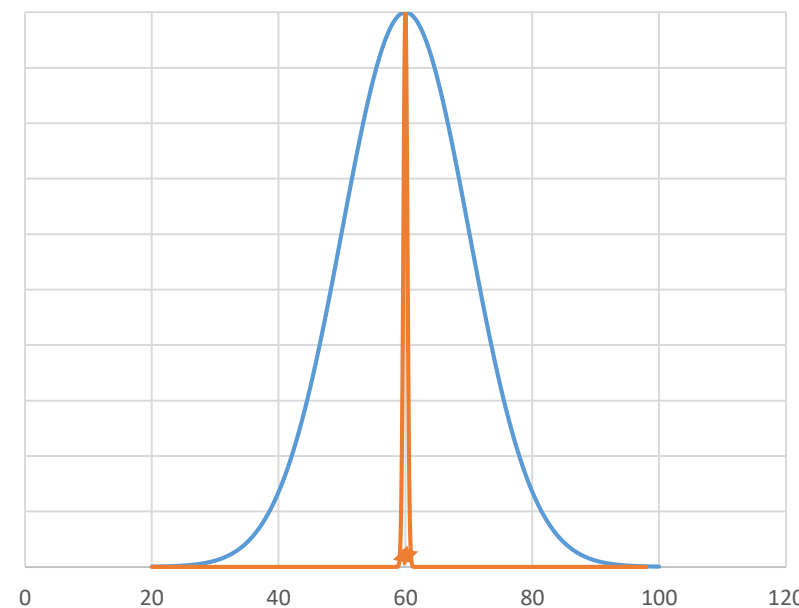
95% IS = 53,8 – 66,2

N = 100



95% IS = 58,0 – 62,0

N = 1000



95% IS = 59,4 – 60,6

Obecný vzorec výpočtu intervalu spolehlivosti

- Interval spolehlivosti lze spočítat pro odhad jakékoliv popisné statistiky (průměr, směrodatná odchylka, procento, korelační koeficient, regresní koeficient, odds ratio atd.)
- Pro danou popisnou statistiku musíme znát odpovídající modelové rozdělení jejího odhadu
- Obecná rovnice pro výpočet hranic intervalu spolehlivosti (v některých případech může být složitější – asymetrické intervaly spolehlivosti, různá rovnice pro dolní a horní hranici):

Bodový odhad ± kvantil modelového rozdělení * střední chyba odhadu

↑
Např. průměr vzorku

↑
V případě průměru a 95% intervalu spolehlivosti to je 2.5% a 97.5% kvantil normálního rozdělení = ± 1.96

↑
V případě průměru je vypočtena jako:

$$s_{\bar{x}} = \frac{s}{\sqrt{N}}$$

Výpočet odhadu průměru

- Bodový odhad průměru daného vzorku
- Střední chyba odhadu průměru

\bar{x}

$$s_{\bar{x}} = \frac{s}{\sqrt{N}}$$

- Interval spolehlivosti

$$\bar{x} - t_{1-\alpha/2}^{v=N-1} \frac{s}{\sqrt{N}} \leq \mu \leq \bar{x} + t_{1-\alpha/2}^{v=N-1} \frac{s}{\sqrt{N}}$$

$$\mu: \bar{x} \pm t_{1-\alpha/2}^{v=N-1} \frac{s}{\sqrt{N}}$$

$$\mu: \bar{x} \pm t_{1-\alpha/2}^{v=N-1} s_{\bar{x}}$$

t – Studentovo rozdělení (používáno namísto normálního při malé velikosti vzorku)

v – stupně volnosti, zde počítány jako N-1

Co je ? $t_{1-\alpha/2}^{v=N-1}$

Kvantil modelového rozdělení, α znamená zastoupení případů, které do intervalu nechceme zahrnout, zde pro 95% interval spolehlivosti je $\alpha = 5\%$, hledáme tedy 97.5% kvantil studentova rozdělení

Statistické tabulky t-rozdělení

- Na rozdíl od tabulek normálního rozdělení musíme zohlednit i stupně volnosti
- Z tohoto důvodu je tabulka konstruována jen pro vybrané hodnoty pravděpodobnosti

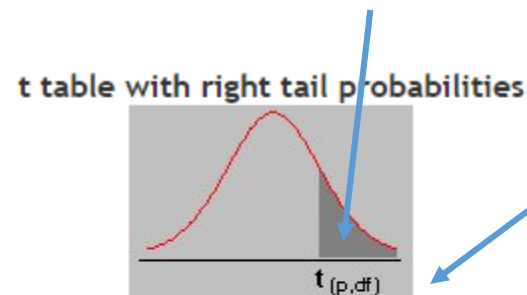


William Sealy Gosset

Publikace pod pseudonymem Student

t rozdělení na základě experimentů s kvasinkami

Hledáme hodnotu **t** (= kvantil rozdělení) pro danou plochu **(pravděpodobnost)** a **stupně volnosti**



Pravděpodobnost (plocha pod křivkou), nejběžněji 0.025 ($2 \cdot 0.025 = 0.05$)

df \ p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869

Stupně volnosti

Odhad průměru a jeho intervalu spolehlivosti – příklad 1

- Provádíme vzorkování populace živočichů a chceme odhadnout průměrnou hodnotu sledované proměnné
- Vzorek: N = 10, průměr (bodový odhad) 61,5, směrodatná odchylka 10,1

- **Jaký je 95% interval spolehlivosti?**

- Střední chyba odhadu $s_{\bar{x}} = \frac{s}{\sqrt{N}} = \frac{10,1}{\sqrt{10}} = 3,207$

- Kvantil modelového rozdělení pro $\alpha=0,05$ (1-0,95)

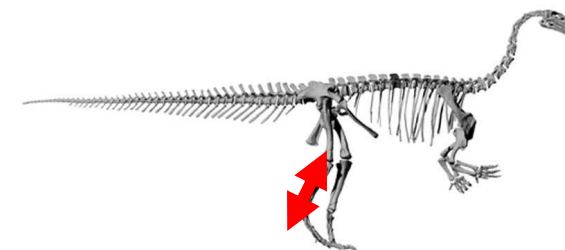
$$t_{1-\alpha/2}^{v=N-1} = t_{1-0,05/2}^{v=10-1} = t_{0,975}^9 = 2,262$$

- 95% interval spolehlivosti – výpočet

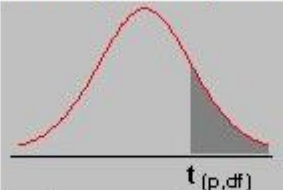
$$\mu: \bar{x} \pm t_{1-\alpha/2}^{v=N-1} \frac{s}{\sqrt{N}} = 61,5 \pm 2,262 * 3,207 = 61,5 \pm 7,256$$

- 95% interval spolehlivosti - výsledek

61,5 (54,2 – 68,7)



t table with right tail probabilities



df \ p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809

- **Při opakovaném vzorkování o N=10 bude odhad průměru s pravděpodobností 0,95 ležet v rozsahu (54,2 – 68,7)**

Odhad průměru a jeho intervalu spolehlivosti – příklad 2

- Provádíme vzorkování populace živočichů a chceme odhadnout průměrnou hodnotu sledované proměnné
- Vzorek: N = 100, průměr (bodový odhad) 61,5, směrodatná odchylka 10,1

- **Jaký je 95% interval spolehlivosti?**

- Střední chyba odhadu $s_{\bar{x}} = \frac{s}{\sqrt{N}} = \frac{10,1}{\sqrt{100}} = 1,014$

- Kvantil modelového rozdělení pro $\alpha=0,05$ (1-0,95)

$$t_{1-\alpha/2}^{v=N-1} = t_{1-0,05/2}^{v=100-1} = t_{0,975}^{99} = 1,960$$

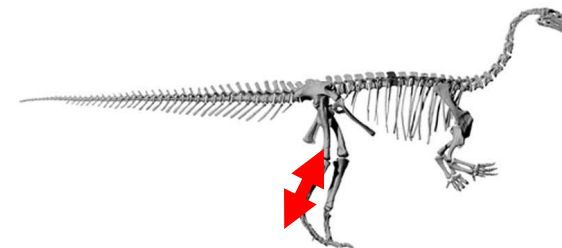
- 95% interval spolehlivosti – výpočet

$$\mu: \bar{x} \pm t_{1-\alpha/2}^{v=N-1} \frac{s}{\sqrt{N}} = 61,5 \pm 1,960 * 1,014 = 61,5 \pm 1,988$$

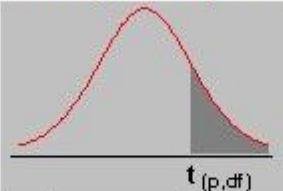
- 95% interval spolehlivosti - výsledek

61,5 (59,5 – 63,5)

- **Při opakovaném vzorkování o N=100 bude odhad průměru s pravděpodobností 0,95 ležet v rozsahu (59,5 – 63,5)**



t table with right tail probabilities



df\p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
inf	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905

Interval spolehlivosti pro odhad rozptylu

- Příklad asymetrického intervalu spolehlivosti; modelovým rozdělením je Pearsonovo (chi-kvadrát rozdělení)

- **Pro rozptyl**

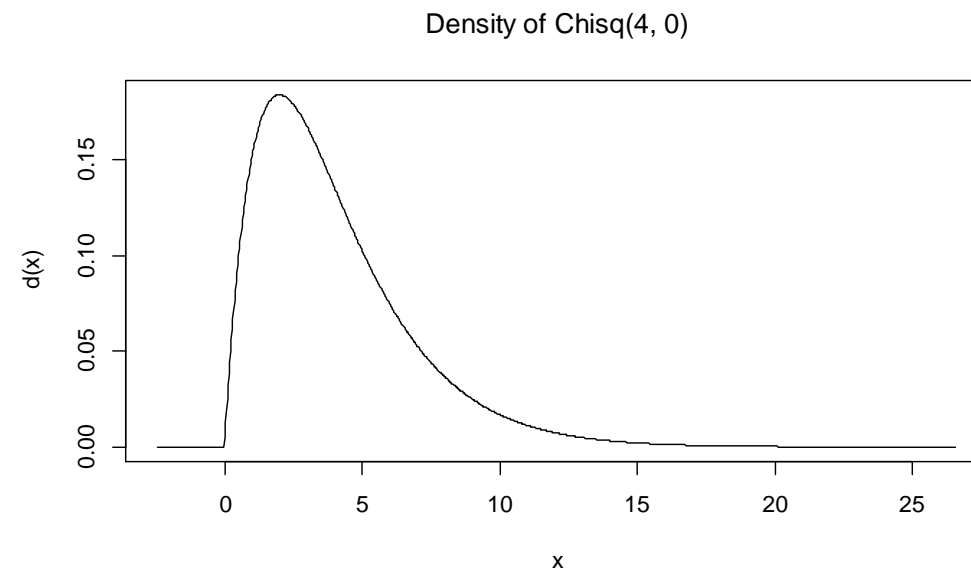
$$\frac{(N-1)s^2}{\chi^2_{\alpha/2, \nu=N-1}} \leq \sigma^2 \leq \frac{(N-1)s^2}{\chi^2_{1-\alpha/2, \nu=N-1}}$$

- **Pro směrodatnou odchylku**

$$\sqrt{\frac{(N-1)s^2}{\chi^2_{\alpha/2, \nu=N-1}}} \leq \sigma \leq \sqrt{\frac{(N-1)s^2}{\chi^2_{1-\alpha/2, \nu=N-1}}}$$

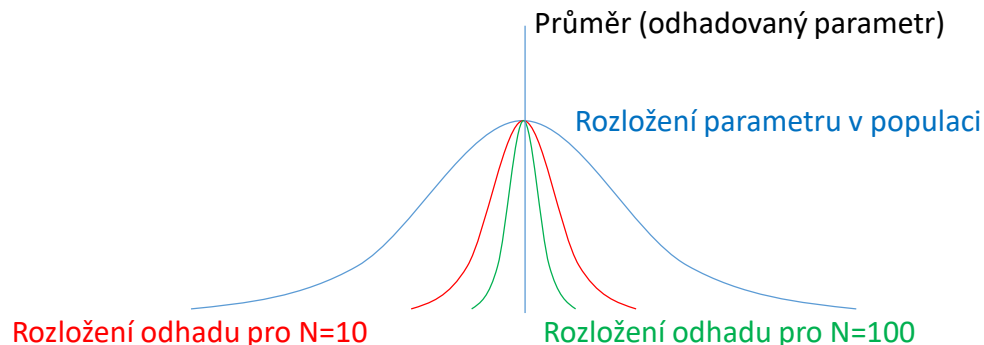
- **Pro střední chybu odhadu průměru**

$$\sqrt{\frac{(N-1)s^2}{N\chi^2_{\alpha/2, \nu=N-1}}} \leq \frac{\sigma}{\sqrt{N}} \leq \sqrt{\frac{(N-1)s^2}{N\chi^2_{1-\alpha/2, \nu=N-1}}}$$



Koncept intervalu spolehlivosti a jeho interpretace: shrnutí

- Při výpočtu odhadu popisné statistiky nás zajímá nejenom její vlastní hodnota (bodový odhad) ale také její rozsah spolehlivosti
- Interval spolehlivosti závisí na:
 - Velikosti vzorku
 - Variabilitě dat
 - Požadované spolehlivosti
- Interval spolehlivosti lze spočítat pro jakoukoliv statistiku (průměr, směrodatná odchylka, korelace, procentuální zastoupení apod.)
- Interval spolehlivosti poskytuje vodítko jak „spolehlivé“ jsou naše výsledky a s jakou pravděpodobností jich je možné opakovaně dosáhnout
- 95% interval spolehlivosti je rozsah hodnot do něž se při opakování studie trefíme s 95% pravděpodobností
- Tvzení, že v rozsahu 95% intervalu spolehlivosti leží s 95% pravděpodobností skutečný průměr populace není pravdivé, skutečný průměr populace neznáme !!!



Poznámka k intervalu spolehlivosti

- Interval spolehlivosti počítá pouze s variabilitou danou náhodným výběrem, nepočítá se zdroji systematického zkreslení.
- **Příklady:**
 - Měření koncentrace polutantu nebo krevního tlaku může být systematicky zkresleno starým měřidlem („technical bias“).
 - Měření koncentrace polutantu může být systematicky zkresleno výběrem pouze čistých nebo pouze kontaminovaných lokalit („selection bias“)
 - Měření krevního tlaku může být systematicky zkresleno tím, že se do studie přihlásí pouze určitá skupina osob („selection bias“)

Základy testování hypotéz

Princip statistického testování hypotéz

Testová statistika a statistická významnost

Chyby statistického testování

Anotace

- Testování hypotéz je po popisné statistice druhým hlavním směrem statistických analýz. Při testování pokládáme hypotézy, které se snažíme s určitou pravděpodobností potvrdit nebo vyvrátit.
- Tzv. nulovou hypotézu lze nejlépe popsat jako situaci, kdy předpokládáme vliv náhody (rozdíl mezi skupinami je pouhá náhoda, vztah dvou proměnných je pouhá náhoda apod.), alternativní hypotéza předpokládá vliv nenáhodného faktoru.
- Výsledkem statistického testu je v zásadě pravděpodobnost nakolik je hodnocený jev náhodný nebo ne, při překročení určité hranice (nejčastěji méně než 5% pravděpodobnost, že jev je pouhá náhoda) deklaruujeme, že pravděpodobnost náhody je pro nás dostatečně nízká abychom jev prohlásili za nenáhodný
- Statistická významnost je ovlivnitelná velikostí vzorku a tak je pouze indicií k prohlášení např. rozdílu dvou skupin pacientů za skutečně významný. V ideální situaci je nezbytné aby rozdíl byl významný nejenom statisticky (=nenáhodný), ale i prakticky (=nejde pouze o artefakt velikosti vzorku).

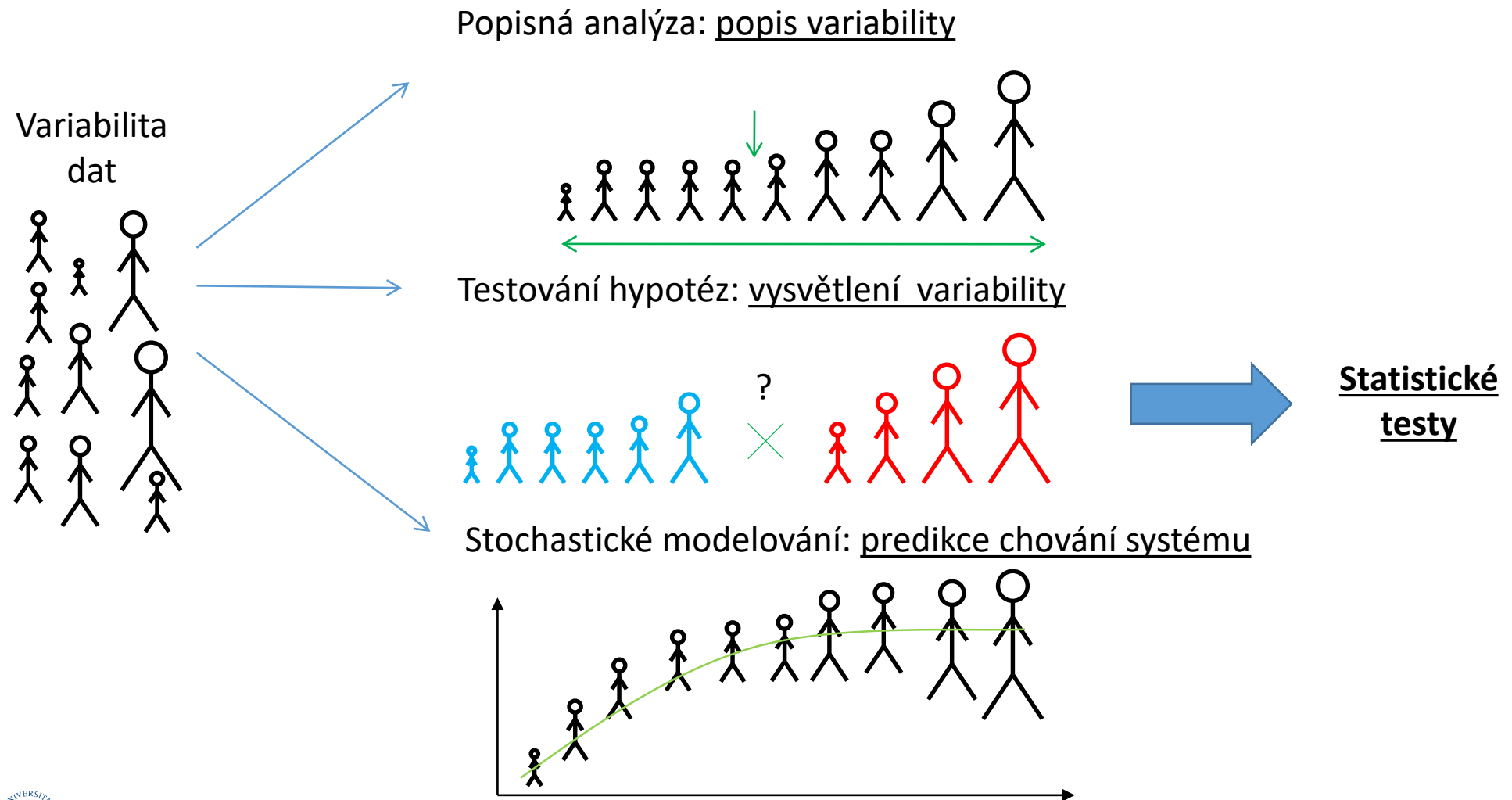
Statistické testování neznamená průkaz kauzality !!!!

- Výsledek statistického testování neznamená kauzální prokázání nebo neprokázání vztahu, jde pouze o indicii k našemu rozhodování.



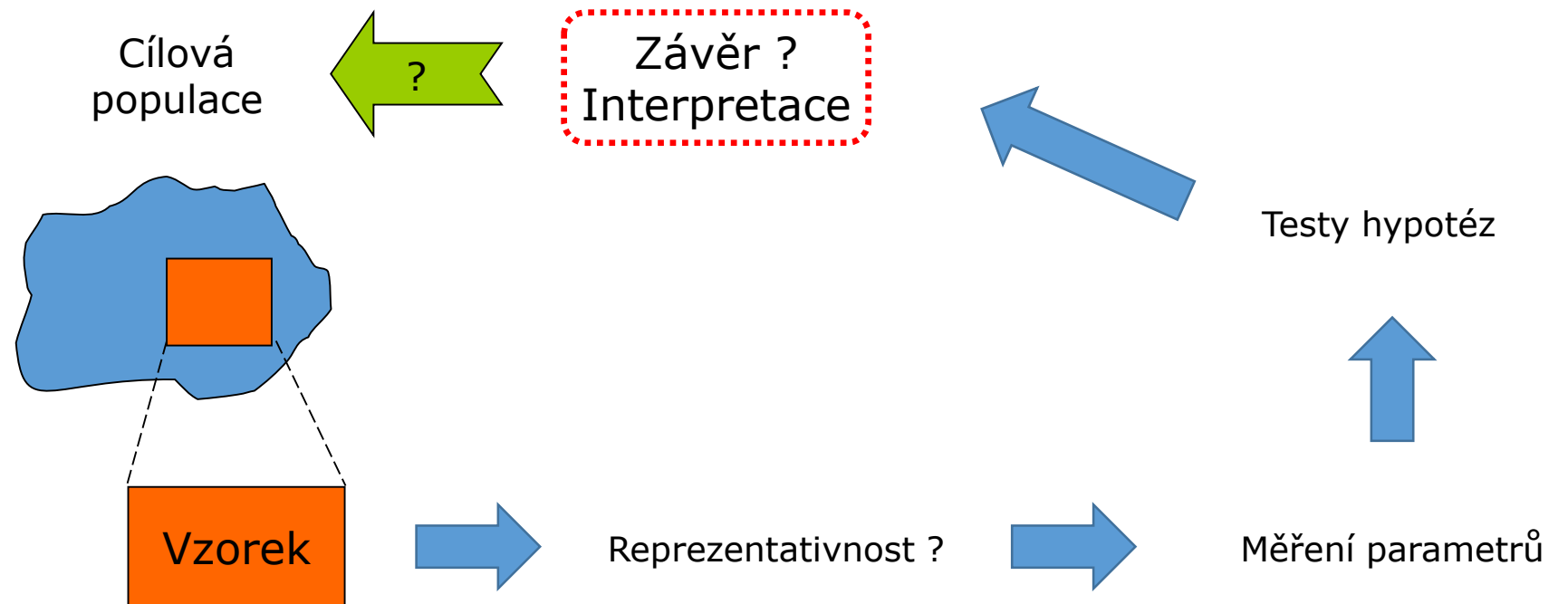
Práce s variabilitou v analýze dat

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



Princip testování hypotéz

- Formulace hypotézy
- Výběr cílové populace a z ní reprezentativního vzorku
- Měření sledovaných parametrů
- Použití odpovídajícího testu → závěr testu
- Interpretace výsledků



Stanovení hypotézy

- **Nulová hypotéza („null hypothesis“)** – tvrzení o neznámých vlastnostech rozdělení pravděpodobnosti sledované náhodné veličiny (znaku, vlastnosti) týkající se cílové populace.
- Nulová hypotéza má tvar: $H_0 : \theta = \theta_0$
- Nulová hypotéza obecně říká, že rozdíl není, popřípadě, že rozdíl je tak malý, že jej můžeme považovat za náhodný -> základní otázkou testování tak je „jak definovat co je pro nás „dostatečně“ náhodné?“
- **Alternativní hypotéza** – tvrzení o neznámých vlastnostech rozdělení pravděpodobnosti sledované náhodné veličiny, které popírá platnost nulové hypotézy. Vymezuje, jaká situace nastává, když nulová hypotéza neplatí.
- Alternativní hypotéza má tvar: $H_1 : \theta \neq \theta_0$
 $H_1 : \theta < \theta_0$
 $H_1 : \theta > \theta_0$

Příklady hypotézy

- Liší se lokality poblíž lidských sídel od lokalit v chráněných rezervacích co do míry znečištění?

Míra znečištění na lokalitách poblíž sídel: θ_1 $H_0 : \theta_1 = \theta_2$

Míra znečištění na lokalitách v rezervacích: θ_2 $H_1 : \theta_1 \neq \theta_2$

- Je efekt snížení systolického tlaku novým antihypertenzivem stejný u hypertoniků, kteří kouří, jako u hypertoniků, kteří nekouří?

Střední hodnota efektu u kuřáků: θ_1 $H_0 : \theta_1 = \theta_2$

Střední hodnota efektu u nekuřáků: θ_2 $H_1 : \theta_1 < \theta_2$

Proč nulová hypotéza vyjadřuje nepřítomnost efektu?

- Nulová hypotéza odráží fakt, že se něco nestalo nebo neprojevalo → je stanovena obvykle jako opak toho, co chceme experimentem prokázat.
- **Nulová hypotéza je postavena tak, abychom ji mohli pomocí pozorovaných hodnot vyvrátit.**
- Pro zamítnutí platnosti nulové hypotézy nám totiž stačí najít jeden příklad, kdy nulová hypotéza neplatí – tím příkladem má být náš náhodný výběr (naše pozorovaná data).
- Zamítnout nulovou hypotézu je jednodušší než nulovou hypotézu potvrdit.

Testování hypotéz

- Testování hypotéz se zabývá rozhodováním o platnosti stanovených hypotéz na základě pozorovaných dat.
- Platnost hypotéz ověřujeme pomocí **statistického testu** – rozhodovacího pravidla, které každému náhodnému výběru přiřadí právě jedno ze dvou možných rozhodnutí – H_0 nezamítáme nebo H_0 zamítáme.

Statistický test

- Testování hypotéz probíhá na základě dat.
- **Testované hypotéze odpovídá statistický test**, respektive testová statistika, která umožní ověřit platnost nulové hypotézy.
- **Testová statistika** je vzorec vycházející z pozorovaných dat s rozdělením pravděpodobnosti, sama tedy má také **rozdělení pravděpodobnosti**. Rozdělení pravděpodobnosti testové statistiky za platnosti H_0 se označuje jako „null distribution“.

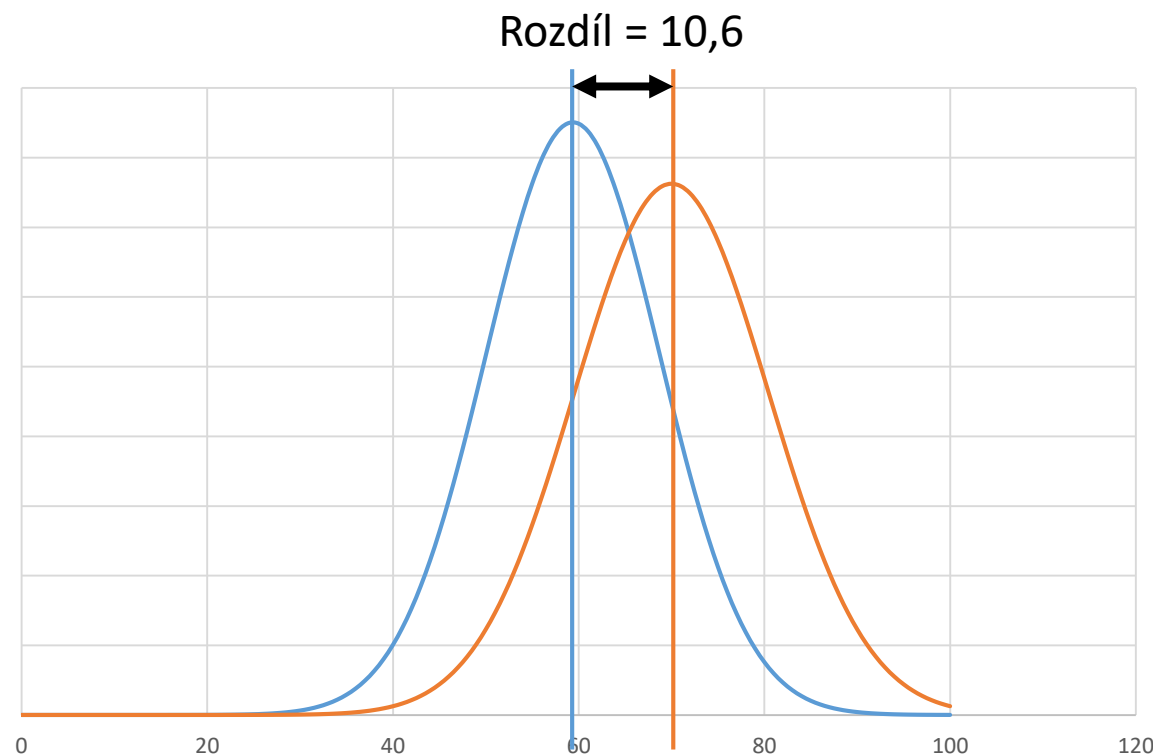
Postup statistického testování

- Formulujeme nulovou hypotézu H_0 (sledovaný efekt je nulový)
- Formulujeme alternativní hypotézu H_A (sledovaný efekt je různý mezi skupinami)
Alternativní hypotéza u parametrických testů může být oboustranná nebo jednostranná.
- Hypotéza musí být stanovena tak abychom mohli vybrat a spočítat tzv. testovou statistiku (např. hypotéza o průměrech bude pravděpodobně řešena pomocí t-testu, jehož testová statistika má t rozdělení)
- Hodnotu testové statistiky vypočítáme na základě pozorovaných hodnot
- Vypočtenou testovou statistiku porovnáme s jejím rozdělením (= rozdělení náhodných rozdílů), posoudíme náhodnost rozdílu a vyslovíme závěr o zamítnutí / nezamítnutí H_0

Na čem závisí hodnota testové statistiky?

- Máme dvě skupiny hodnot, každá je popsána svojí velikostí, průměrem a směrodatnou odchylkou – co ovlivňuje významnost rozdílu jejich průměrů?

N = 100
Průměr = 59,4
SD = 9,4

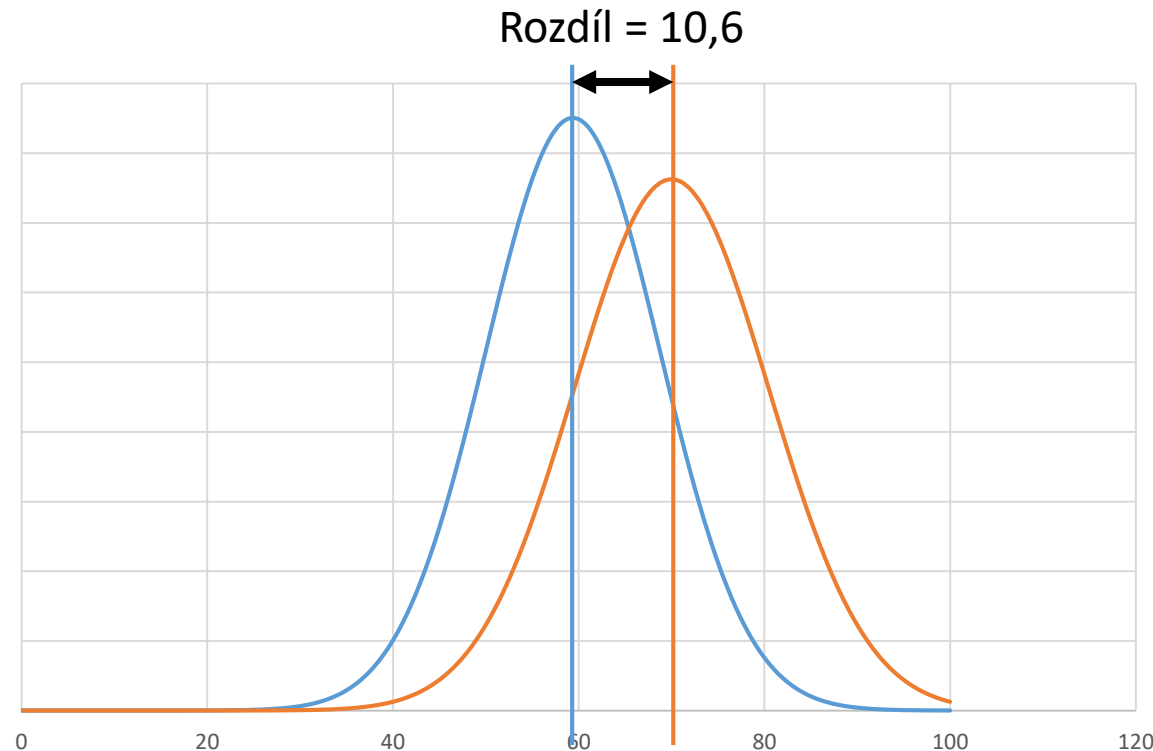


N = 100
Průměr = 70,0
SD = 10,5

Na čem závisí hodnota testové statistiky?

- Máme dvě skupiny hodnot, každá je popsána svojí velikostí, průměrem a směrodatnou odchylkou – co ovlivňuje významnost rozdílu jejich průměrů?

N = 100
Průměr = 59,4
SD = 9,4

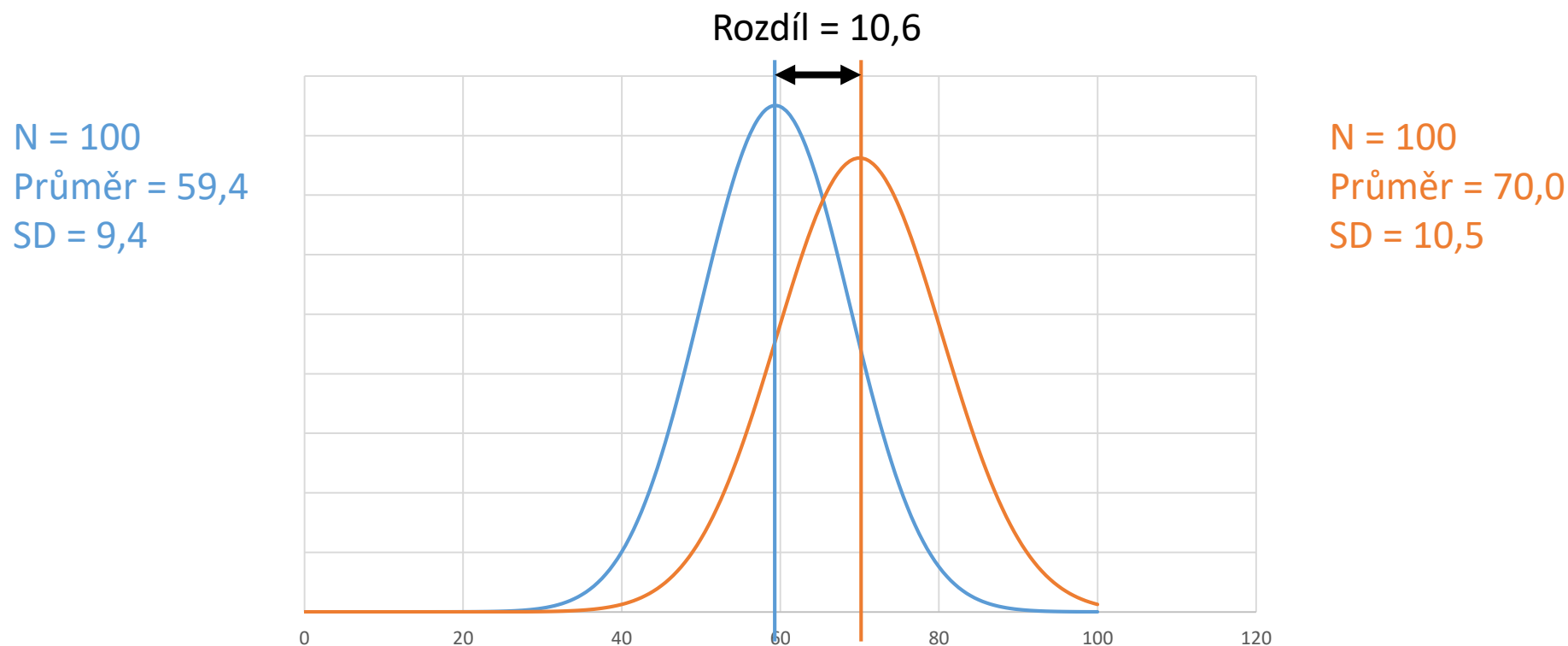


N = 100
Průměr = 70,0
SD = 10,5

- Na velikosti vzorku (větší vzorek = větší významnost) a směrodatné odchylce (větší variabilita = menší významnost) - ovlivňují spolehlivost s jakou odhadujeme srovnávané průměry
- Na velikosti rozdílu mezi srovnávanými průměry (větší rozdíl = větší významnost)

Testová statistika

- Testová statistika kombinuje velikost rozdílu s dalšími charakteristikami dat (velikost vzorku, variabilita atd.), jde vlastně o rozdíl vážený dalšími charakteristikami
- Hodnota testové statistiky je ve vazbě na významnost rozdílu
- Pro finální rozhodnutí o významnosti rozdílu je nezbytné testovou statistiku porovnat s jejím rozdělením náhodných rozdílů (= jaké by bylo rozdělení této statistiky, kdyby byl rozdíl náhodný)



Dva způsoby získání rozdělení testové statistiky

- Testová statistika představuje rozdělení náhodných rozdílů, lze ji získat dvěma způsoby
- **Aproximací na modelové rozdělení**
 - „standardní“ postup, výhodou je snadný výpočet, citlivé na nedodržení předpokladů o rozložení dat
 - Různé testy mají své rozdělení náhodných rozdílů popsány různými modelovými rozděleními (např. t-test pomocí t-rozdělení, test dobré shody pomocí Pearsonova (chi-kvadrát) rozdělení)
- **Permutační metody**
 - Rozdělení náhodných rozdílů je získáno pomocí počítačové simulace buď všech možných nebo zadaného počtu náhodných situací
 - Vhodné pro malé velikosti vzorku nebo situace, kdy není možná aproximace na modelová rozdělení
 - Náročné na výpočetní výkon (v současnosti stále menší problém)
 - Výukově názorné

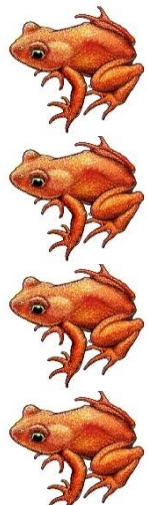
Způsoby testování

- Testování H_0 proti H_A na hladině významnosti α můžeme provést třemi různými způsoby:
 1. Kritický obor (označení W) neboli obor zamítnutí H_0 ,
 2. Interval spolehlivosti,
 3. P-hodnota.

Příklad: permutační testování

Hodnotíme velikost dvou druhů žab, od každého druhu jsme vzorkovali 100 jedinců.

N=100



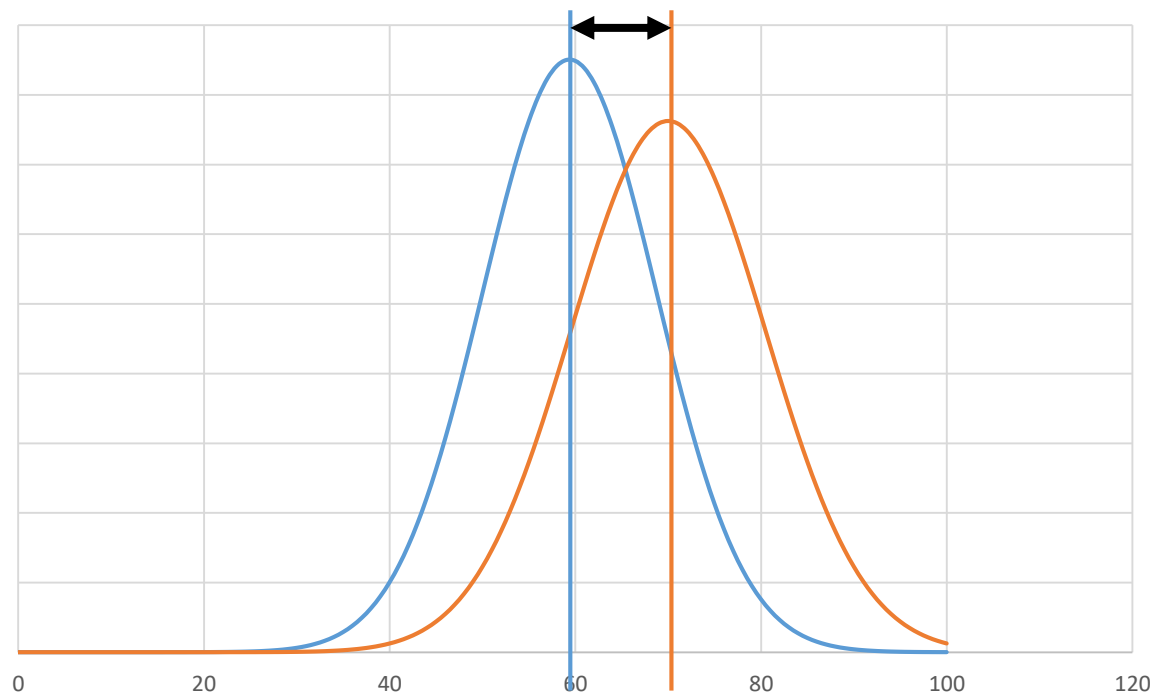
N = 100
Průměr = 59,4
SD = 9,4

Rozdíl ???

N=100



Rozdíl = 10,6



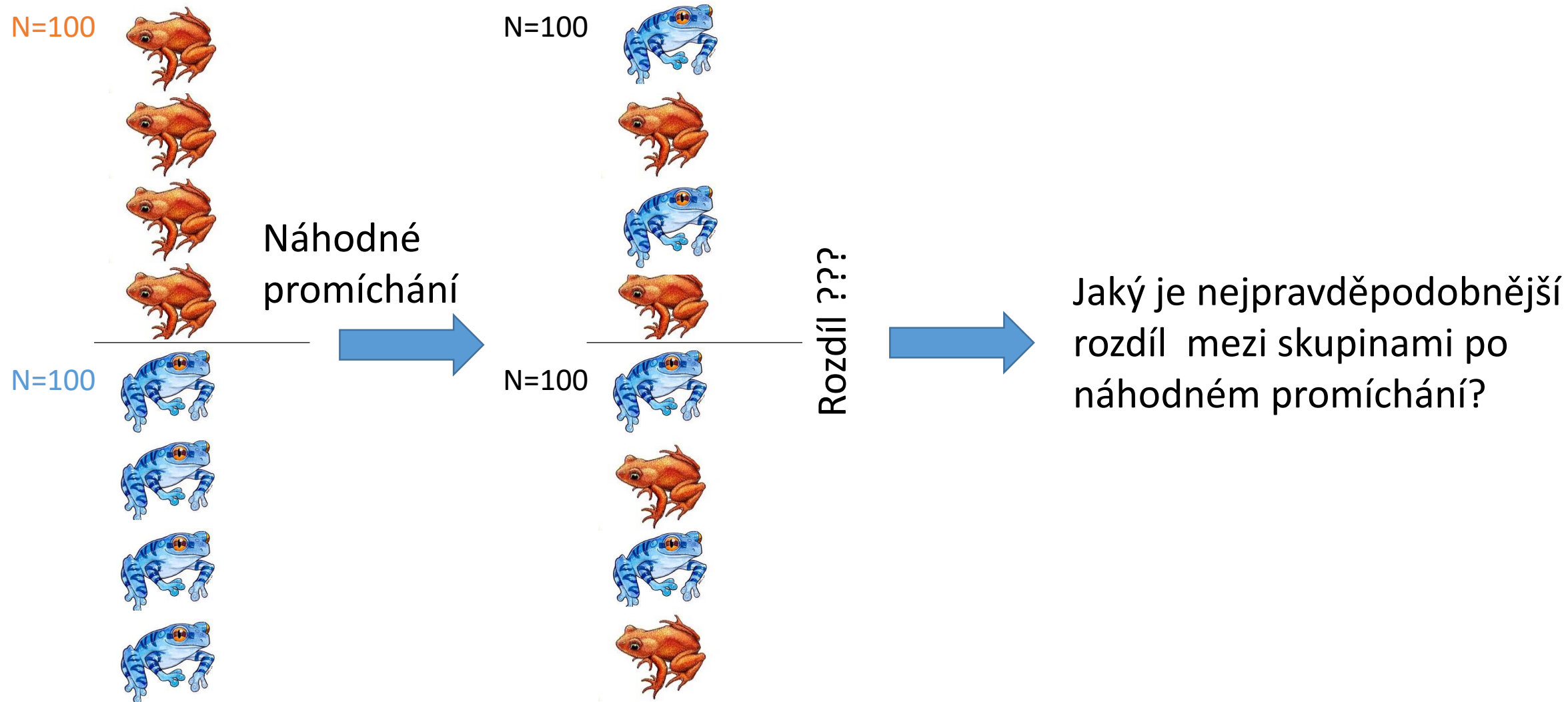
N = 100
Průměr = 70,0
SD = 10,5

Jak zjistit, zda pozorovaný rozdíl je daný pouhou náhodou?

Nasimulujeme si ho !!!!

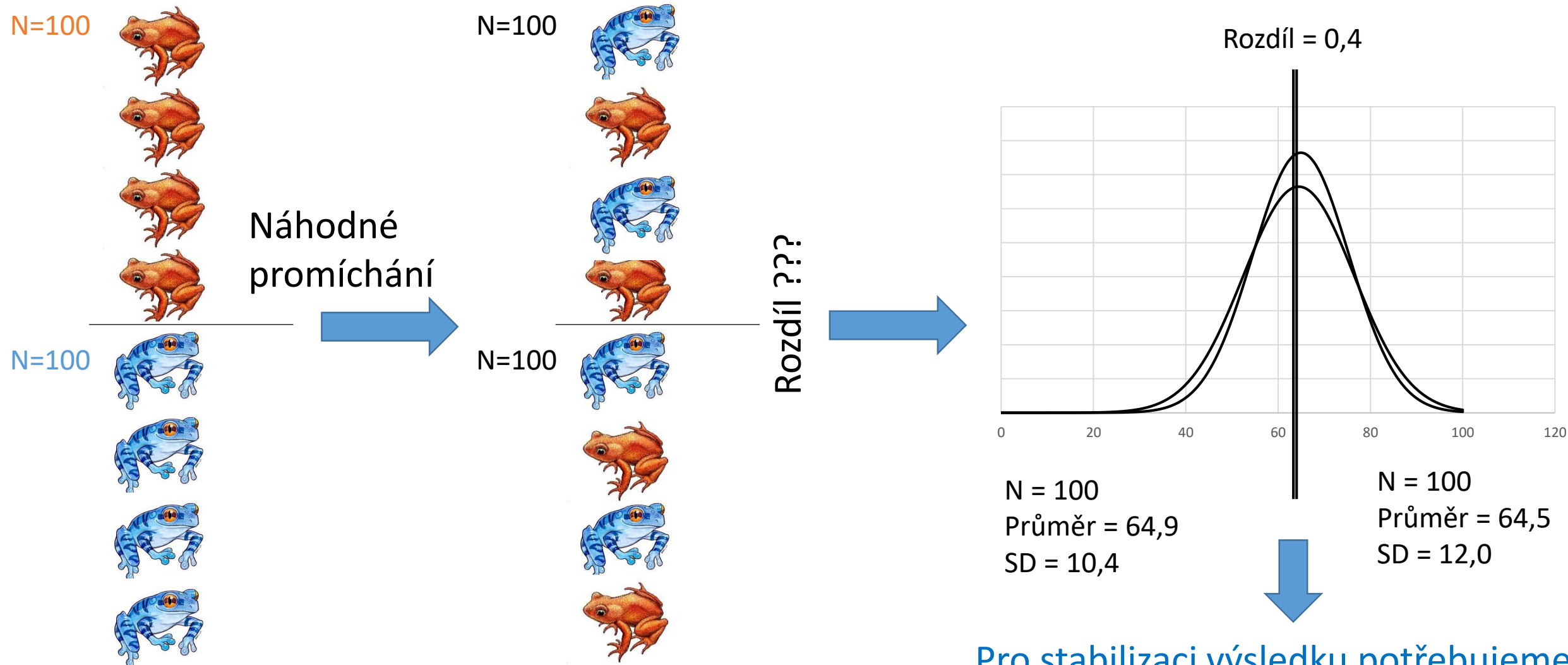
Příklad: permutační testování

Hodnotíme velikost dvou druhů žab, od každého druhu jsme vzorkovali 100 jedinců.



Příklad: permutační testování

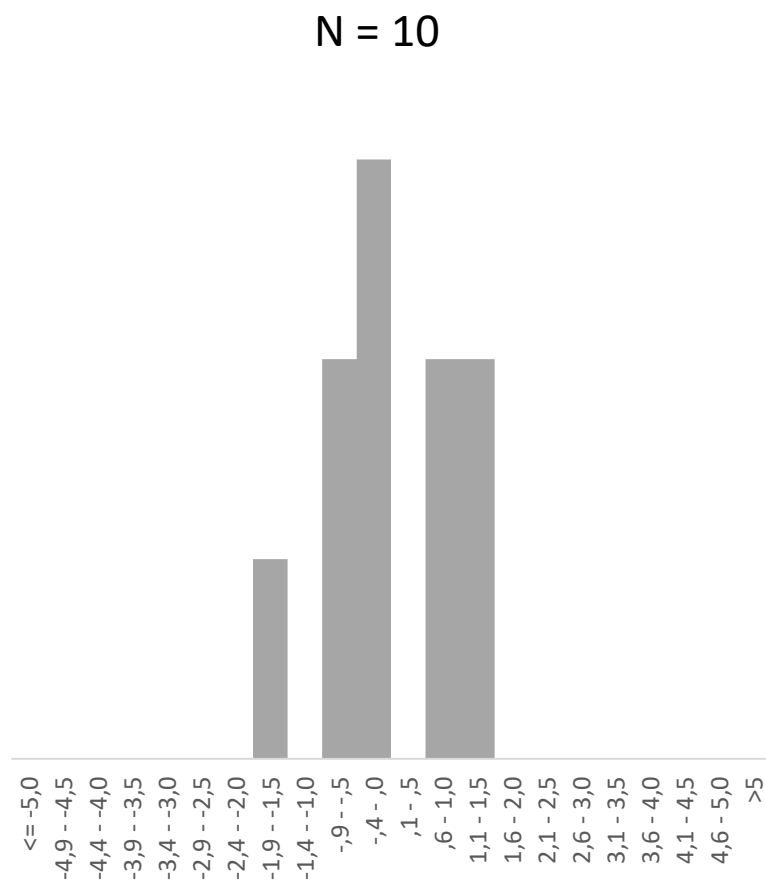
Hodnotíme velikost dvou druhů žab, od každého druhu jsme vzorkovali 100 jedinců.



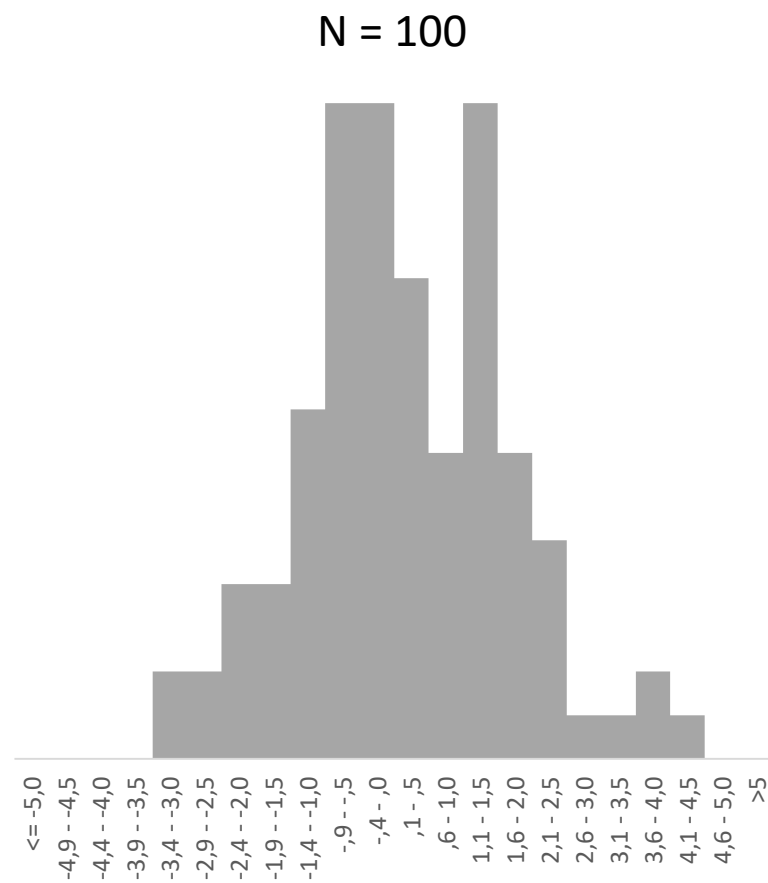
Pro stabilizaci výsledku potřebujeme velký počet permutací.

Výsledky při různém počtu permutací

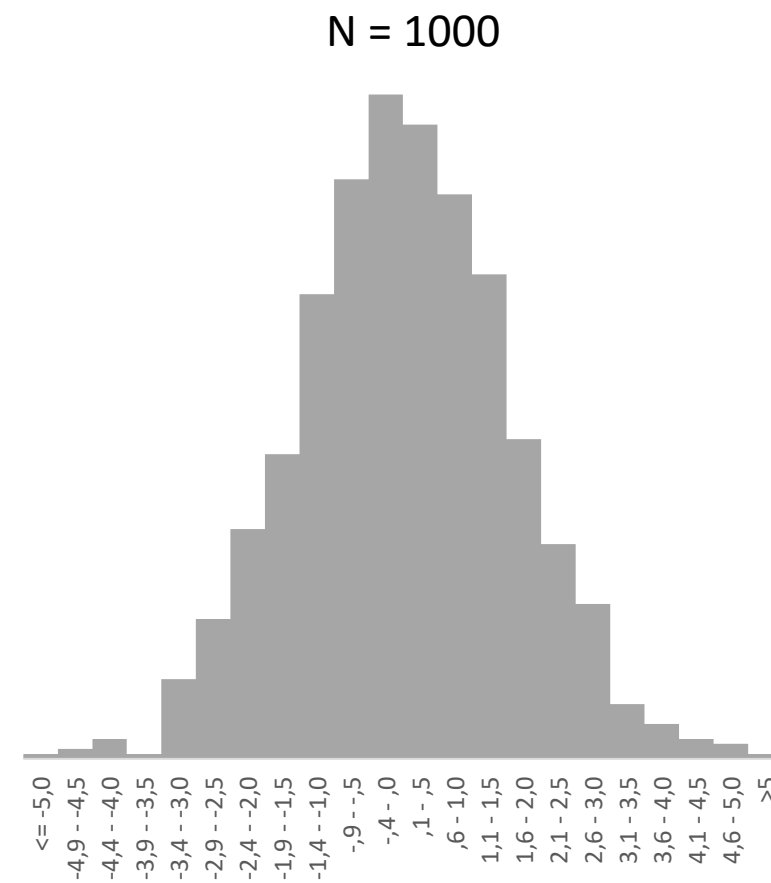
- Se zvyšujícím počtem permutací pozorujeme vytváření rozdělení náhodných rozdílů



Náhodné rozdíly



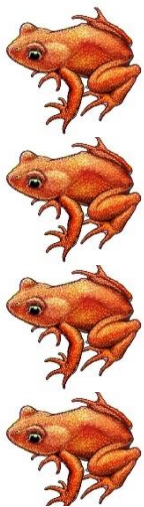
Náhodné rozdíly



Náhodné rozdíly

Náhodné rozdíly vs. pozorovaný rozdíl

N=100



N=100

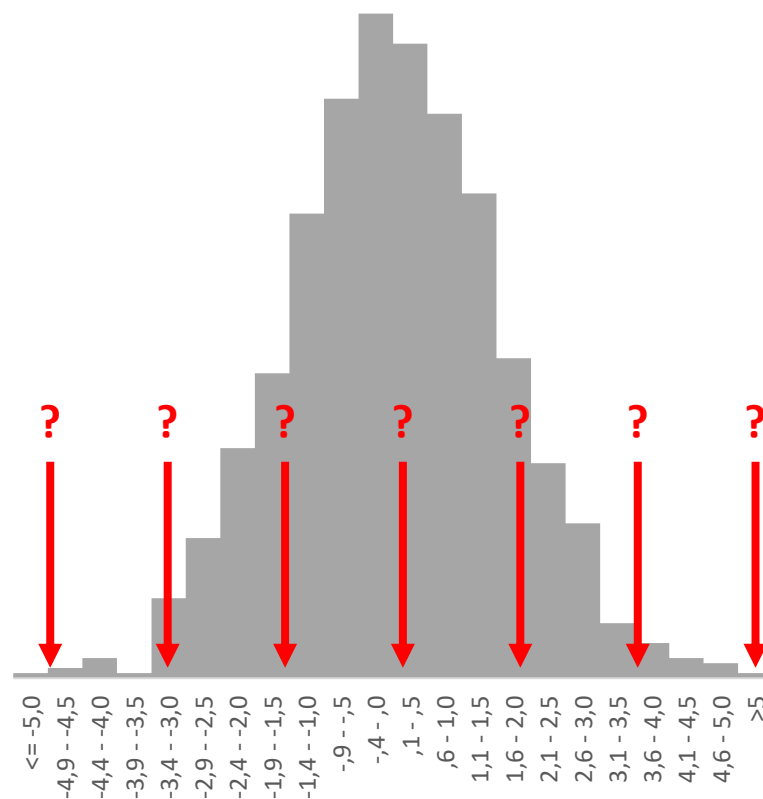


Rozdíl = 10,6



- Reálný rozdíl porovnáme s rozložením náhodných rozdílů

N = 1000



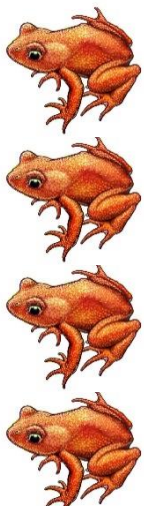
Náhodné rozdíly

Rozložení náhodných rozdílů a jeho využití pro testování

- Stanovíme si kritický obor testové statistiky = s jakou pravděpodobností náhodného vzniku pozorovaného rozdílu jsme schopni se smířit při zamítnutí nulové hypotézy (tedy prohlášení, že rozdíl nepovažujeme za náhodný)
- Nejběžněji se používá kritický obor testové statistiky vedoucí k pravděpodobnosti náhodného rozdílu 0.05 nebo 0.01 (tzv. **hladina statistické významnosti, nejde o přírodní zákon, pouze o domluvu**)
- Náš skutečný rozdíl porovnáme s rozložením náhodných rozdílů a stanoveným kritickým oborem této statistiky
- Pokud skutečný rozdíl leží v kritickém oboru, říkáme, že na dané hladině významnosti zamítáme nulovou hypotézu
- Pro danou hodnotu testové statistiky jsme schopni určit i přesnou pravděpodobnost s jakou existují náhodné rozdíly větší než je náš pozorovaný rozdíl = pravděpodobnost, že námi pozorovaný rozdíl je pouhá náhoda

Statistická významnost pozorovaného rozdílu

N=100



N=100



Rozdíl = 10,6

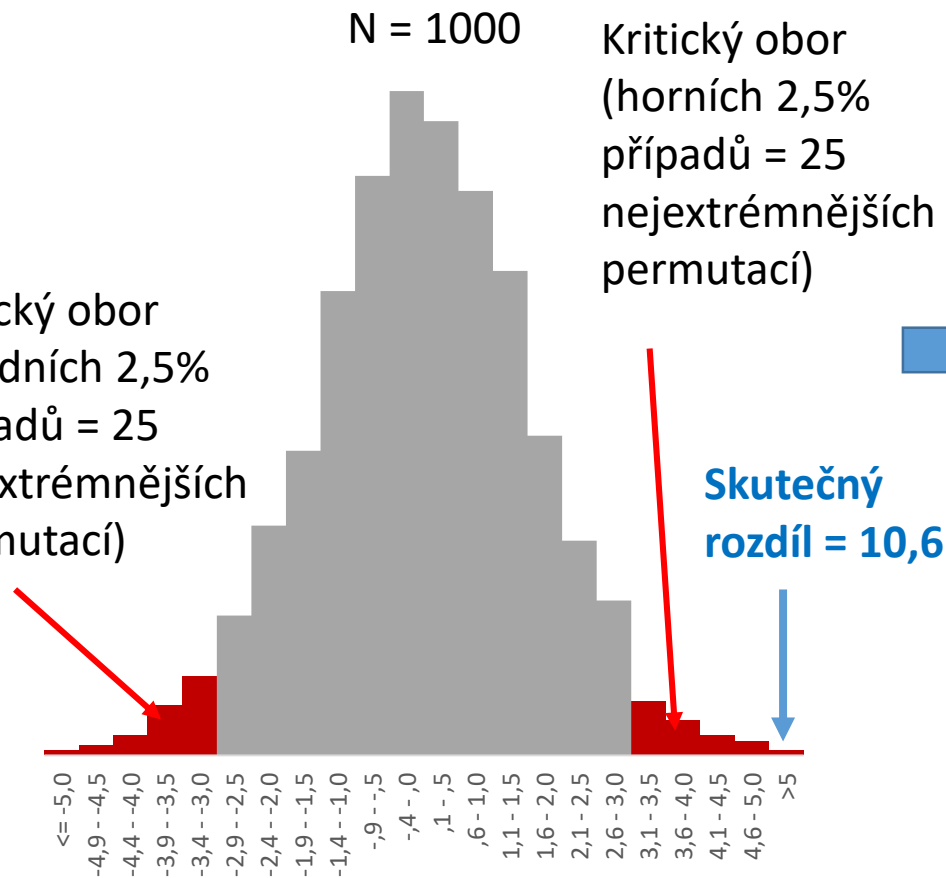


- Jako hladinu statistické významnosti budeme uvažovat 0.05 (5%)

Kritický obor (spodních 2,5% případů = 25 nejextrémnějších permutací)

N = 1000

Kritický obor (horních 2,5% případů = 25 nejextrémnějších permutací)



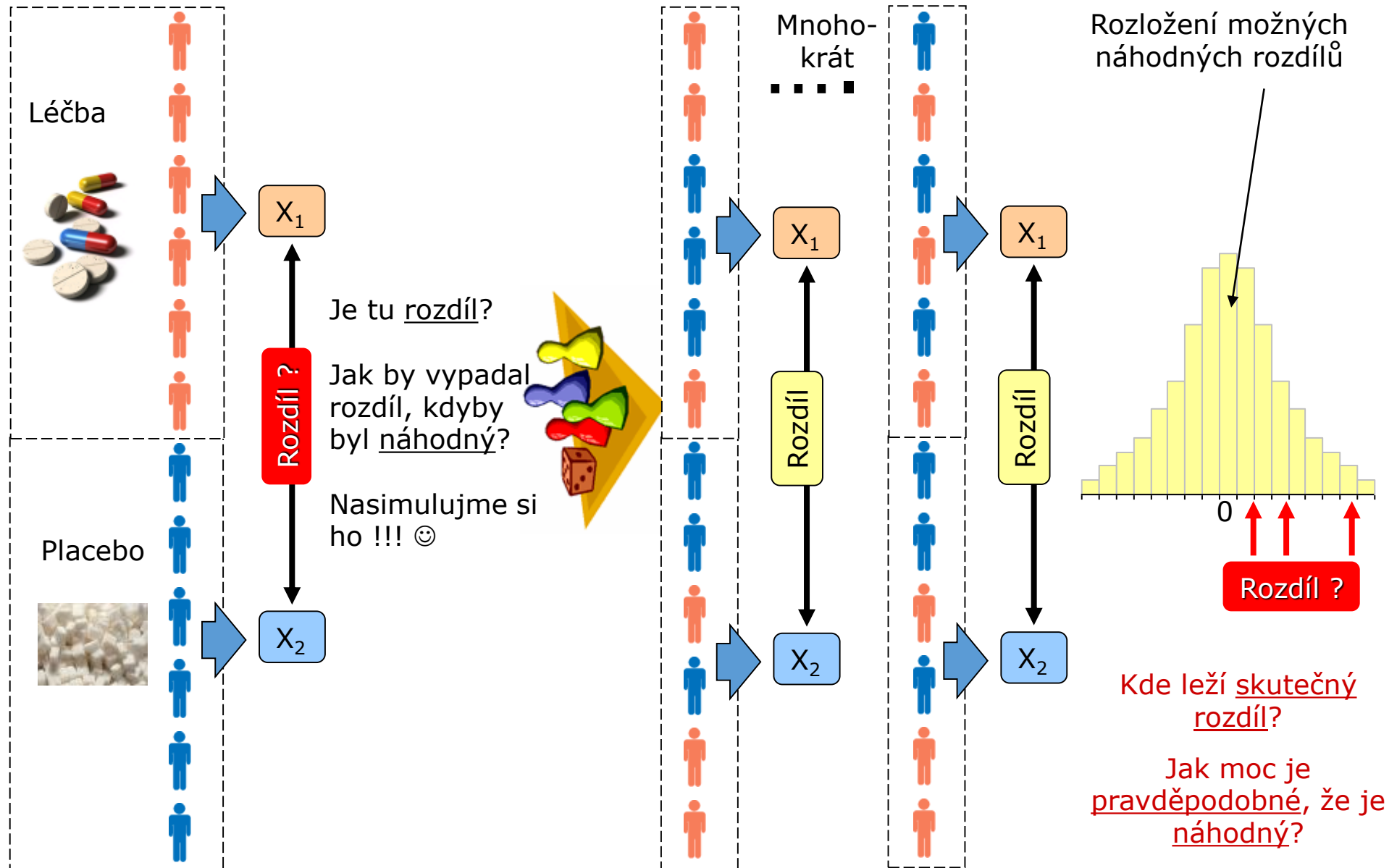
Skutečný rozdíl = 10,6



- Skutečný rozdíl leží v kritickém oboru testové statistiky = **zamítáme nulovou hypotézu o shodě průměru obou skupin**
- Existuje pouze jeden náhodný rozdíl vzniklý permutacemi větší než je skutečný rozdíl = pravděpodobnost, že pouhou náhodou existuje větší rozdíl než je námi pozorovaný je $1/1000 = 0,001$ = **statistická významnost námi pozorovaného rozdílu je $p=0,001$.**

Náhodné rozdíly

Co znamená náhodný rozdíl? Shrnutí.



Kde leží skutečný rozdíl?

Jak moc je pravděpodobné, že je náhodný?

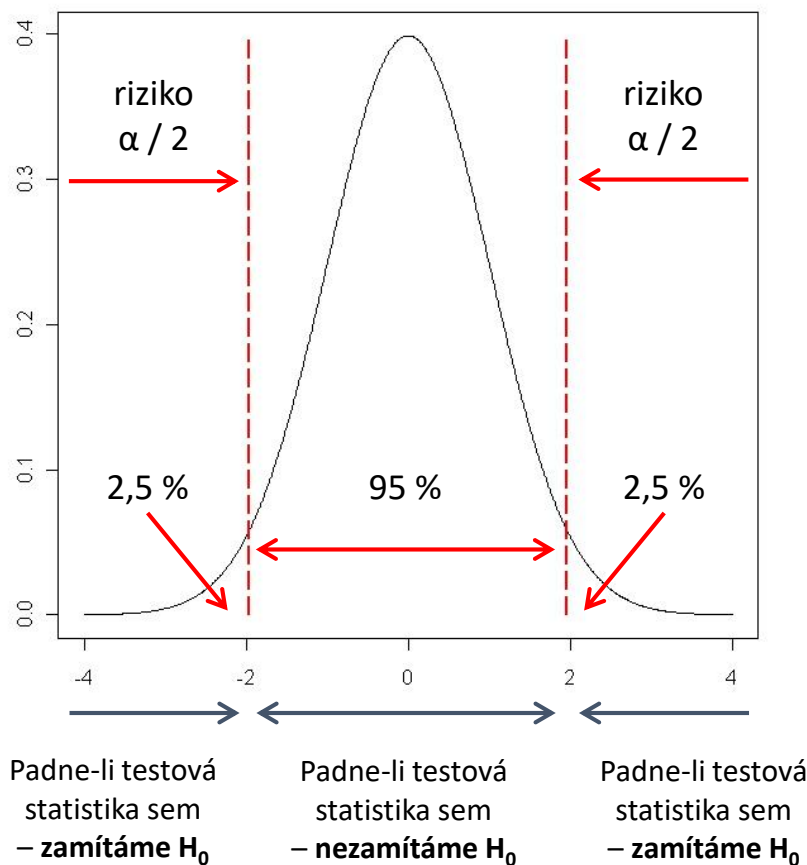
Zamítnutí / nezamítnutí nulové hypotézy

- Hodnotu testové statistiky srovnáme s kvantilem (kritickou hodnotou) jejího rozdělení odpovídajícím zvolené hladině významnosti testu α .
- Představuje-li pozorovaná hodnota testové statistiky extrémnější (méně pravděpodobnou) hodnotu v rámci rozdělení odpovídajícího nulové hypotéze než je kritická hodnota (kvantil) odpovídající zvolenému riziku α , pak nulovou hypotézu zamítáme.

Zamítnutí / nezamítnutí nulové hypotézy

Oboustranný test při $\alpha = 0,05$

$$H_0 : \theta_1 = \theta_2 \quad H_1 : \theta_1 \neq \theta_2$$



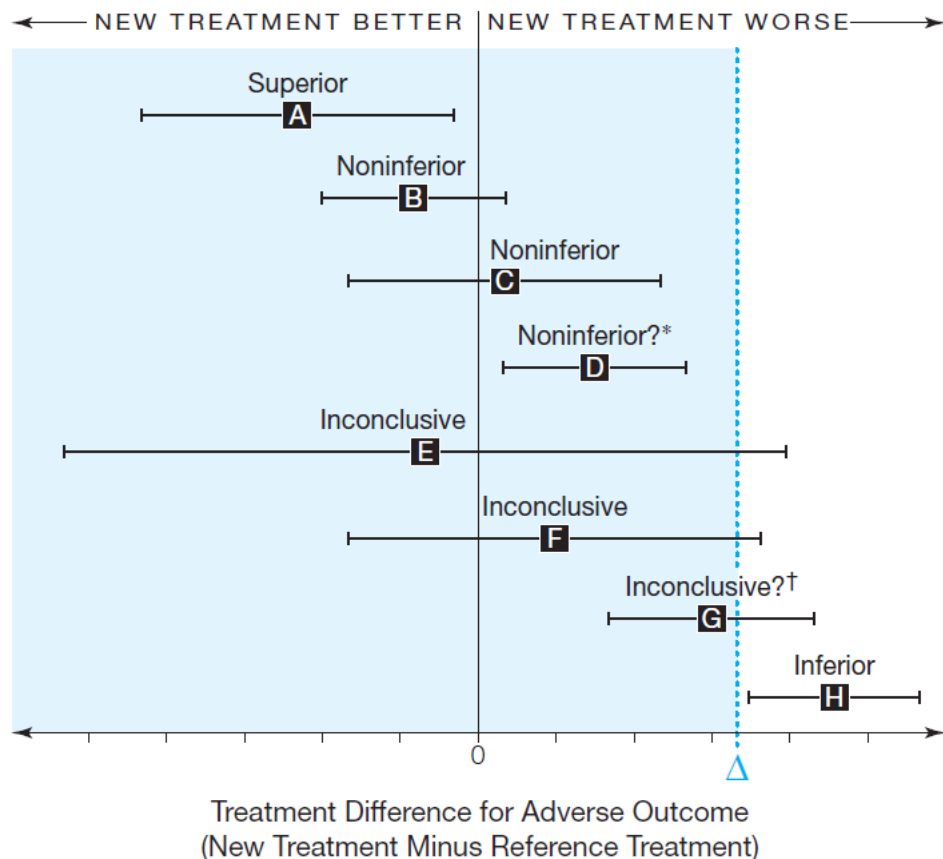
Rozdělení náhodných rozdílů:

- Buď příslušné modelové rozdělení
- Nebo výsledek simulace

Zamítnutí nulové hypotézy:

- Naše testová statistika spadá do kritického oboru
- Odvozená přesná hodnota p je menší než s kritickým oborem spjaté p

Testování pomocí intervalů spolehlivosti

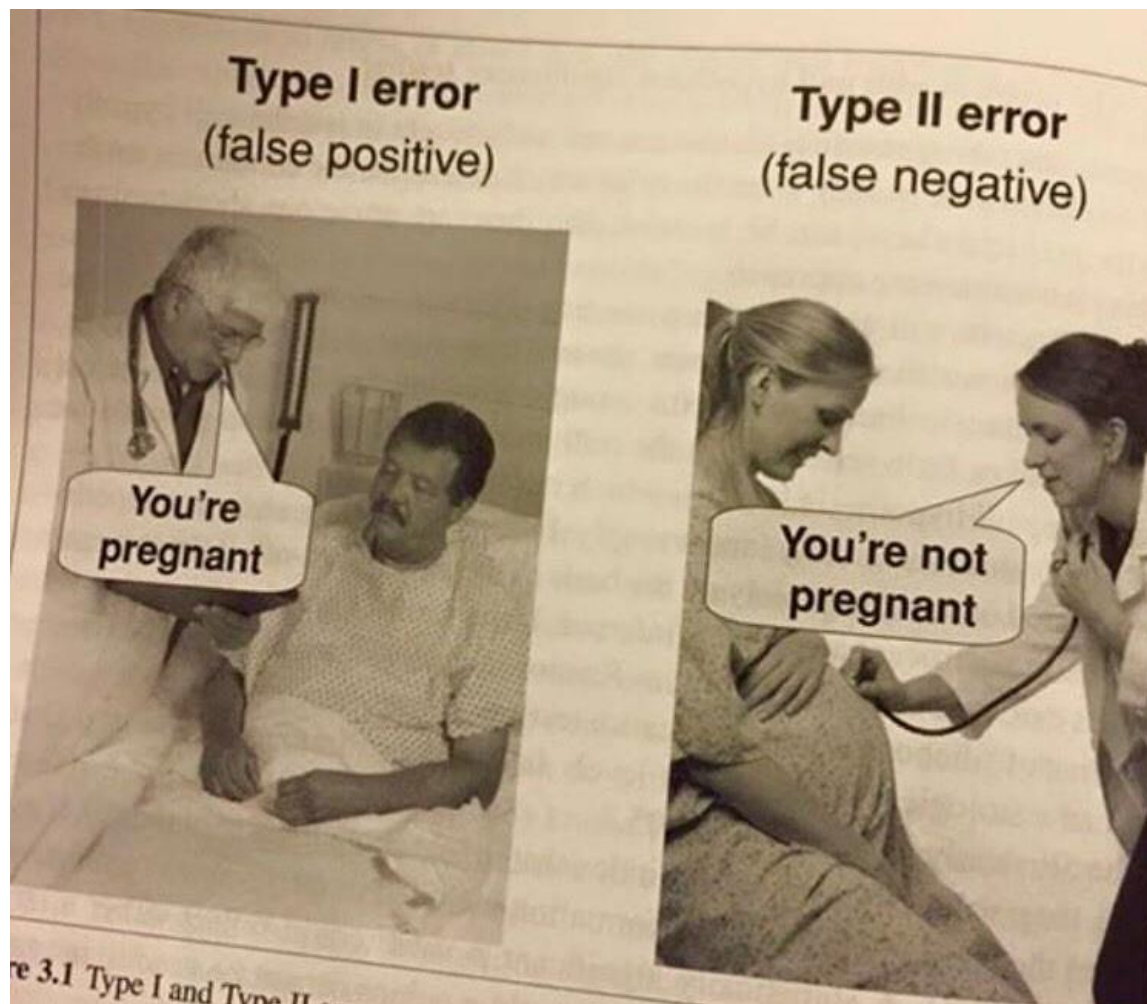


- Principem testování pomocí intervalů spolehlivosti je výpočet intervalu spolehlivosti pro daný rozdíl nebo míru vztahu proměnných a porovnání s referenční hodnotou (např. 0 v případě rozdílu).
- Pokud interval neobsahuje tuto referenční hodnotu, jde o ekvivalent prokázání statistické významnosti rozdílu na dané hladině významnosti (95% interval spolehlivosti je ekvivalentní hladině významnosti 0.05)

Source: Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ; CONSORT Group. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. JAMA. 2006 Mar 8;295(10):1152-60.

Statistics and Informatics Services Group, Department of Reproductive Health and Research, World Health Organization, Geneva.

Možné chyby při testování hypotéz



Co se při rozhodování může stát

- Vzhledem k nulové hypotéz máme čtyři možnosti výsledku rozhodovacího procesu:

Rozhodnutí	Skutečnost	
	H_0 platí	H_0 neplatí
H_0 nezamítneme	správné přijetí platné nulové hypotézy	chyba II. druhu
H_0 zamítneme	chyba I. druhu	správné zamítnutí neplatné nulové hypotézy

- Při rozhodování se můžeme mýlit, můžeme se dopustit dvou chybných úsudků.

Analogie se soudním procesem

- Ctíme presumpci nevinny = předpokládáme, že nulová hypotéza platí.
- **Požadujeme důkaz pro prokázání viny = na základě dat chceme ukázat, že nulová hypotéza neplatí.**
- Když nám bude stačit málo důkazů, zvýší se procento odsouzených nevinných = **chyba I. druhu**, ale zároveň se zvýší i procento odsouzených, kteří jsou skutečně vinni = **správné zamítnutí neplatné nulové hypotézy**.
- Když budeme požadovat hodně důkazů, zvýší se procento nevinných, kteří budou osvobozeni = **správné přijetí platné nulové hypotézy**, ale zároveň se zvýší i procento vinných, kteří budou osvobozeni = **chyba II. druhu**.

Pravděpodobnost výsledků rozhodovacího procesu

Rozhodnutí	Skutečnost	
	H_0 platí	H_0 neplatí
H_0 nezamítneme	správné rozhodnutí $P = 1 - \alpha$	chyba II. druhu $P = \beta$
H_0 zamítneme	chyba I. druhu $P = \alpha$	správné rozhodnutí $P = 1 - \beta$

- Jak je vidět z analogie se soudním procesem, nelze zároveň minimalizovat α i β . V praxi je nutné více hlídat $\alpha \rightarrow$ předem stanovíme maximální hranici pro α (hladina významnosti testu, „level of significance“) a za této podmínky minimalizujeme β .

Co znamená „padnutí testové statistiky“

- Je-li hodnota testové statistiky větší než kvantil příslušný riziku α , pak mohly nastat dvě situace:

- 1. buď H_0 platí a my jsme pozorovali málo pravděpodobný jev**
- 2. nebo H_0 neplatí**

- My pracujeme s rizikem α , tedy málo pravděpodobné jevy jsou součástí našeho rizika, proto v tomto případě volíme možnost 2 a zamítáme H_0 .

Chyby statistického testu jako důsledek našeho rozhodnutí

- Samotná statistická významnost znamená pouze pravděpodobnost toho, že námi pozorovaný rozdíl nebo vztah proměnných je daný pouhou náhodou
- V okamžiku, kdy na základě této pravděpodobnosti provedeme rozhodnutí o neplatnosti nulové hypotézy, smiřujeme se s pravděpodobností (odpovídající dané statistické významnosti), že toto rozhodnutí je chybné a ve skutečnosti nulová hypotéza platí (rozdíl je daný pouhou náhodou)
- Každé naše rozhodnutí o zamítnutí nulové hypotézy v sobě skrývá hada chyby I. druhu



P-hodnota

- P-hodnota vyjadřuje pravděpodobnost za platnosti H_0 , s níž bychom získali stejnou nebo extrémnější hodnotu testové statistiky (samozřejmě vzhledem k jednostrannosti nebo oboustrannosti testu).
- Platí tedy, že čím nižší p-hodnota testu je, tím menší nám tento test indikuje pravděpodobnost, že platí nulová hypotéza. Jinak řečeno, vyjde-li nám při vyhodnocení statistického testu p-hodnota „blízká nule“ (standardně jsou opět přijímány dvě hranice: 5 % a 1 %), znamená to, že naše nulová hypotéza má velmi malou oporu v pozorovaných datech a můžeme ji zamítnout.

P-hodnota

- Výslednou p-hodnotu tedy srovnáme se zvolenou hladinou významnosti α s tím, že nulová hypotéza je zamítána ve chvíli, kdy p-hodnota testu klesne pod tuto hladinu.
- Dá se tedy říci, že ve chvíli, kdy riziko falešně pozitivního výsledku v souvislosti se zamítnutím nulové hypotézy klesne pod vybranou hladinu (např. 5 % nebo 1 %), pak ji zamítáme.
- P-hodnotu lze chápat jako číselný indikátor platnosti nebo neplatnosti nulové hypotézy vyjádřený na pravděpodobnostní škále. A jako každý indikátor, může i p-hodnota indikovat špatný výsledek, neboť si stále musíme uvědomovat, že nám hrozí jak chyba I. druhu, tak chyba II. druhu.

Síla testu

- Pravděpodobnost chyby II. druhu značíme β .
- $1 - \beta$ se nazývá síla testu a vyjadřuje pravděpodobnost, že zamítneme H_0 ve chvíli, kdy H_0 opravdu neplatí.
- Snažíme se sílu testu optimalizovat při zachování hladiny významnosti testu $\alpha \rightarrow$ princip výpočtu velikosti experimentálního vzorku před provedením studie
- Optimalizovat sílu testu a velikost vzorku předem není triviální, můžeme narazit na spoustu problémů – biologické limity, etické limity, finanční limity.

Faktory ovlivňující sílu testu

- **Velikost vzorku:** čím více pozorování (informace o platnosti nulové hypotézy), tím větší má test sílu. Stejně jako u intervalů spolehlivosti, síla testu roste s odmocninou z n .
- **Velikost efektu (účinku):** velikost rozdílu v neznámých parametrech také ovlivňuje sílu testu. Vždy je jednodušší identifikovat jako významný velký efekt, např. velký rozdíl ve středních hodnotách objemu prostaty dvou populací. Naopak je těžší prokázat jako významný menší efekt (menší rozdíl).
- **Variabilita dat:** variabilita dat zvyšuje variabilitu odhadů a ztěžuje tak rozhodnutí o H_0 . Čím více jsou pozorované hodnoty variabilní, tím více dat bude potřeba pro přesný odhad velikosti účinku (rozdílu).
- **Hladina významnosti:** snížíme-li hladinu významnosti testu (např. zvolíme 0,01 místo 0,05), bude obtížnější H_0 zamítnout → sníží se síla testu.