Přednáška 3

Informace a rozdělení dat

Jak vznikají informace

Rozdělení dat

Anotace

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

Vznik informací: pojmy I

Skutečnost

Pozorovatel

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

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

Skutečnost + Jevové pole = Měřitelný prostor

Vznik informací: pojmy II

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

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

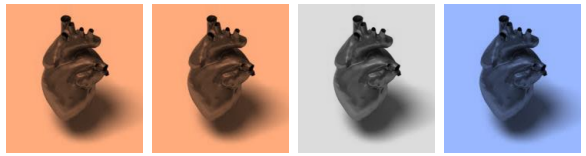
Vznik informací: vzorkování

Statistika hovoří o realitě prostřednictvím výběru z cílové populace

Statistické předpoklady korektního vzorkování je nutné dodržet

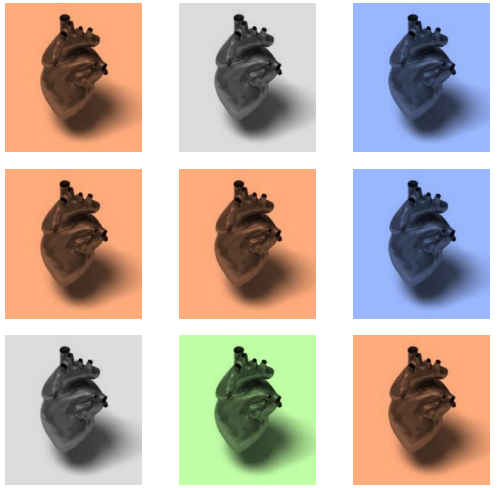
Náhodný výběr z cílové populace

Representativnost: struktura vzorku musí maximálně reflektovat realitu



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

Cílová populace

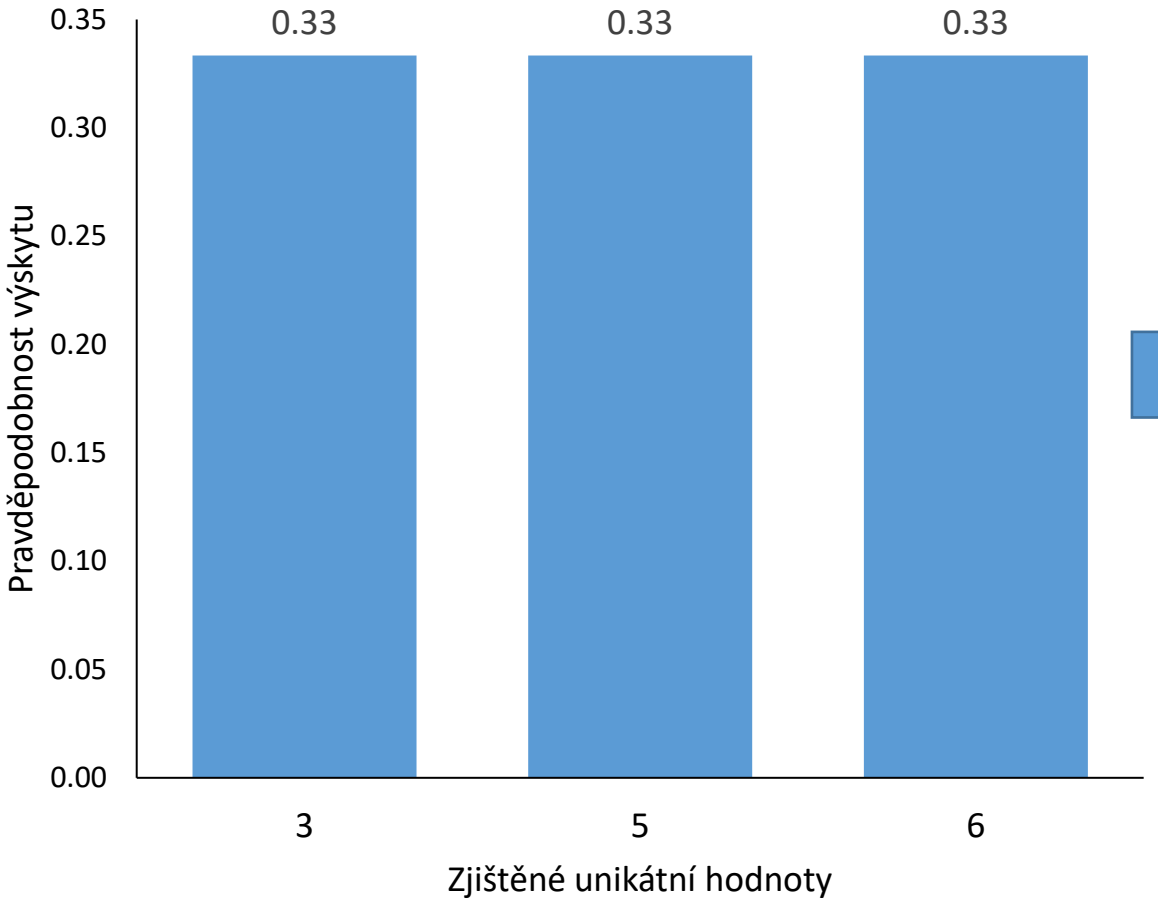


Příklad vzorkování

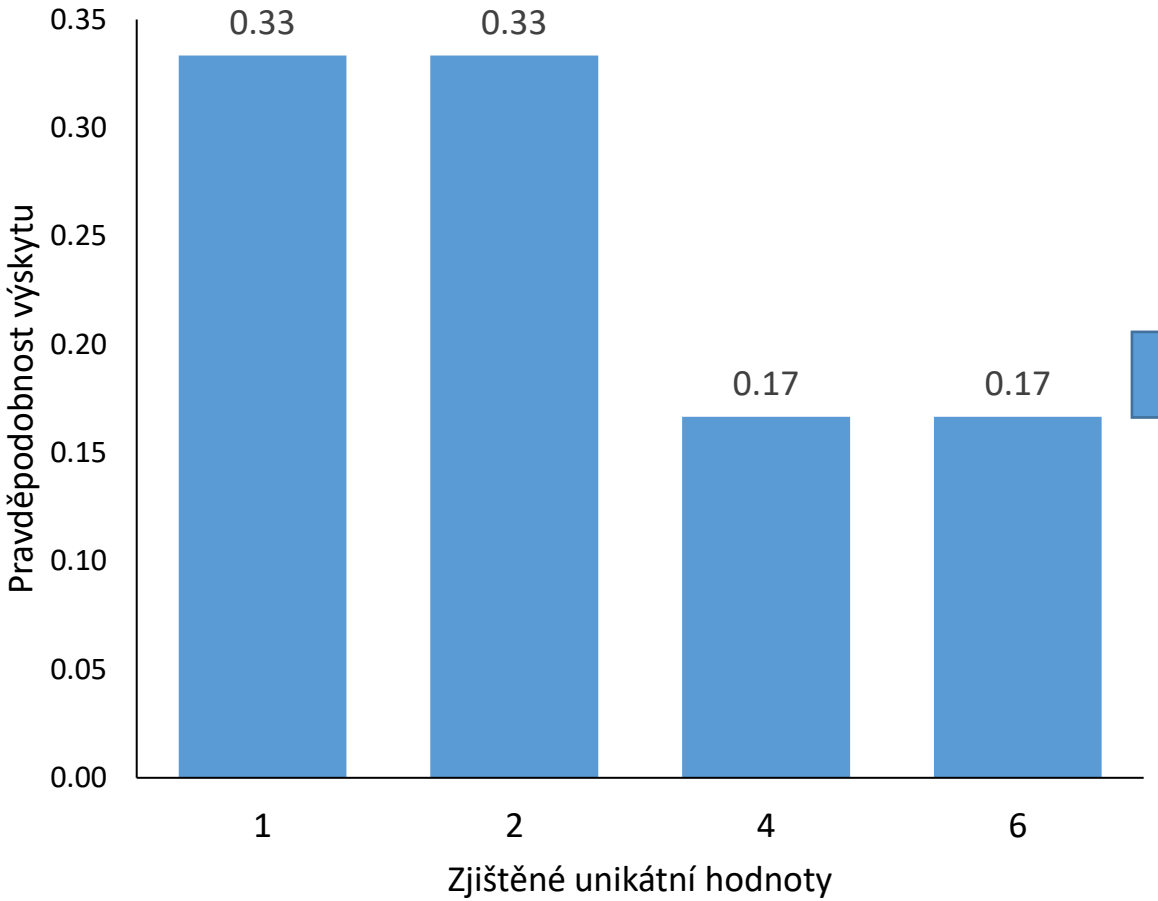
- Na základě vzorkování chceme zjistit vlastnosti nějakého jevu
- Naší cílovou populací budou hody kostkou s neznámými vlastnostmi

- Chceme zjistit vlastnosti neznámé použité kostky

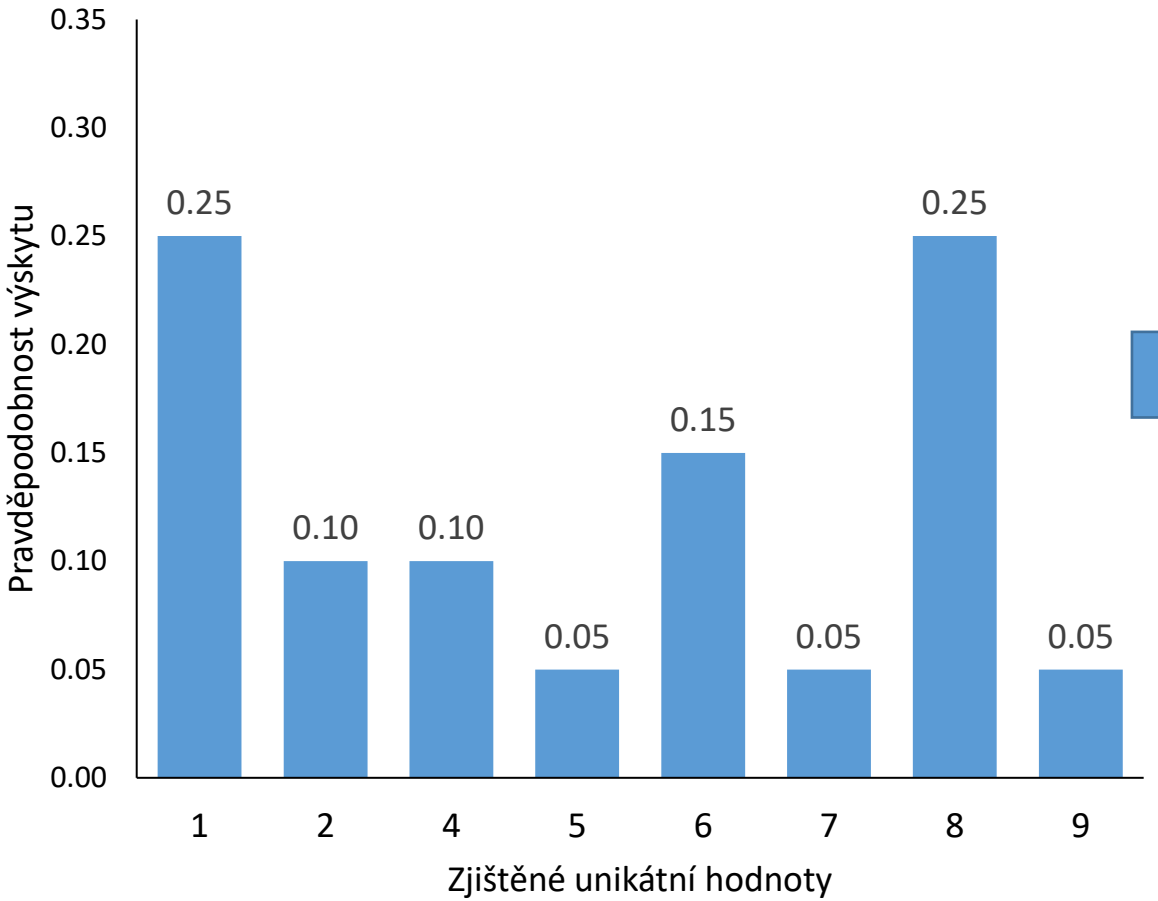
Příklad vzorkování: N=3

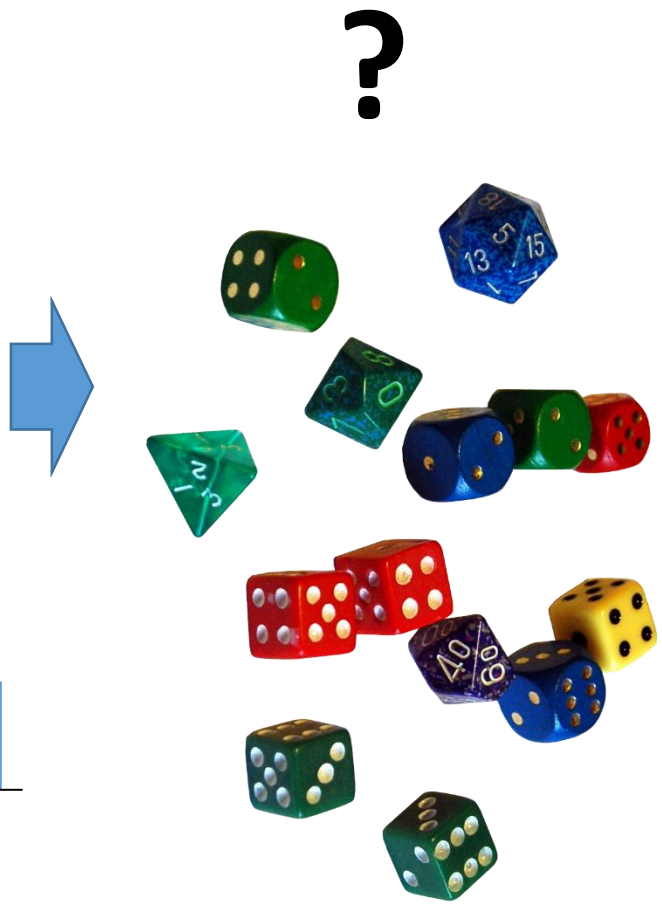


Příklad vzorkování: N=6

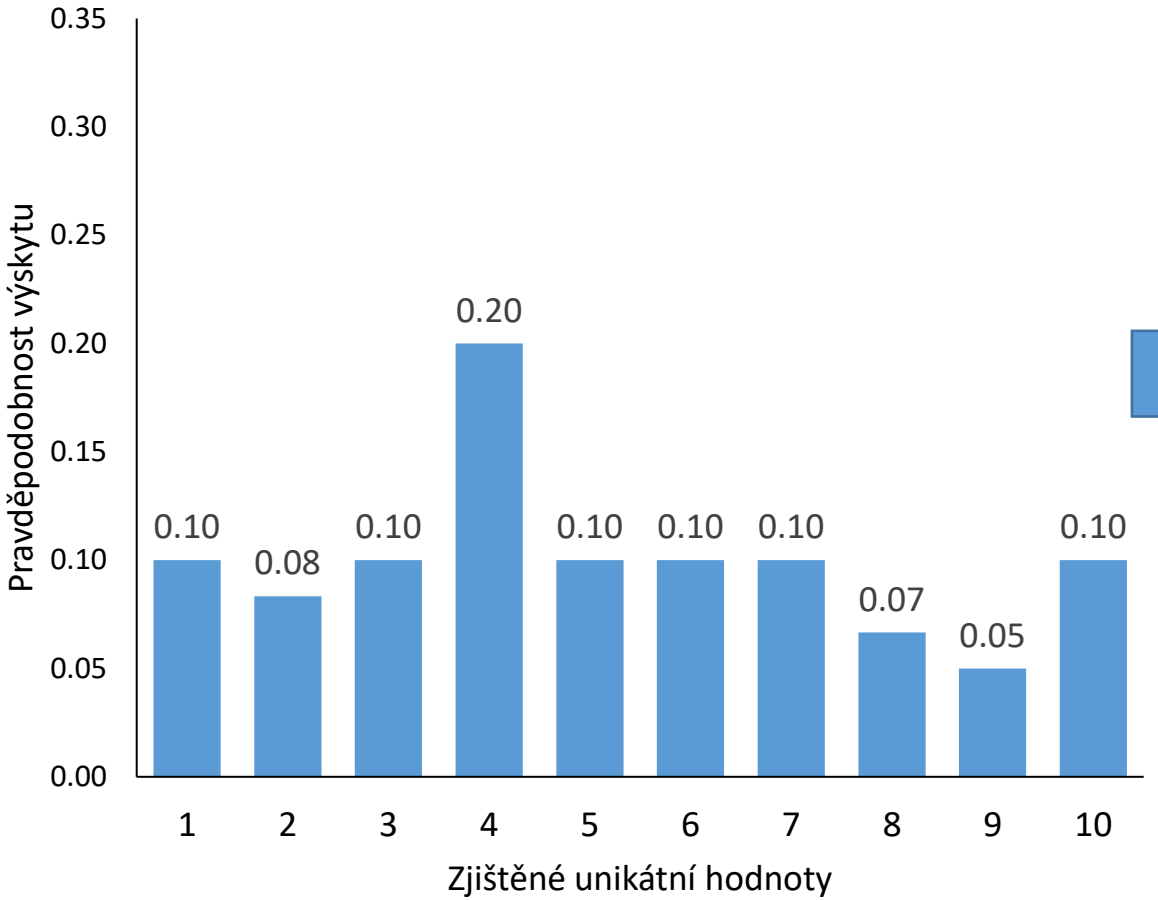


Příklad vzorkování: N=20

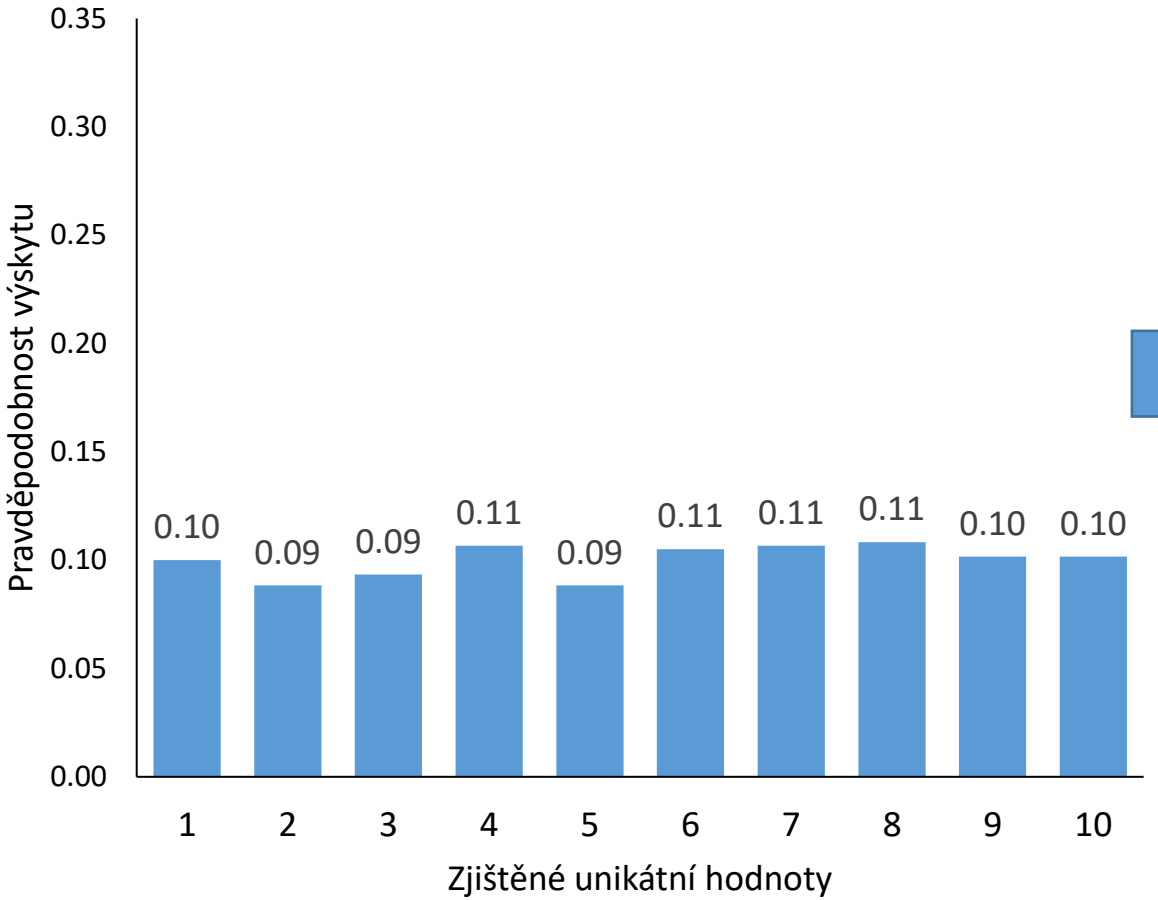


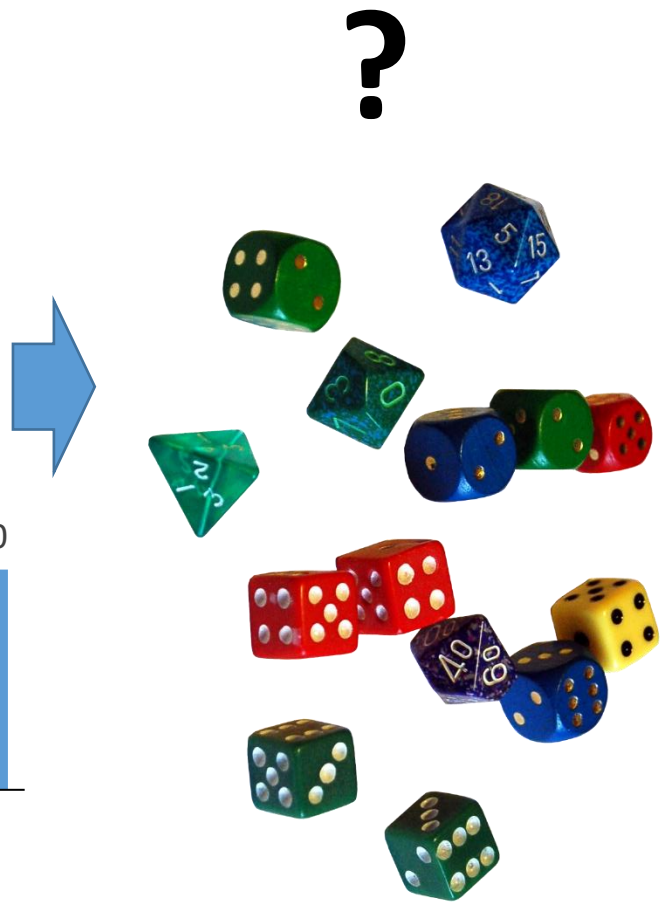


Příklad vzorkování: N=60

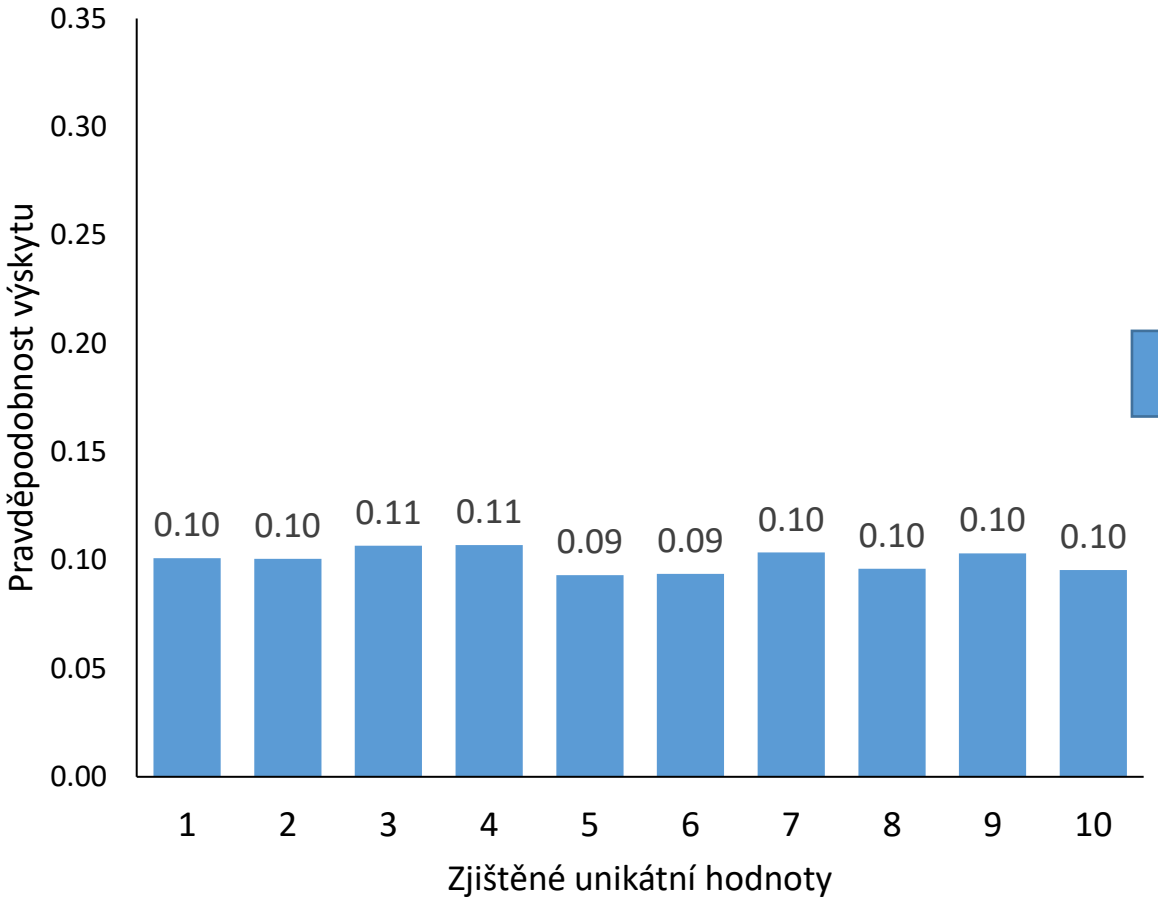


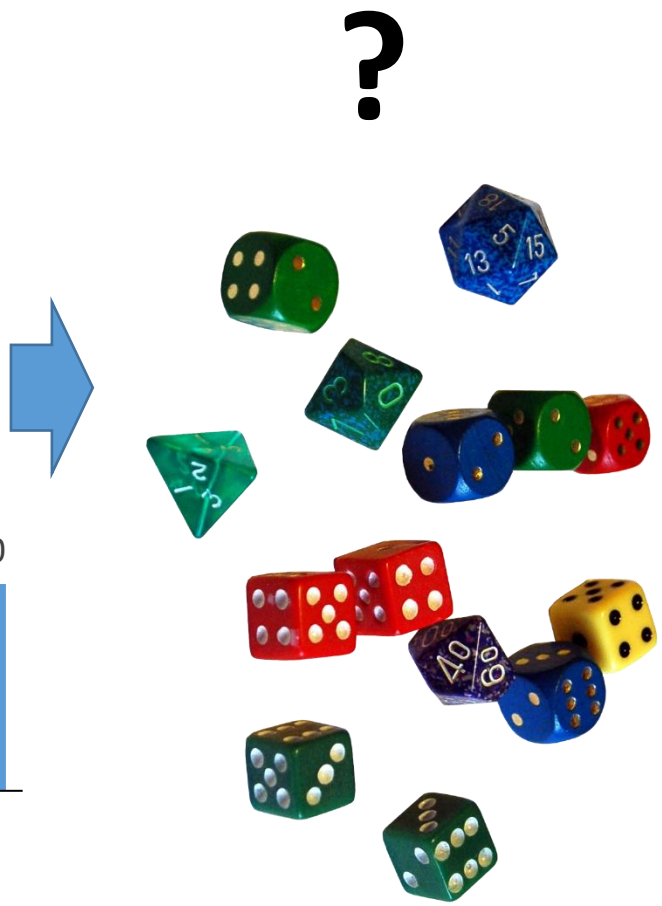
Příklad vzorkování: N=600



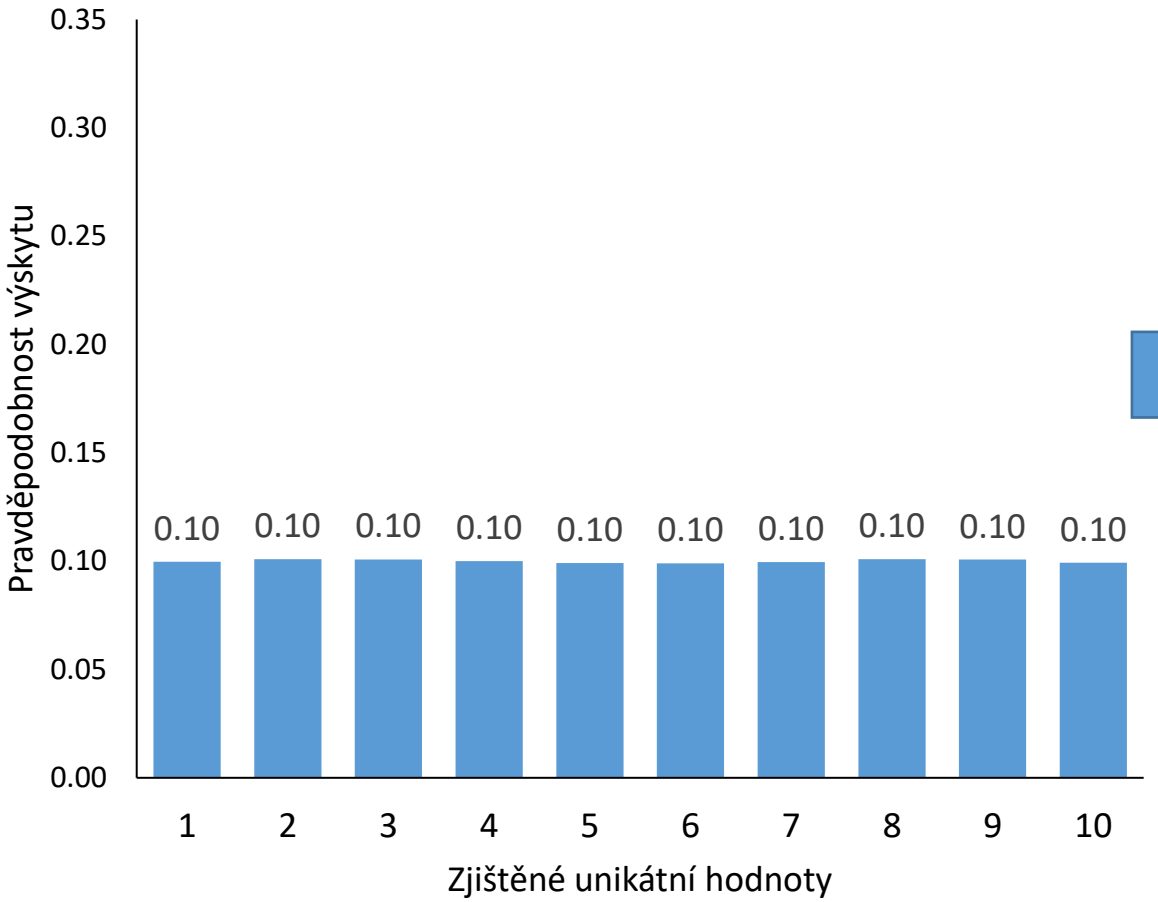


Příklad vzorkování: N=6 000





Příklad vzorkování: N=60 000

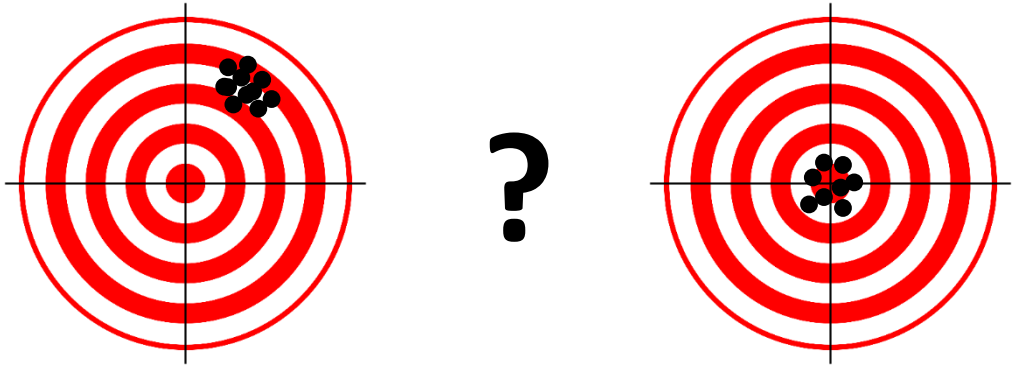


?

Příklad vzorkování: závěr

- Sledovaný jev má pravděpodobně tvar desetistěnné kostky

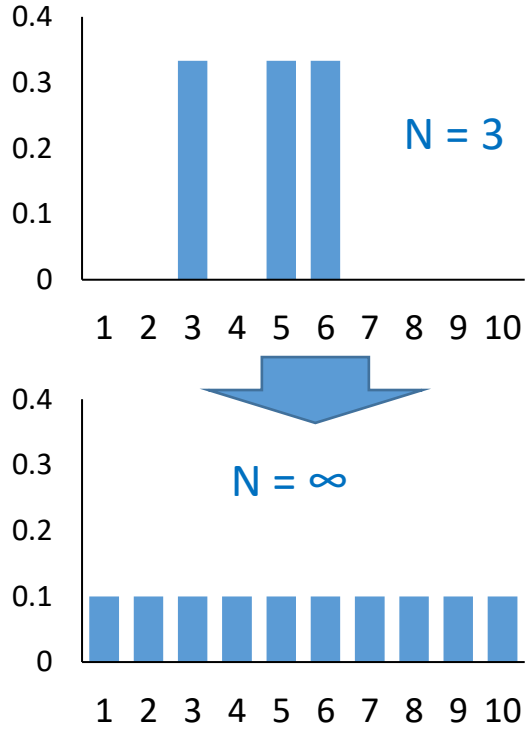
- U složitých stochastických systémů se pravda získá až po odvedení značného množství experimentální práce: musíme dát systému šanci se projevit
- Při realizaci náhodného experimentu roste se zvyšujícím se počtem opakování pravdivá znalost systému (výsledky se stávají stabilnější a spolehlivější)
- Diskutabilní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



Empirický zákon velkých čísel

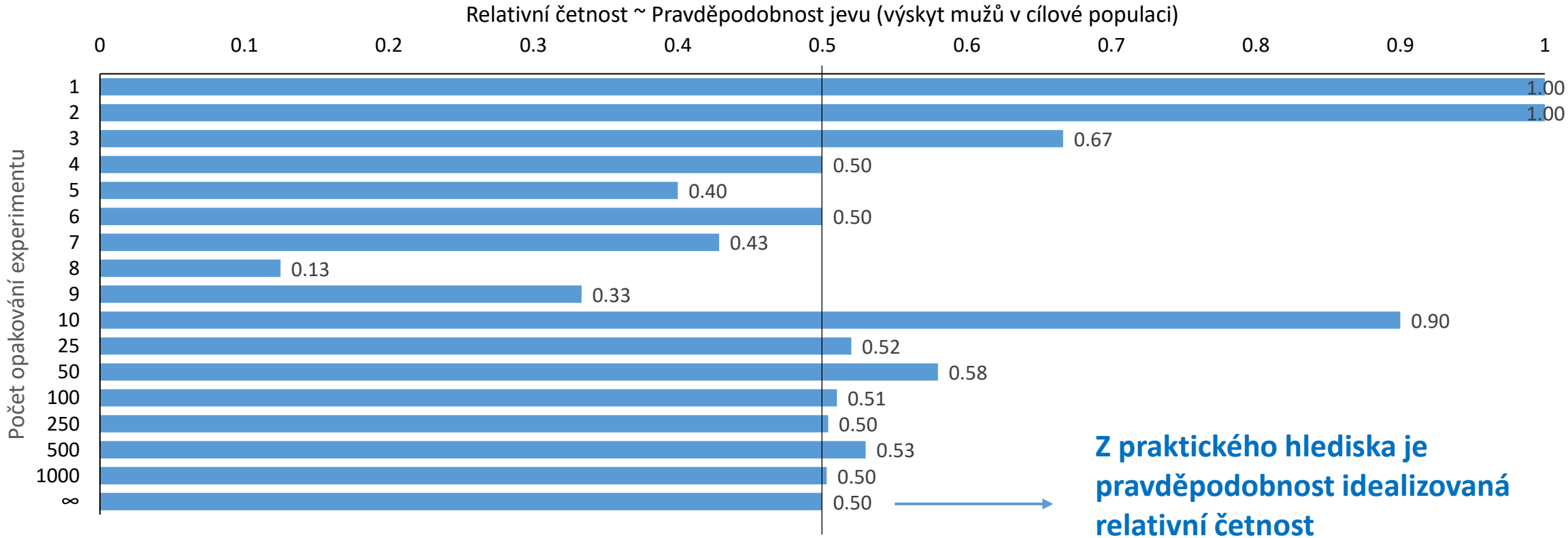
- Při opětovné nezávislé realizaci téhož náhodného experimentu se podíl výskytů sledovaného jevu mezi všemi dosud provedenými realizacemi zpravidla ustaluje kolem konstanty.
- Pravděpodobnost je libovolná reálná funkce definovaná na jevovém poli A (např. hody kostkou), která každému jevu A (např. strany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

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



Empirický zákon velkých čísel: příklad

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

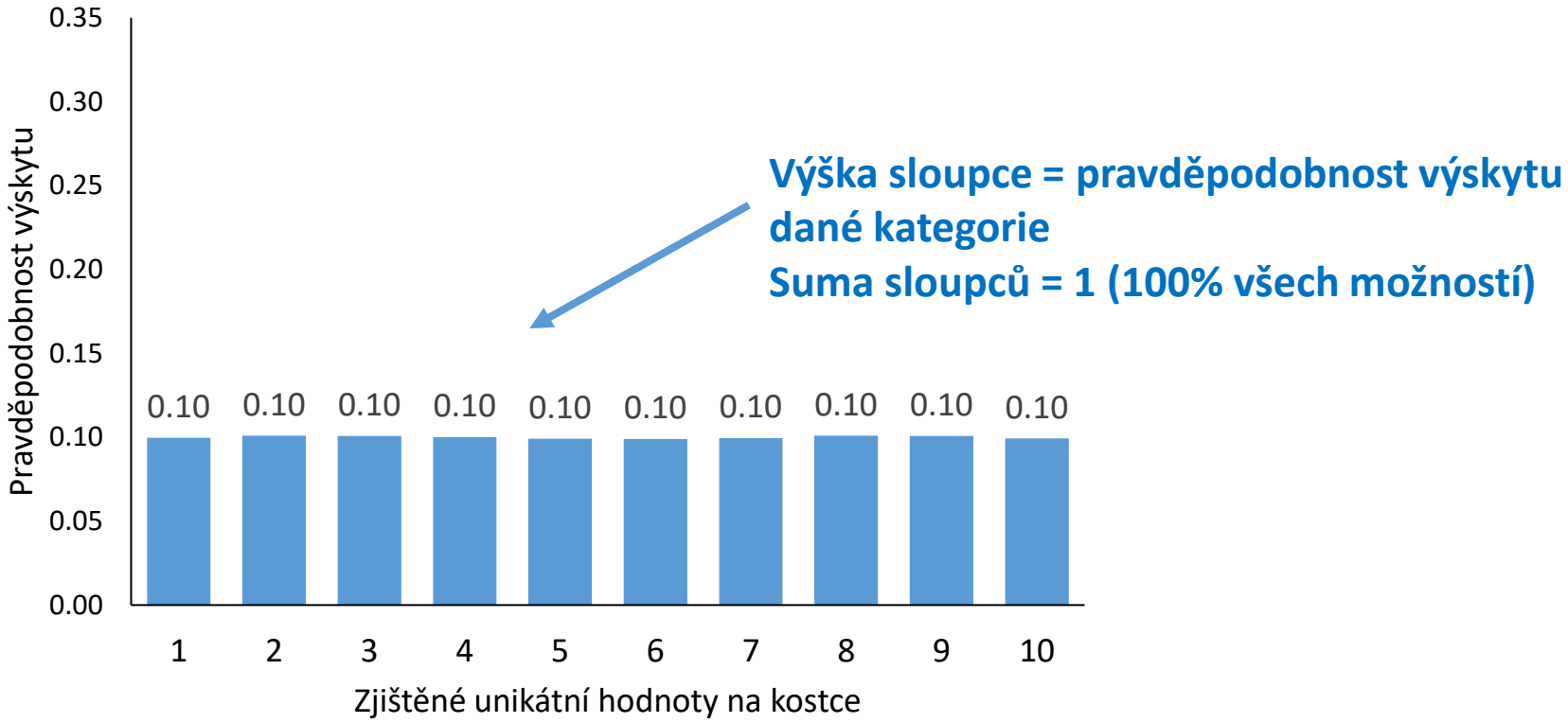


Z praktického hlediska je pravděpodobnost idealizovaná relativní četnost

P=0.5

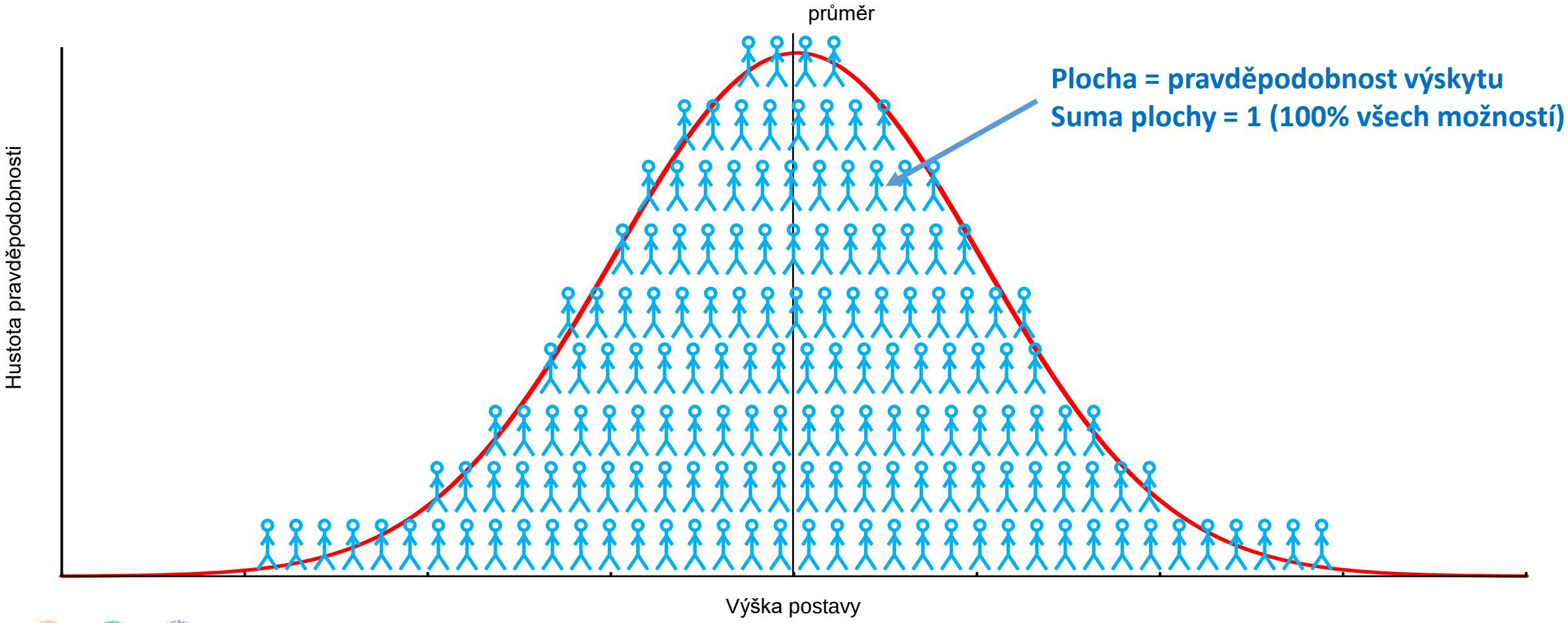
Pravděpodobnost výskytu jevu – rozložení kategoriálních dat

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



Pravděpodobnost výskytu jevu – rozložení spojitých dat

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



Základní typy dat

Spojitá a kategoriální data

Základní popisné statistiky

Grafický popis dat

Anotace

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

Jak vznikají data?

- Záznamem skutečnosti…

Jak vznikají data?

- Záznamem skutečnosti…

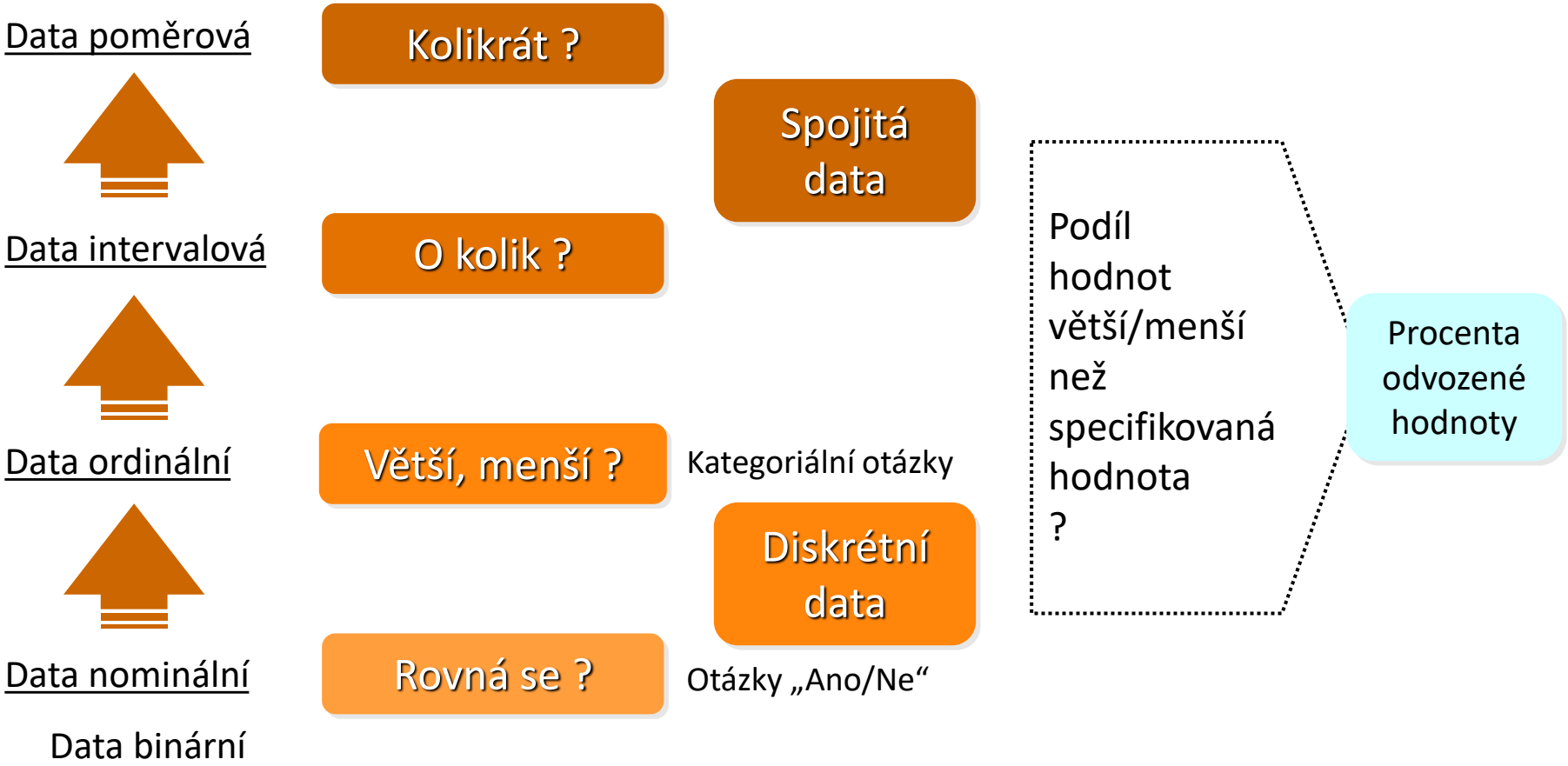
… **kterou chceme dále studovat** → smysluplnost?

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

… více či méně dokonalým → kvalita?

(variabilita = informace + chyba)

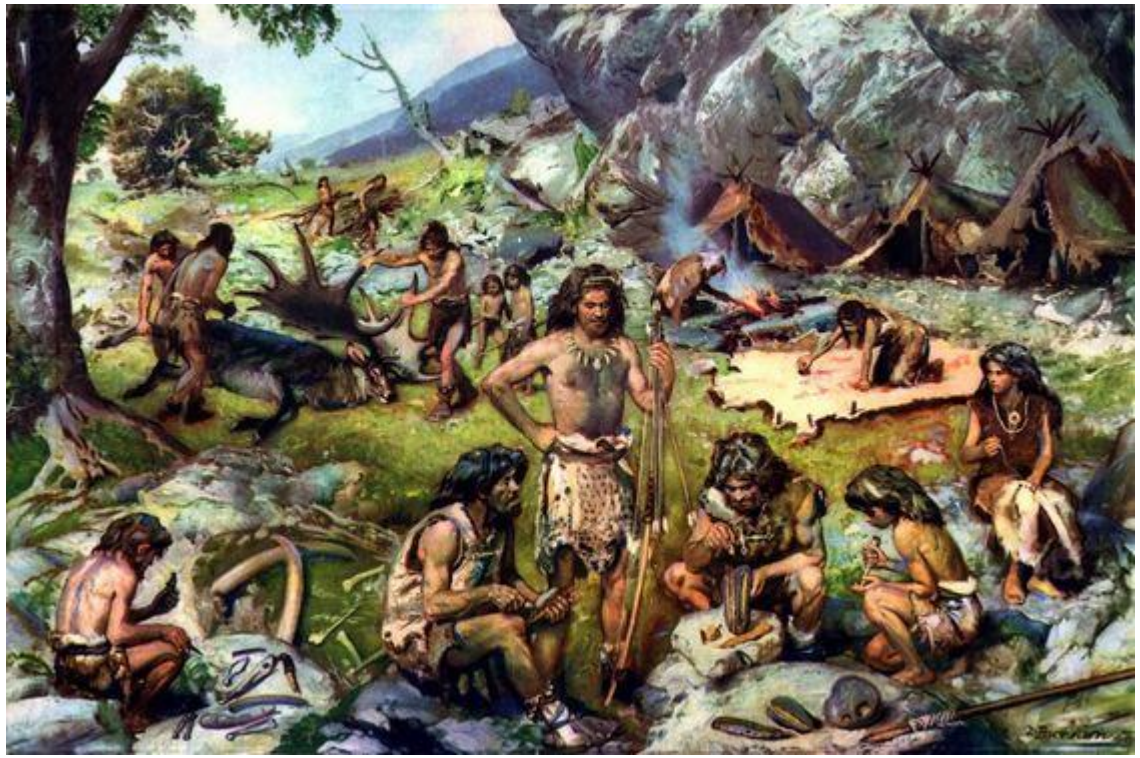
Jak vznikají informace - různé typy dat znamenají různou informaci



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

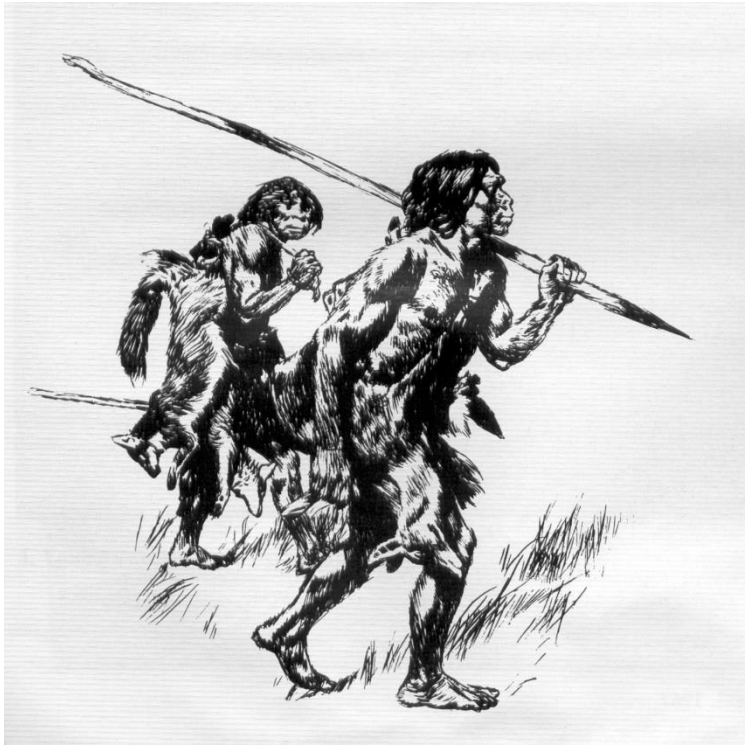
Typy dat a jejich informační hodnota

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



Cílová populace

- Vzorkujeme 3 kategorie sledované proměnné kořist



Kořist

Veverka *Jelen* *Mamut*

Binární data – chytili jsme něco?

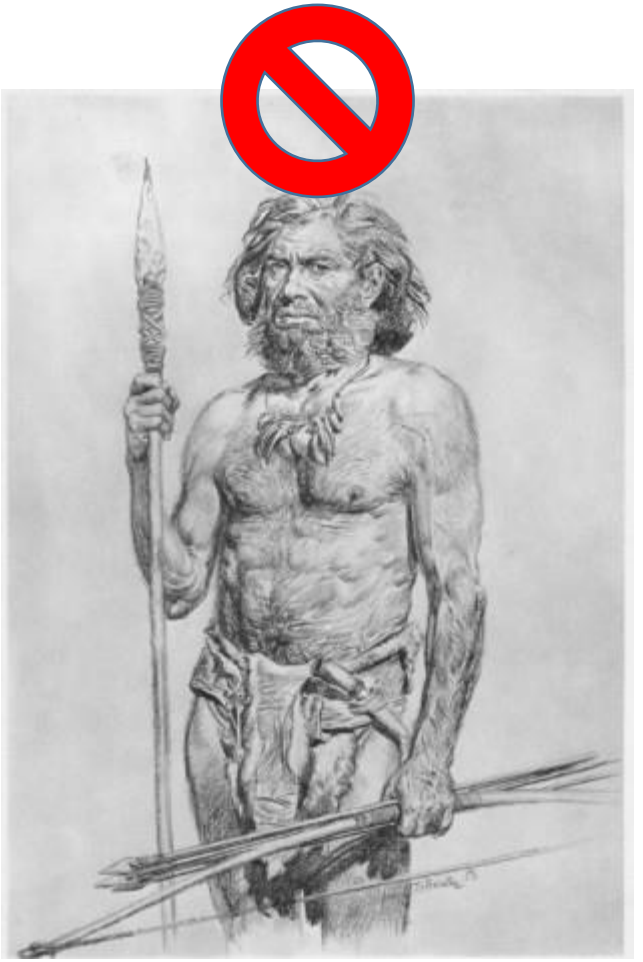
- Informačně nejméně obsáhlá jsou data binární

Hodnotíme dva možné stavy:

Přinesl x nepřinesl kořist

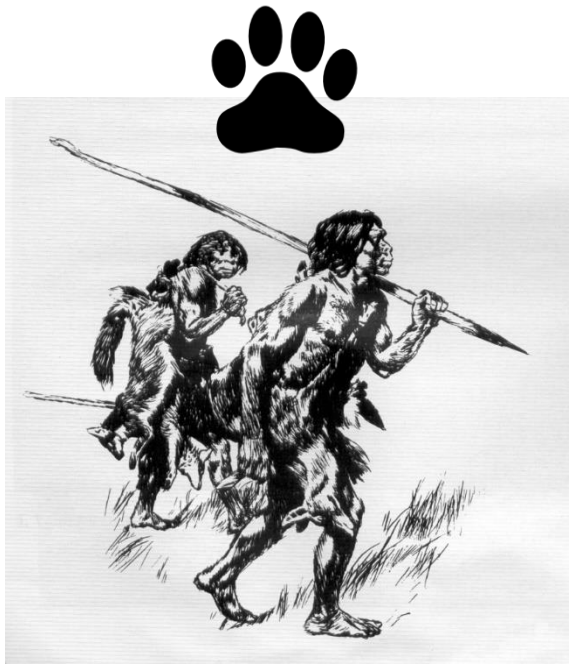
Jak můžeme popsat:

?



Binární data – chytili jsme něco?

- Informačně nejméně obsáhlá jsou data binární

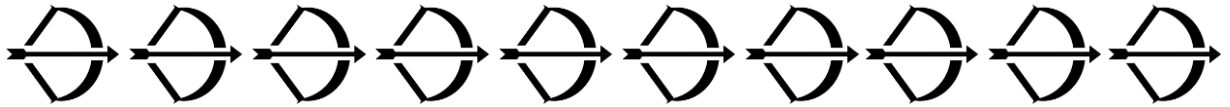


Hodnotíme dva možné stavy:

Přinesl x nepřinesl kořist

Jak můžeme popsat:

Celkový počet lovů (báze hodnocení)

 N=10

Počet úlovků (absolutní četnost)

 N=7

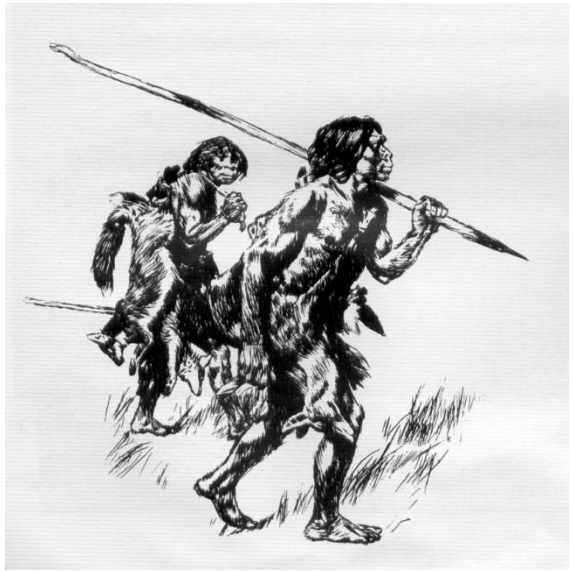
Podíl úspěšných lovů (relativní četnost) nebo nejčetnější kategorie (modus)

 N = 7 (70%)

Jsou binární data dostatečná za všech okolností?

Kategoriální data – co jsme chytili?

- Více informací získáme z dat kategoriálních



Hodnotíme několik možných stavů:

Jak můžeme popsat:

Celkový počet lovů (báze hodnocení)

N = 4 (40%)

Počet různých kategorií úlovků (absolutní četnost)

N = 1 (10%)

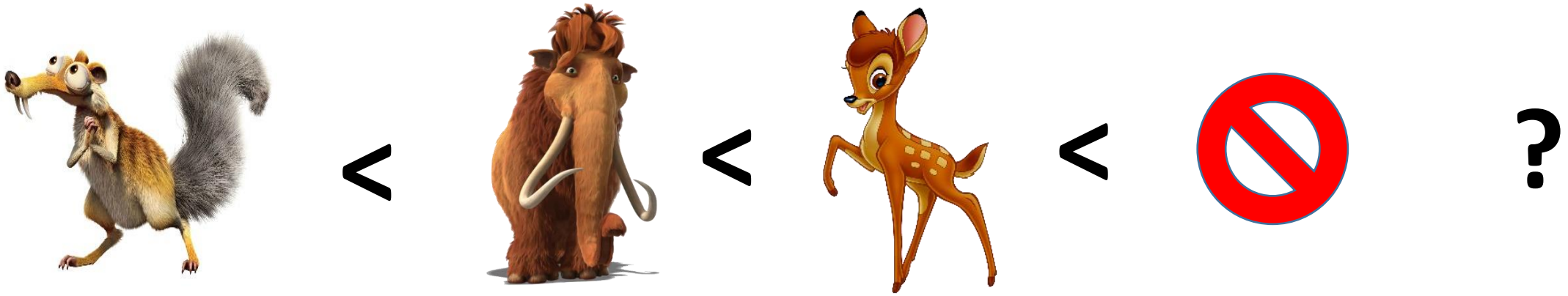
Podíl úspěšných lovů různých kategorií úlovků (relativní četnost) nebo nejčetnější kategorie (modus)

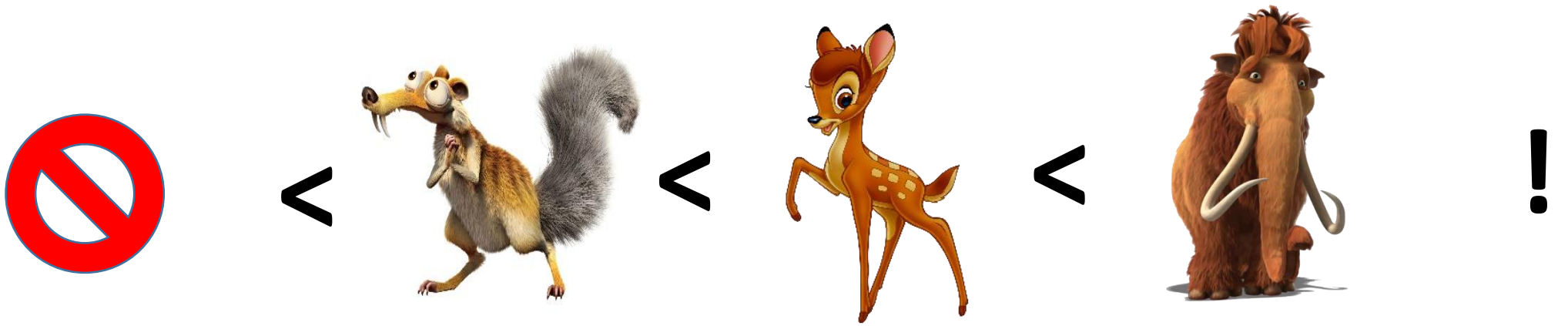
N = 2 (20%)

N = 3 (30%)

Jsou kategoriální data dostatečná za všech okolností?

Jsou kategorie seřaditelné?





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

Jsou kategoriální data dostatečná za všech okolností?

Pozor na medián u ordinálních dat

- Je medián vždy vhodným ukazatelem středu ordinálních dat?

Medián?

Vs.

Medián?

Pozor na medián u ordinálních dat







Vs.

Medián je shodný !







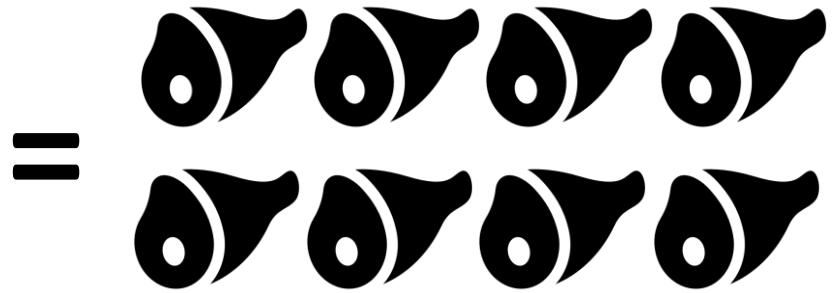
- Medián je shodný, nicméně interpretace dat je odlišná
- Možnost a formální správnost výpočtu statistiky neznamená, že jde o vhodnou metodu.

Kvantitativní data – jaký je objem kořisti ?

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





Typy dat: shrnutí

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

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

Kvalitativní data lze dělit dále

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

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

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

Kvalitativní data – příklady

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

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

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

Jak vznikají informace – popis různých typů dat

Statistika středu

Data poměrová

Data intervalová

Data ordinální

Data nominální

Data binární

PRŮMĚR

MEDIÁN

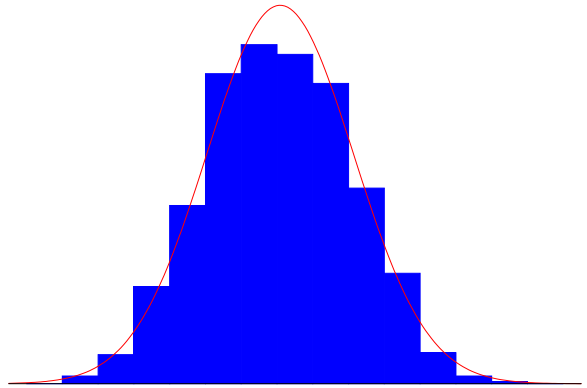
MODUS

Absolutní a relativní četnosti

Spojitá data

Diskrétní data

- Kvantitativní data - četnost hodnot rozložení v jednotlivých intervalech.

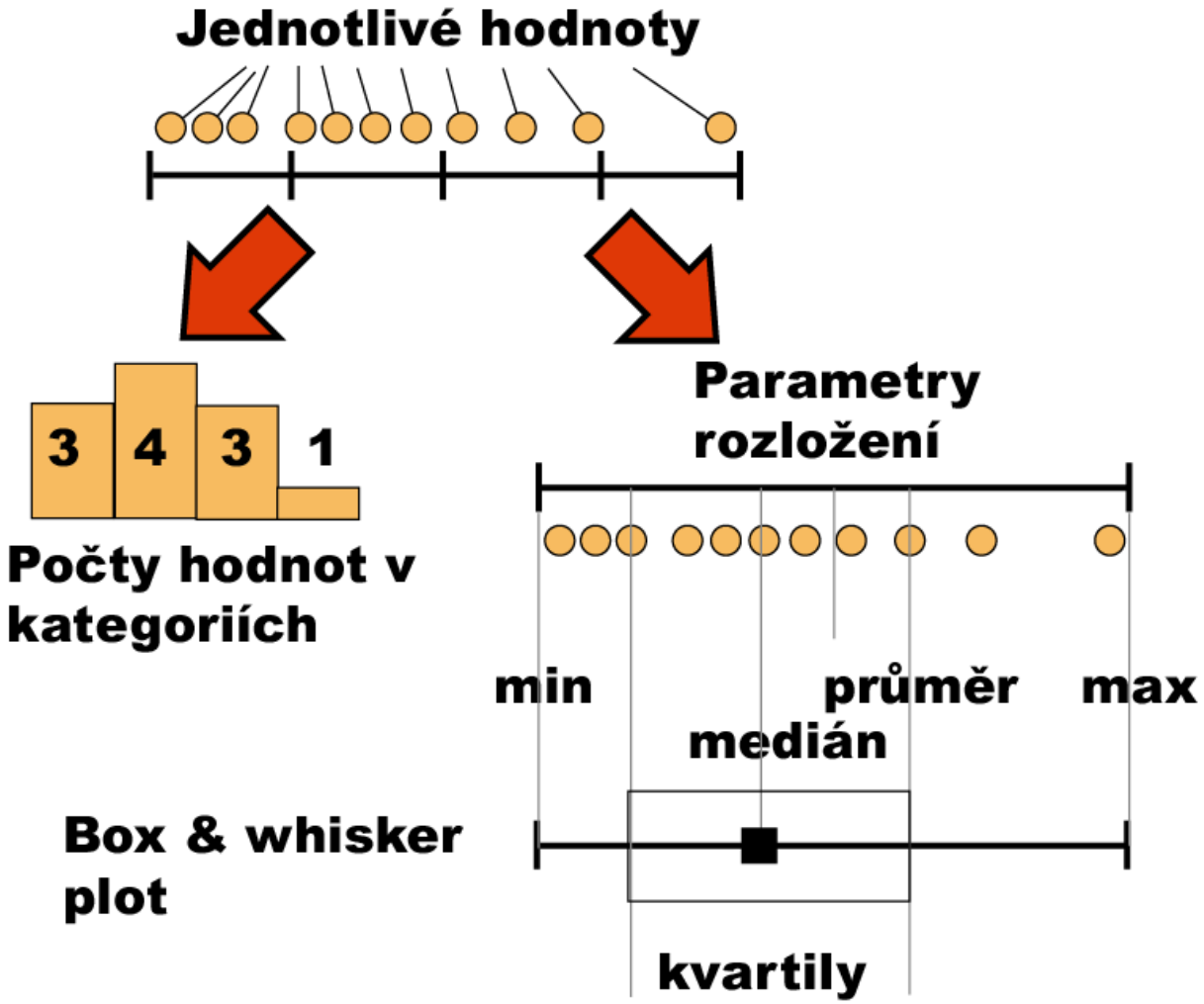


- Kvalitativní data - tabulka s četností jednotlivých kategorií.

Kategorie	Četnost
B	5
C	8
D	1

Řada dat a její vlastnosti

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



Odvozená data: pozor na odvozené indexy

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

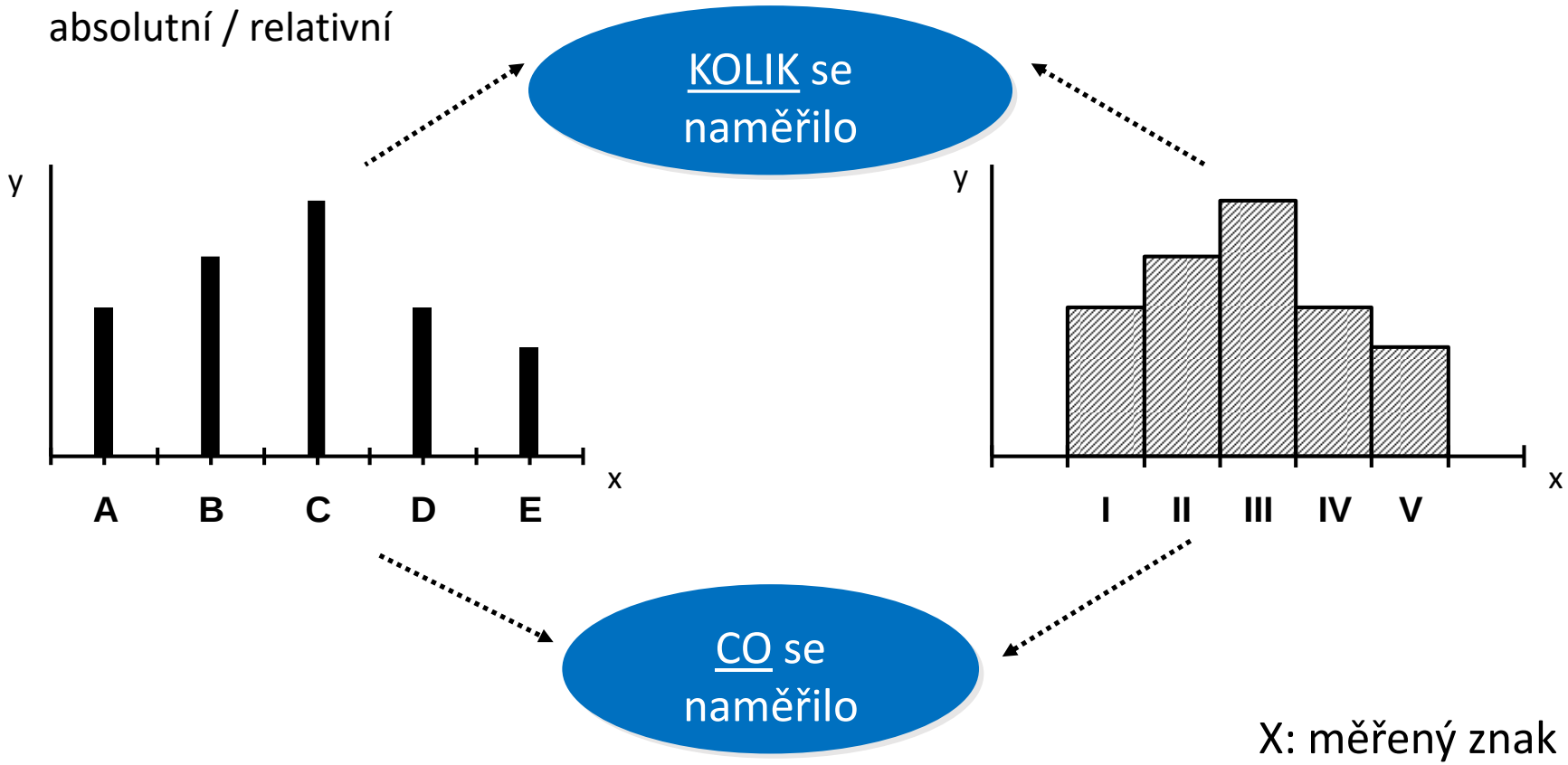
- Popsáno průměrem a rozsahem min-max
  - X: 1,2 : (1,15 - 1,24) → + / - 3,8 %
  - Y: 1,8 : (1,75 - 1,84) → + / - 2,5 %

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

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

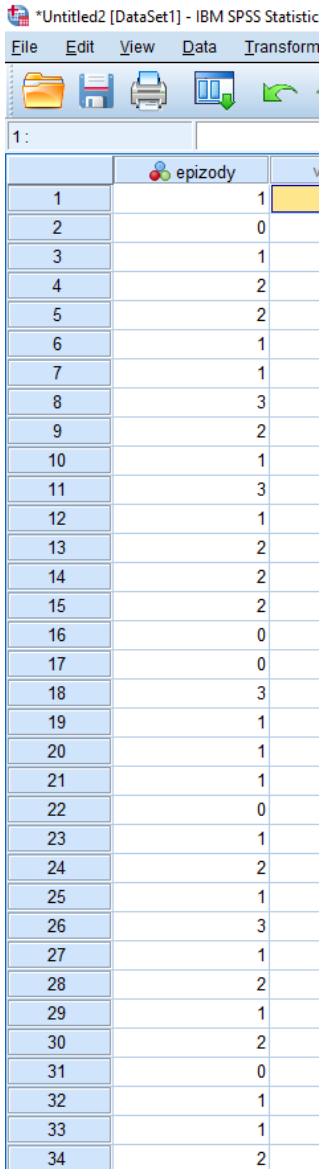
Vznik informací: opakovaná měření informují rozložením hodnot

Y: frekvence -
absolutní / relativní



Frekvenční sumarizace - základní nástroj popisu dat: kvalitativní data

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



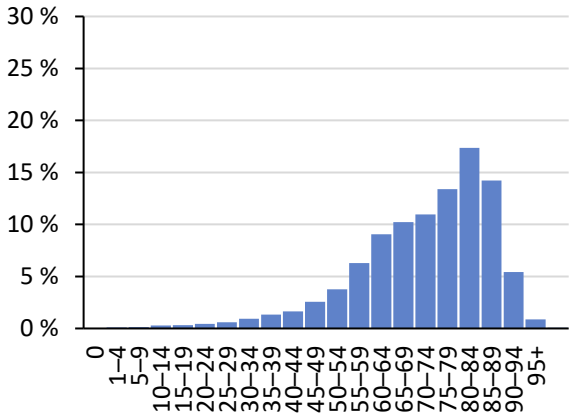
epizody

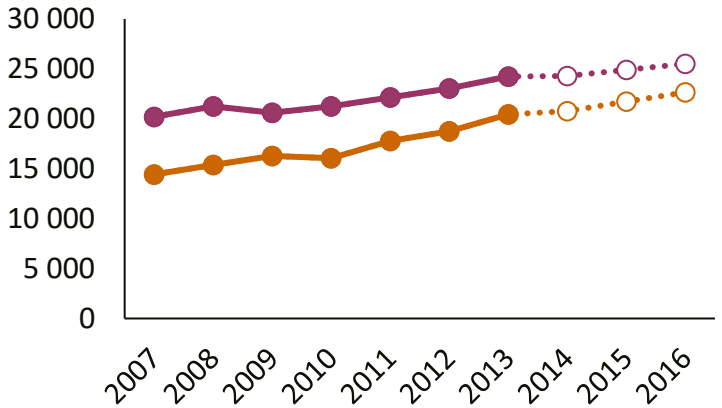
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	22	22,0	22,0	22,0
	1	27	27,0	27,0	49,0
	2	29	29,0	29,0	78,0
	3	22	22,0	22,0	100,0
	Total	100	100,0	100,0	

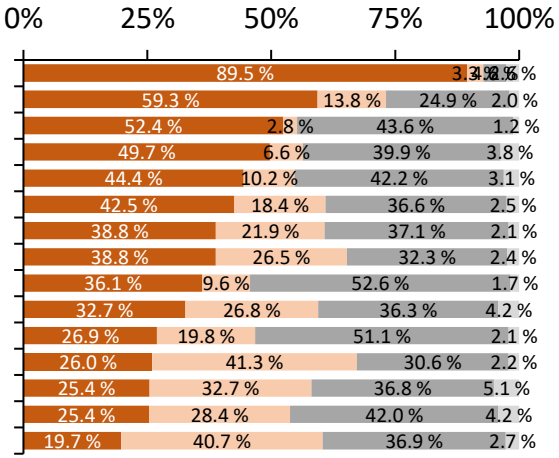
- Tabulka ukazuje unikátní hodnoty v datech
- Frequency = počet hodnot v kategorii (absolutní četnost)
- Percent = procentuální zastoupení kategorie (relativní četnost)
- Valid percent = procentuální zastoupení kategorie (bez započtení chybějících hodnot)
- Cumulative percent = kumulativní procentuální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

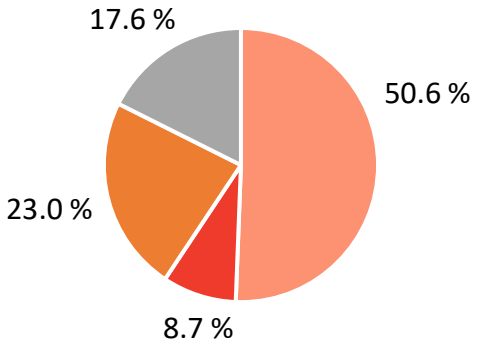
Vizualizace frekvenční tabulky kvalitativních dat

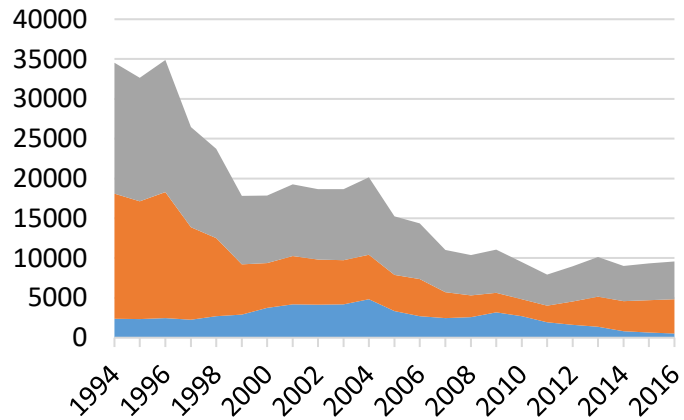
- Libovolné grafy umožňující vizualizaci počtů a procent (koláčový, páskový, sloupcový, čárový)

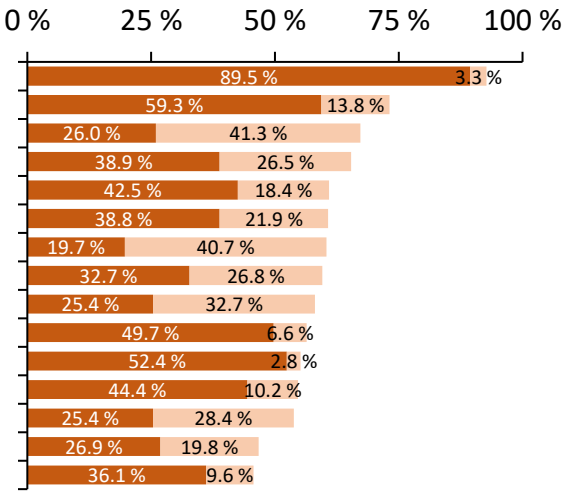






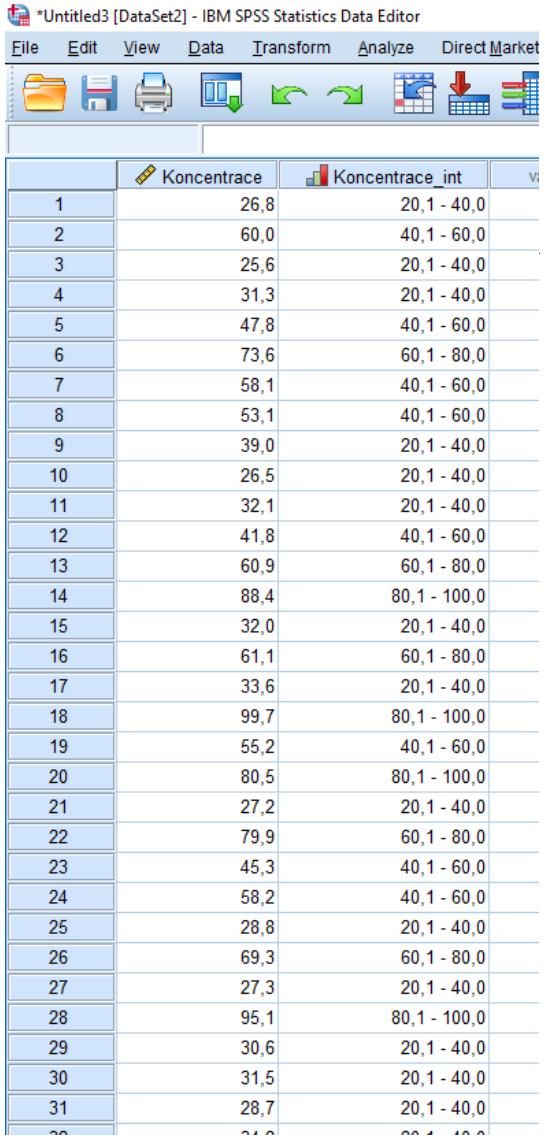






Frekvenční sumarizace - základní nástroj popisu dat: kvantitativní data

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



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

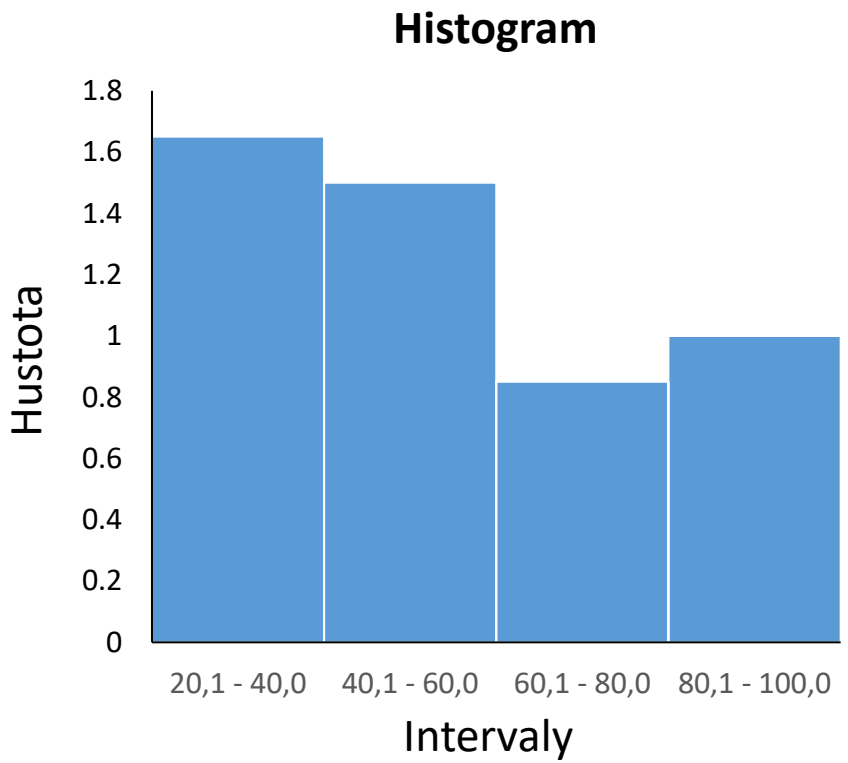
Koncentrace intervaly

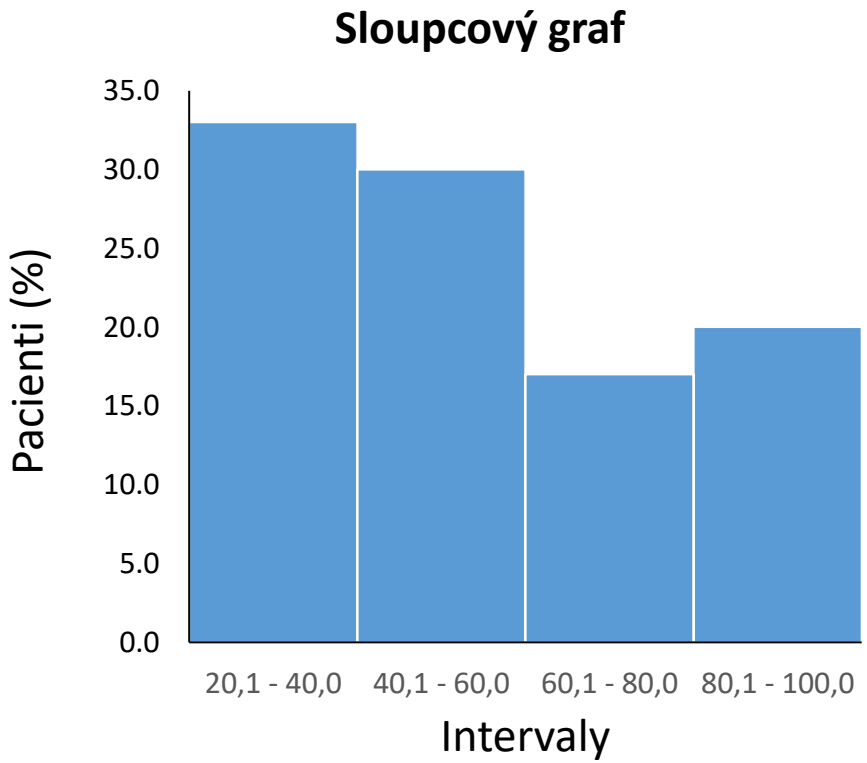
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	20,1 - 40,0	33	33,0	33,0	33,0
	40,1 - 60,0	30	30,0	30,0	63,0
	60,1 - 80,0	17	17,0	17,0	80,0
	80,1 - 100,0	20	20,0	20,0	100,0
	Total	100	100,0	100,0	

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

Vizualizace frekvenční tabulky kvantitativních dat

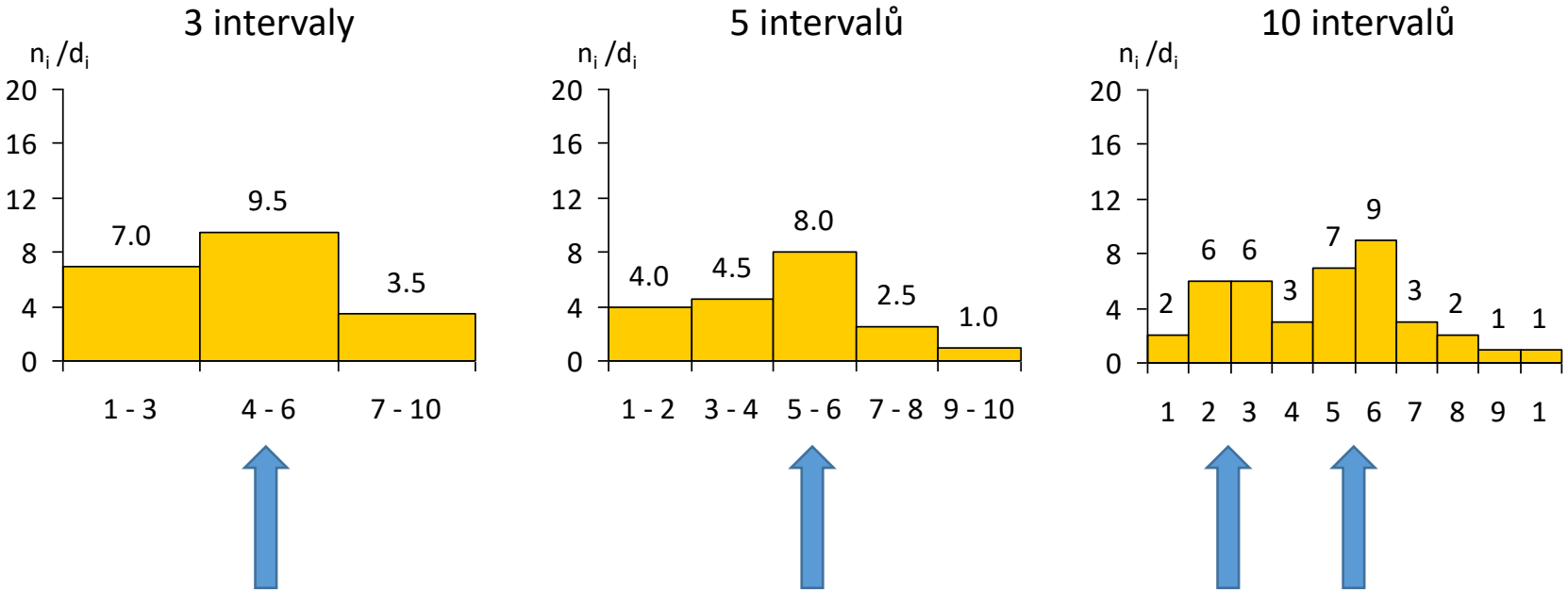
- Základním nástrojem vizualizace spojitých dat založeným na frekvenční tabulce je histogram
- Na rozdíl od sloupcového grafu představuje vizualizovanou hodnotu plocha sloupce, nikoliv jeho výška





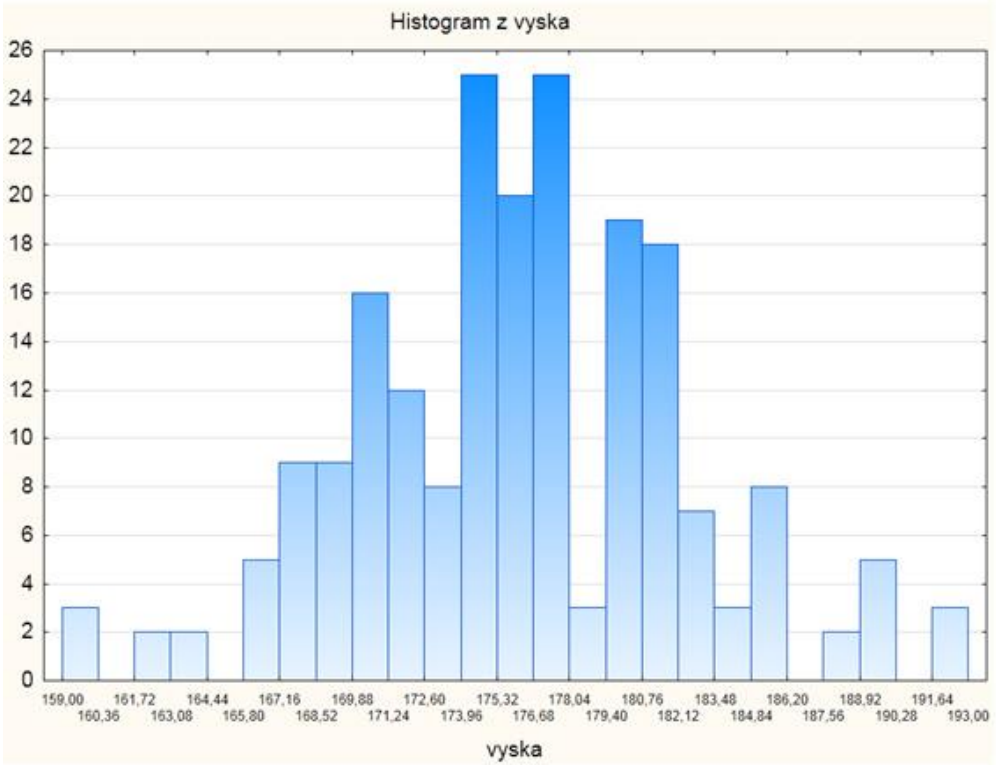
Histogram: vliv kategorizace dat

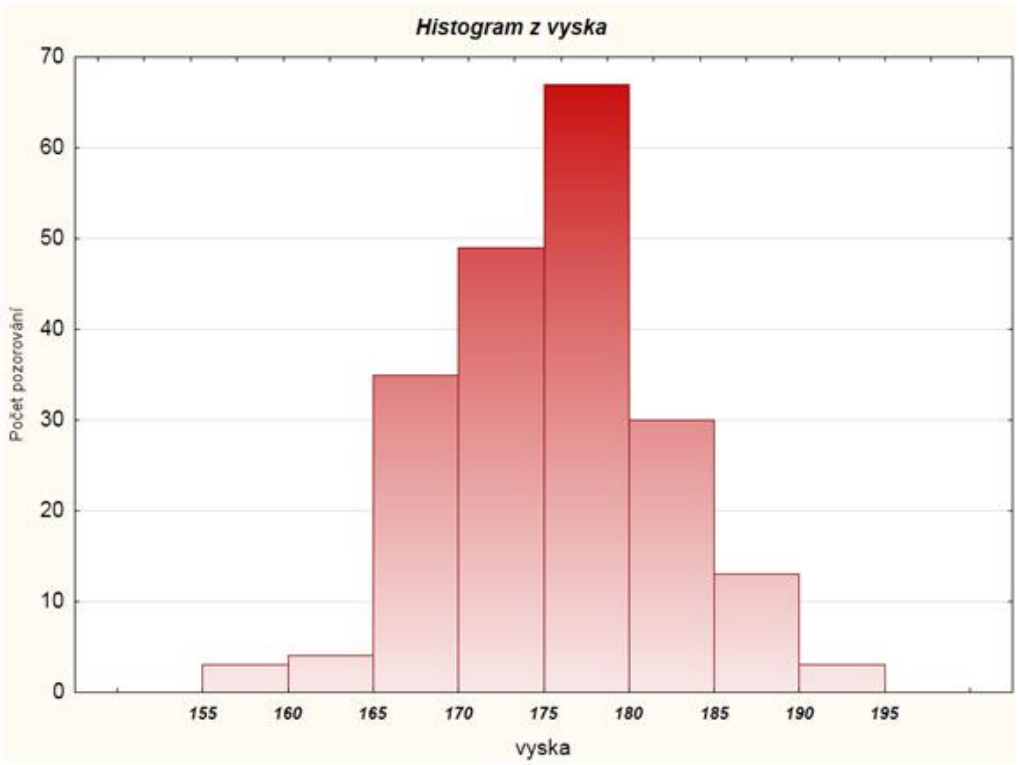
- Počtem zvolených intervalů v histogramu rozhodujeme o tom, jak bude vypadat. Při malém počtu můžeme přehlédnout důležité prvky v datech, při velkém zase může být informace roztříštěná.



Histogram: vliv kategorizace dat

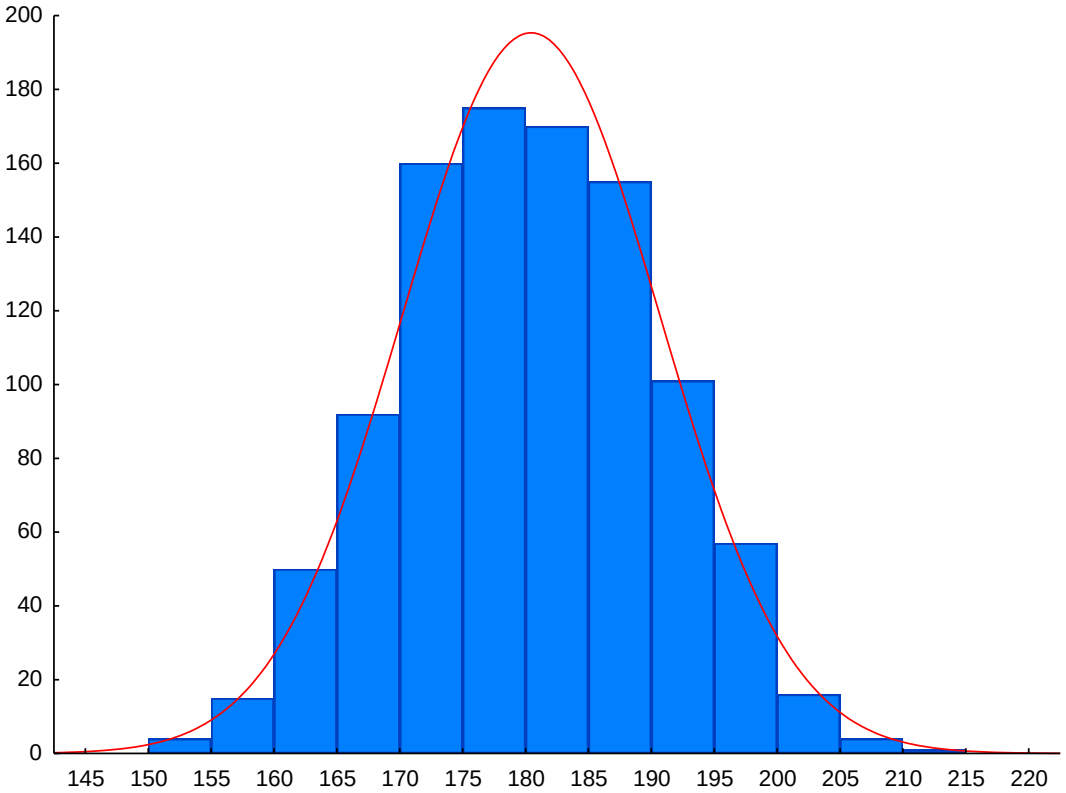
- Výběr počtu kategorií – důležitý pro interpretaci
- Ruční nebo automatický výběr – různé algoritmy (závisí na velikosti vzorku a variabilitě dat)



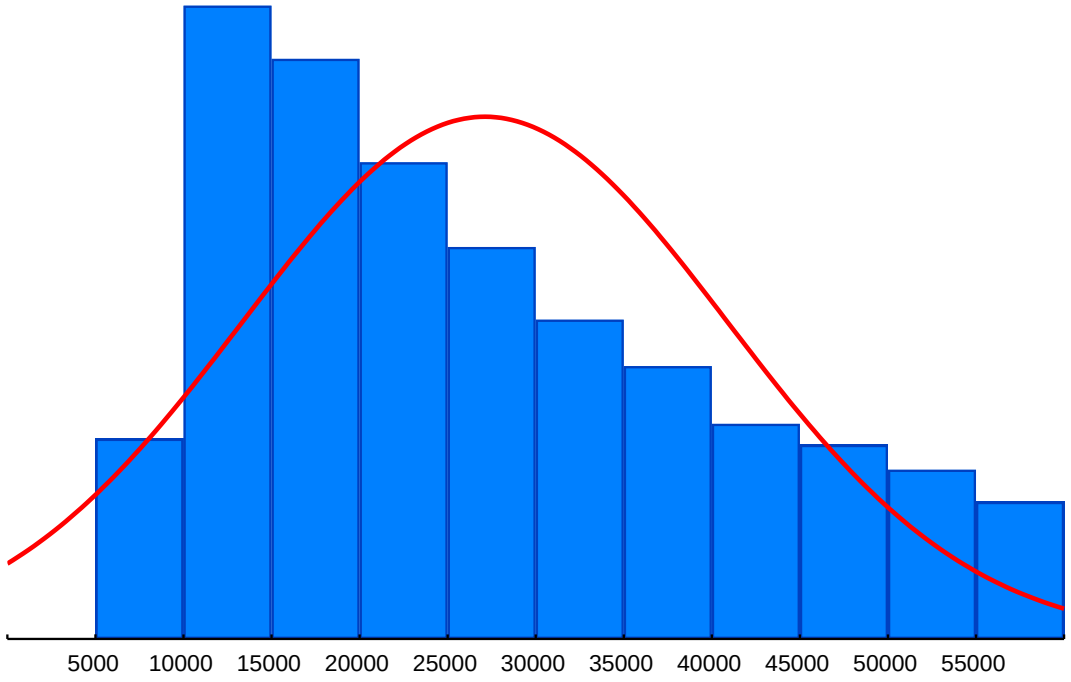


Histogram: nástroj posouzení rozložení dat

- Histogram reálných dat má vazbu na modelové rozdělení



?

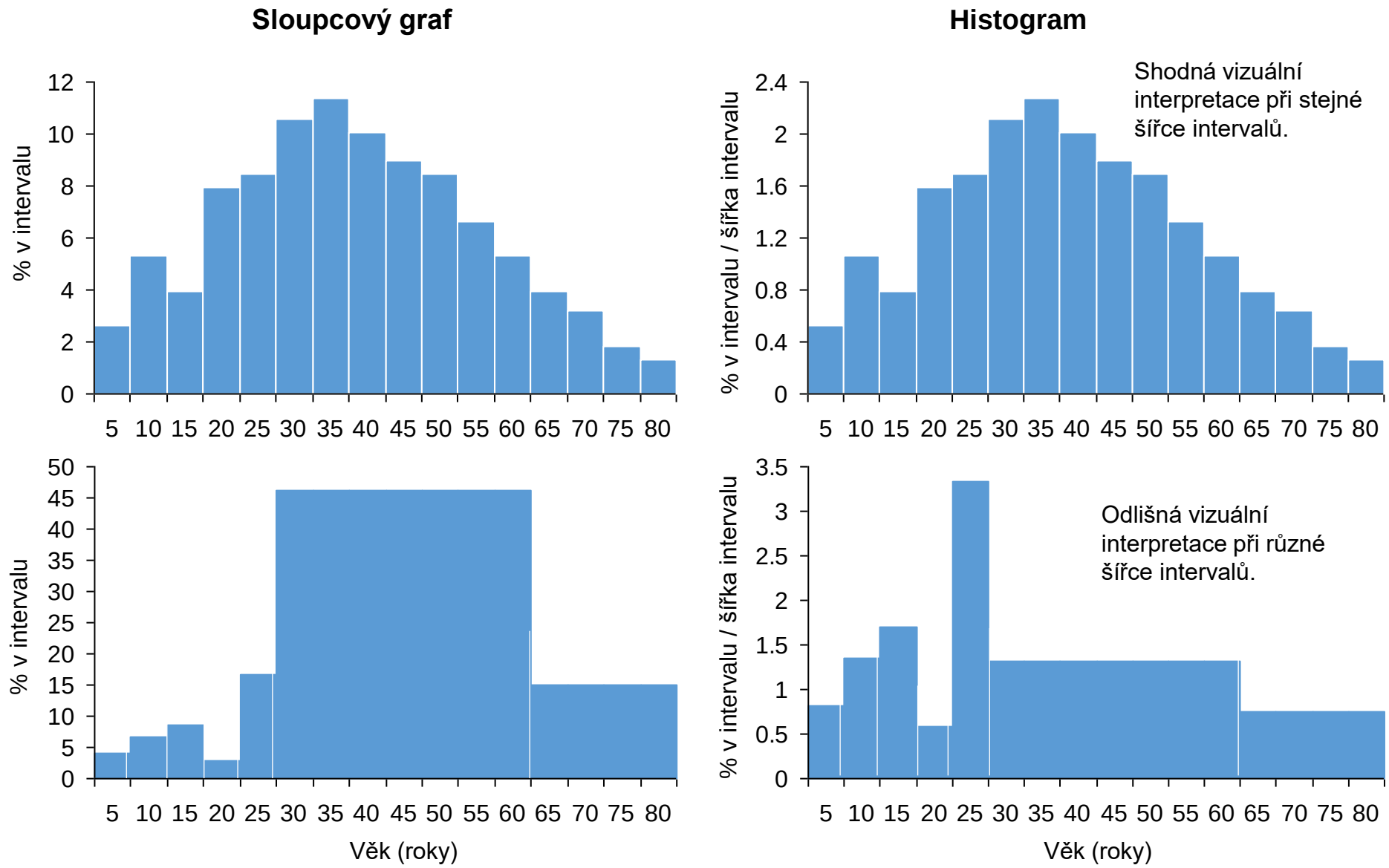


Proč je důležité vědět co je to skutečný histogram I

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

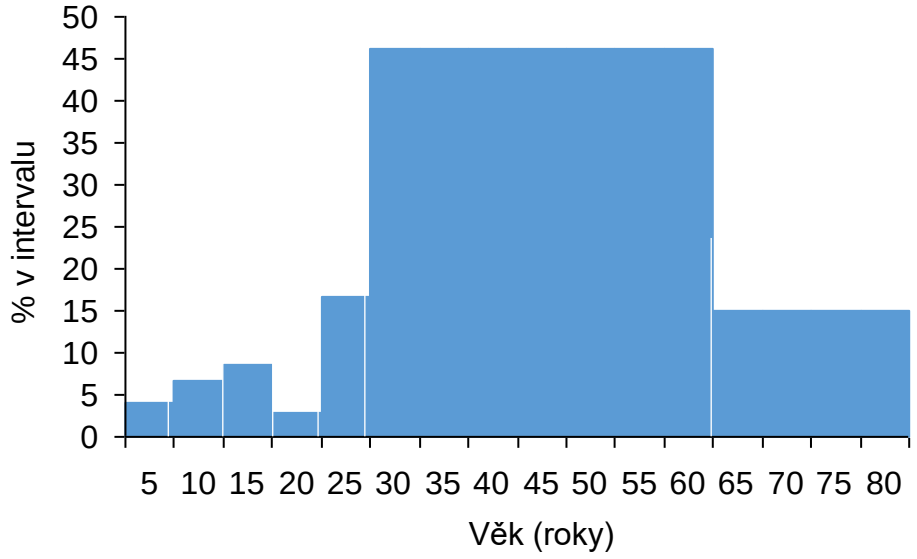
- V praxi se nicméně často používá namísto „pravého“ histogramu sloupcový graf (i výrobci statistických SW)
- V případě stejné šířky intervalů interpretační problém nevzniká (při různé šířce intervalu vypínají SW některé volby = nastavení pro pokročilé uživatele)

Histogram a sloupcový graf

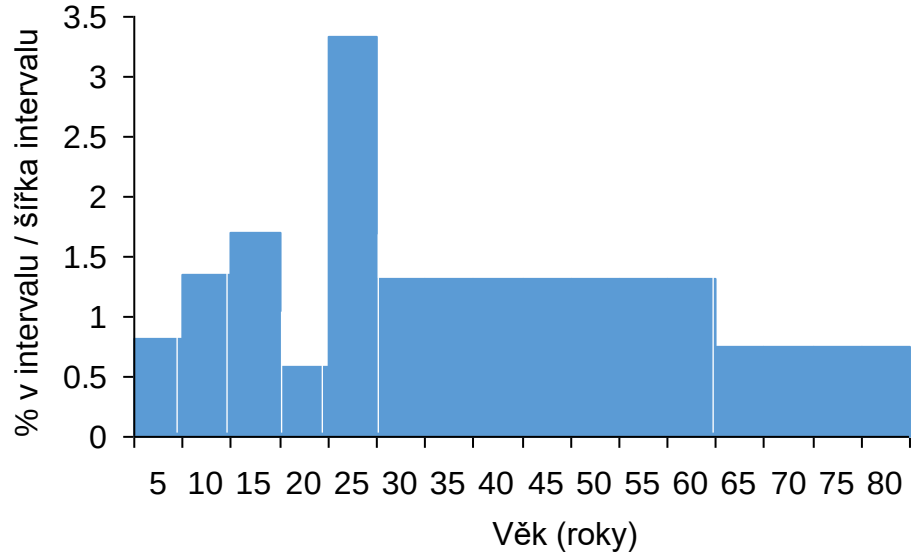


Příklad: věk účastníků vážných dopravních nehod

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



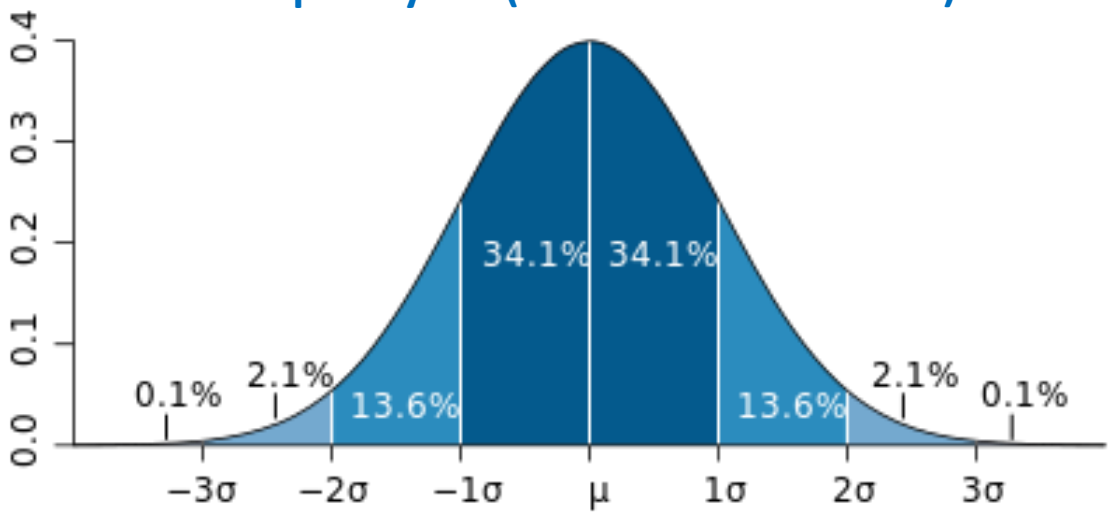
Věk	N	%
0 - 4	28	4,1%
5 - 9	46	6,7%
10-15	58	8,5%
16 - 19	20	2,9%
20 - 24	114	16,6%
25 - 59	316	46,1%
> 60	103	15,0%

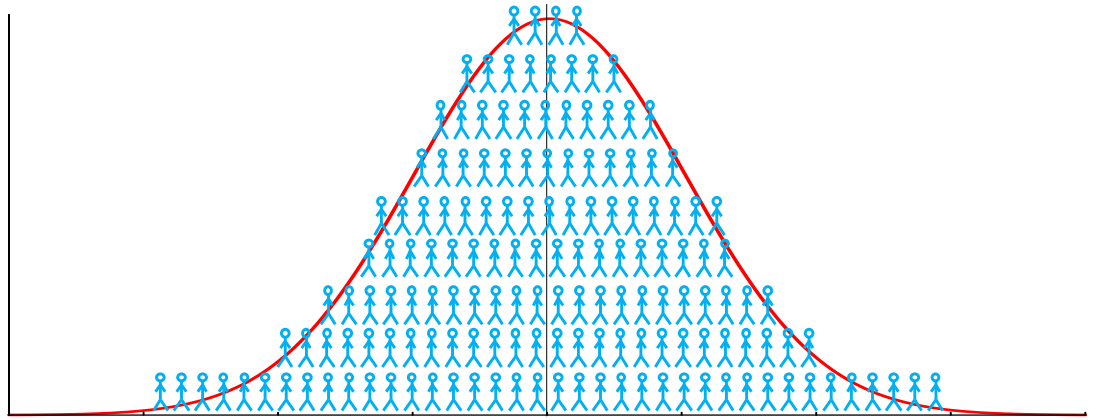


Proč je důležité vědět co je to skutečný histogram II

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

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





Příklad: optimalizace skladových zásob oblečení

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

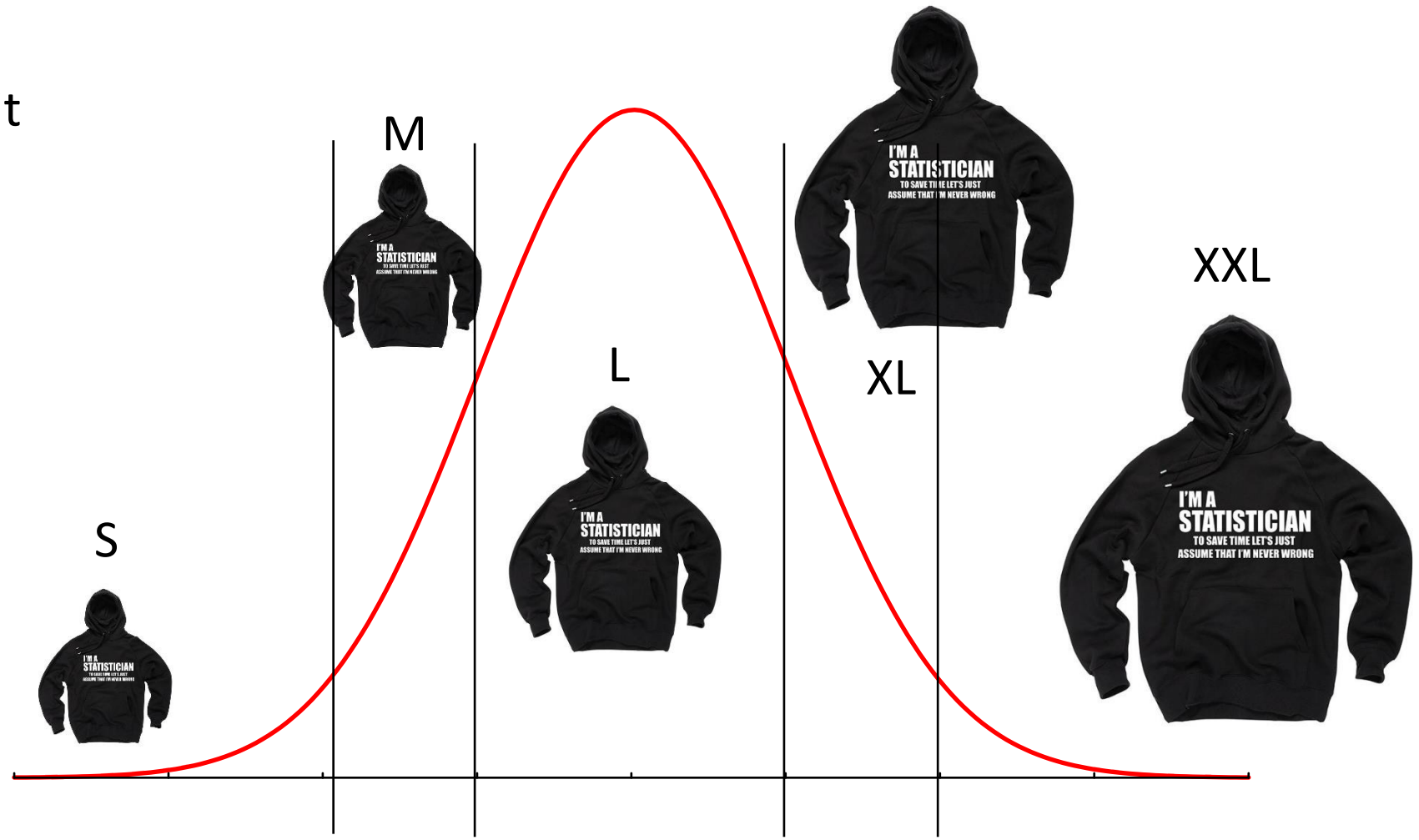
S M L XL XXL

Příklad: optimalizace skladových zásob oblečení

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

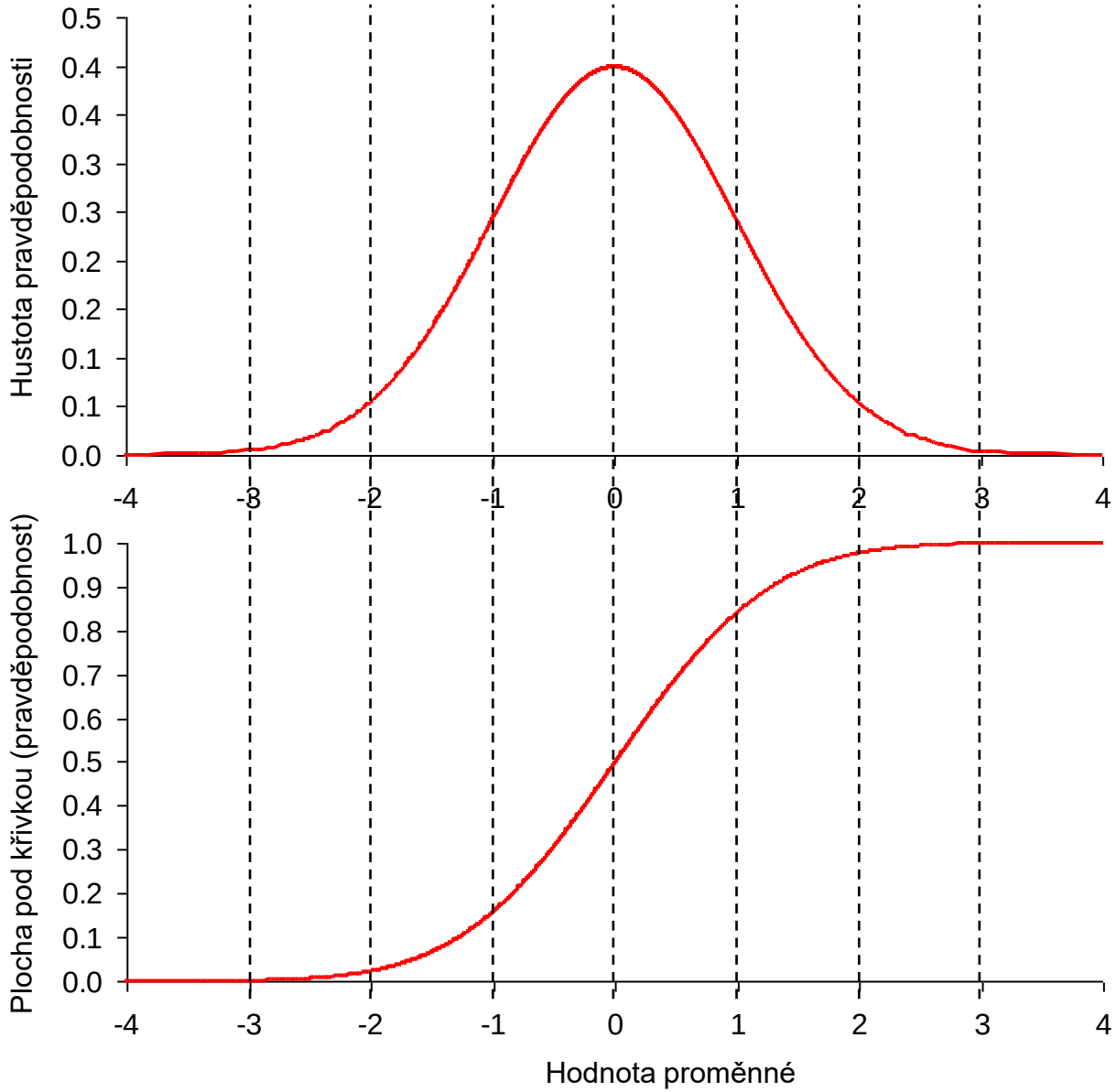
- Integrovat?

- Lze jednodušeji?



Normální rozdělení a jeho distribuční funkce

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

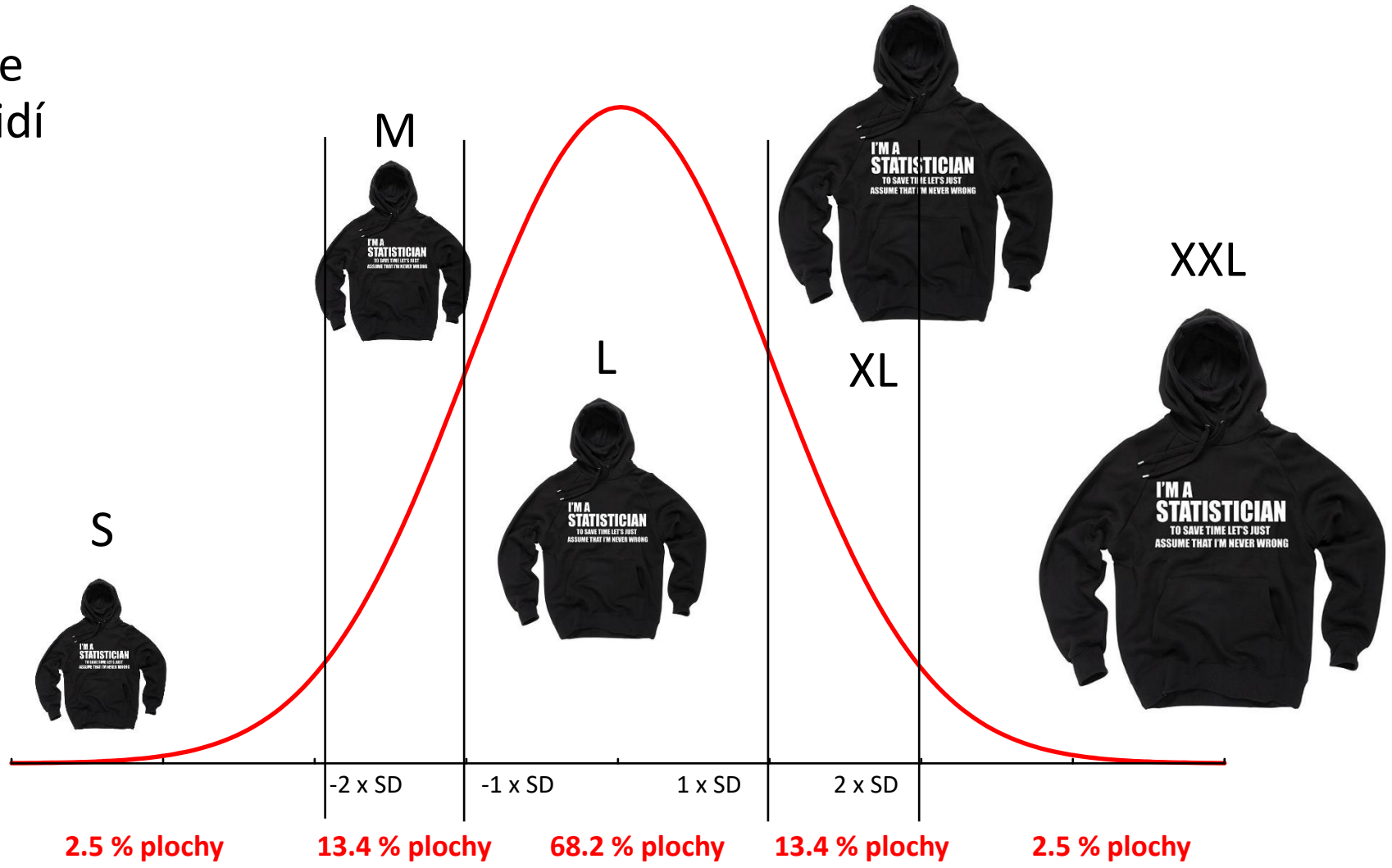


Normální rozdělení

Distribuční funkce normálního rozdělení

Příklad: optimalizace skladových zásob oblečení

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



Velikost člověka relevantní k velikosti oblečení