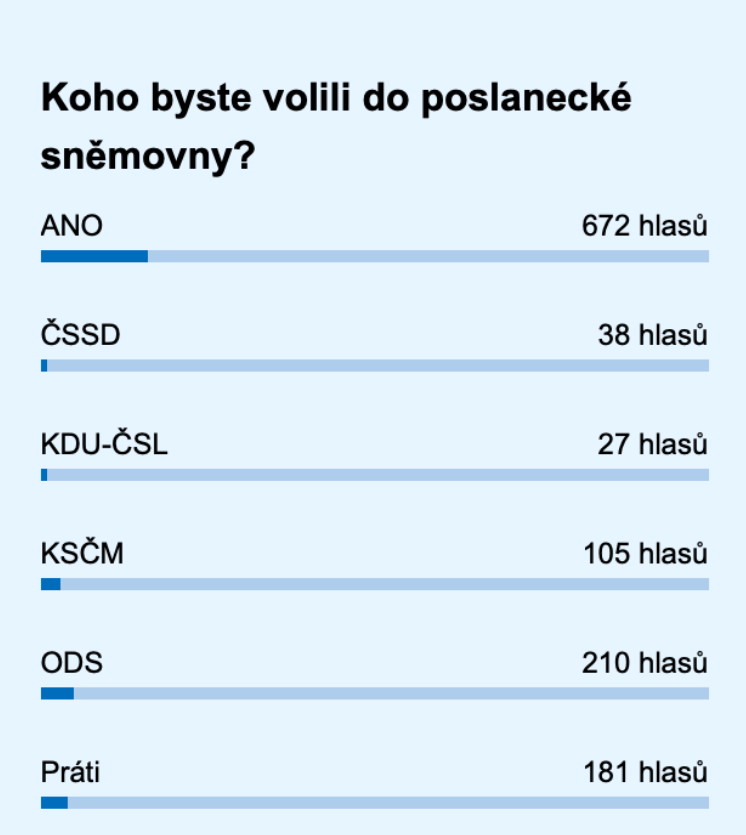
Statistické usuzování

ANALÝZA DAT

18. BŘEZNA 2021

Statistické usuzování v médiích



tn.nova.cz

facebook.com

Pravděpodobnost – terminologie

- experiment – zopakovatelný postup pro vytvoření pozorování (např. hod mincí)
- jev – možný výsledek experimentu (panna, nebo orel)
- množina jevů – množina všech možných jevů ($\Omega = \{panna, orel\}$)
- pravděpodobnost jevu je jeho dlouhodobá relativní frekvence; udává se buď v intervalu <0,1> (matematický zápis) nebo v procentech mezi 0 % a 100 % (intuitivní zápis)

- jestliže P(panna) = 0,5, tak se tento jev bude objevovat přibližně při polovině experimentů, které jsou opakovány donekonečna (https://www.geogebra.org/m/KkqY94aZ)
- jestliže je experiment opakován konečným počtem pokusů, tak se přibližný odhad pravděpodobnosti bude vylepšovat s rostoucím počtem experimentů

Koncept pravděpodobnosti

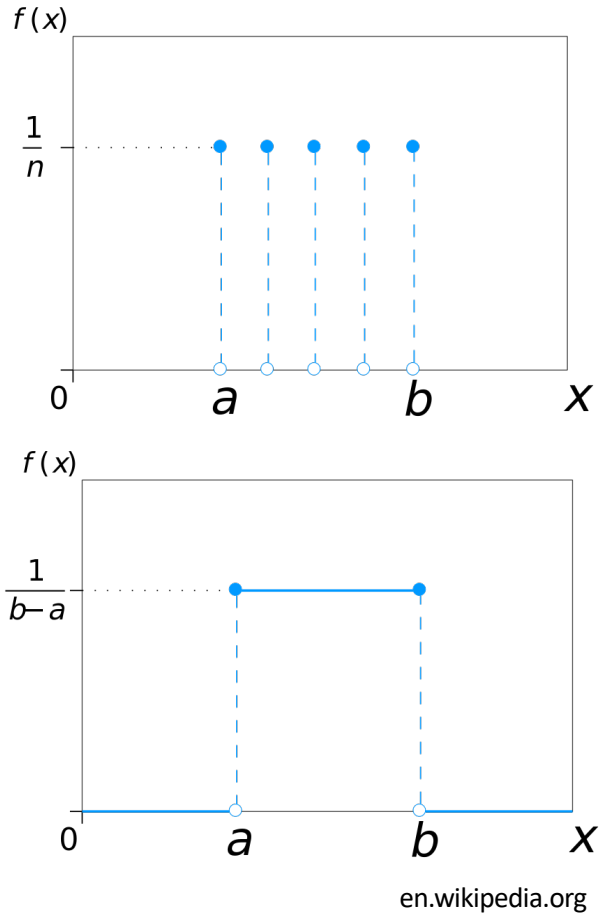
- pravděpodobnost jevu A označíme P(A) a definujeme ji jako
  - $P(A) = \frac{\text{počet příznivých jevů A}}{\text{počet možných událostí}} = \frac{|A|}{|\Omega|}$
- intuitivně využíváme pravděpodobnost jako připisování reálných čísel ke každému jevu způsobem, aby byl součet těchto čísel roven číslu 1
- příklad: 3 lidé volí stranu A, 2 lidé stranu B a 5 lidí stranu C
  - tedy P(A) = 0,3; P(B) = 0,2; P(C) = 0,5

- u pravděpodobností platí
  a) $P(A) \geq 0$ (pravděpodobnosti nejsou negativní)
  b) $P(\Omega) = 1$ (celková pravděpodobnost všech jevů je rovna číslu 1)
  c) jestliže jevy A_1, A_2,…, A_k jsou vzájemně výlučné, tak sjednocením pravděpodobností je jejich součet

$$P(A_1 \cup A_2 \dots \cup A_k) = P(A_1) + P(A_2) \dots P(A_k)$$

Rovnoměrné rozdělení hodnot (*uniform*)

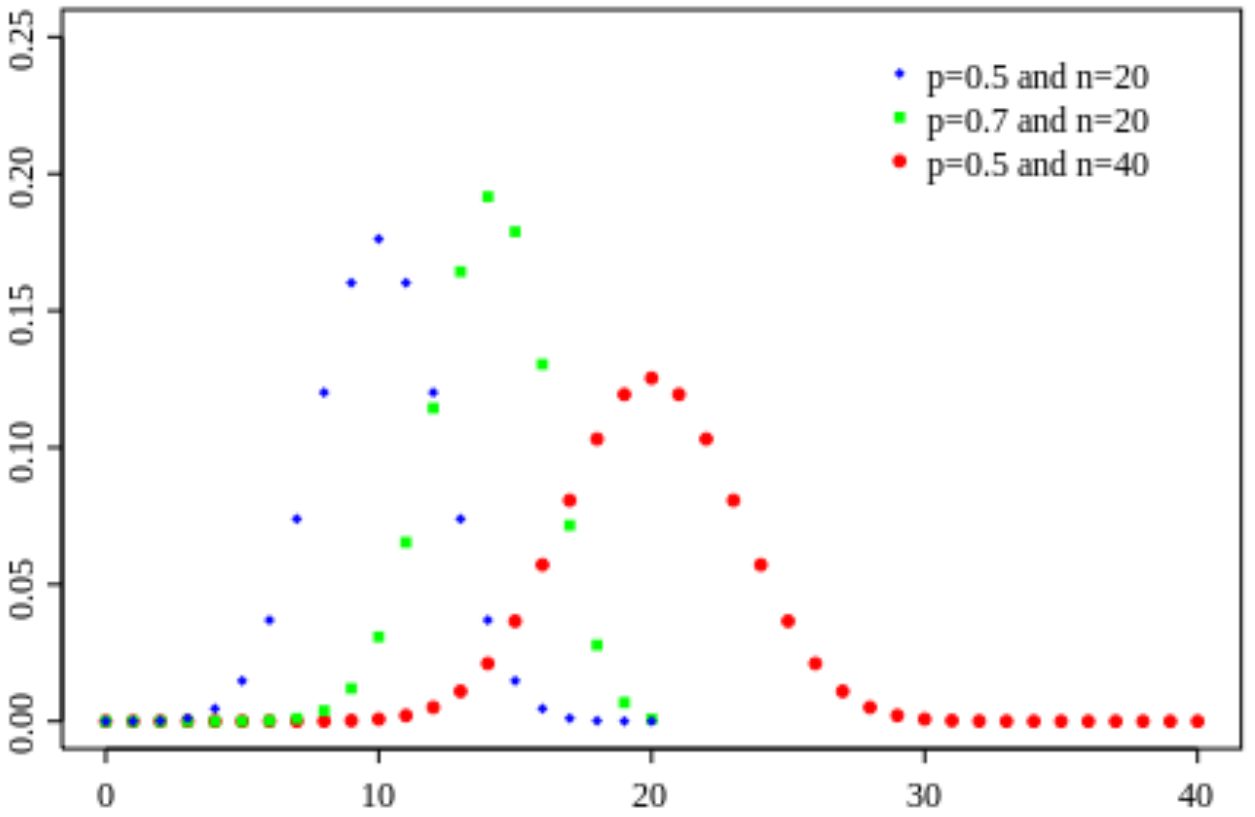
- přiřazuje všem hodnotám veličiny stejnou pravděpodobnost
- může mít nespojitou i spojitou podobu
- nespojité rovnoměrné rozložení popisuje veličinu, která může nabývat n hodnot se stejnou pravděpodobností (příkladem hod kostkou)
- spojité rovnoměrné rozložení přiřazuje všem hodnotám veličiny v daném intervalu stejnou pravděpodobnost



en.wikipedia.org

Binomické rozdělení

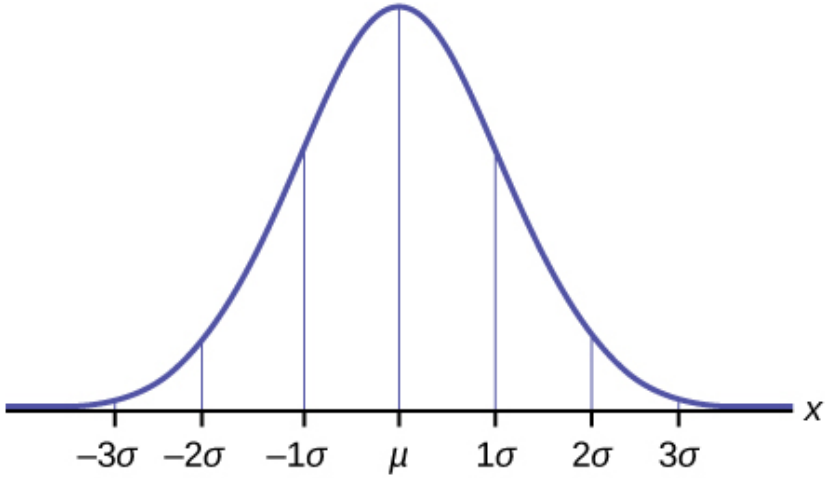
- nespojitá distribuce pravděpodobností úspěchu na sobě nezávislých pozorování s dichotomickým výsledkem (úspěch vs. neúspěch)
- v jednom pozorování může být i více experimentů (např. více hodů mincí)
- Bernoulliho schéma – v jednom pozorování jeden experiment
- určujícími parametry jsou počet pokusů (n) a pravděpodobnost úspěchu (p)



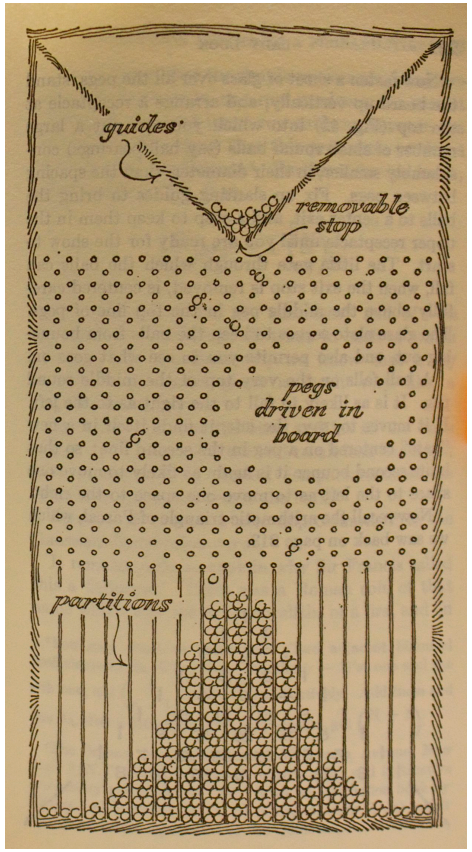
en.wikipedia.org

Normální rozložení

- kontinuální distribuce dat, která znázorňuje data shromážděná okolo průměru
- unikátně určena svým průměrem/mediánem/modem μ a rozptylem σ^2
- normální rozložení je důležité kvůli centrálnímu limitnímu teorému (viz dále)



expii.com

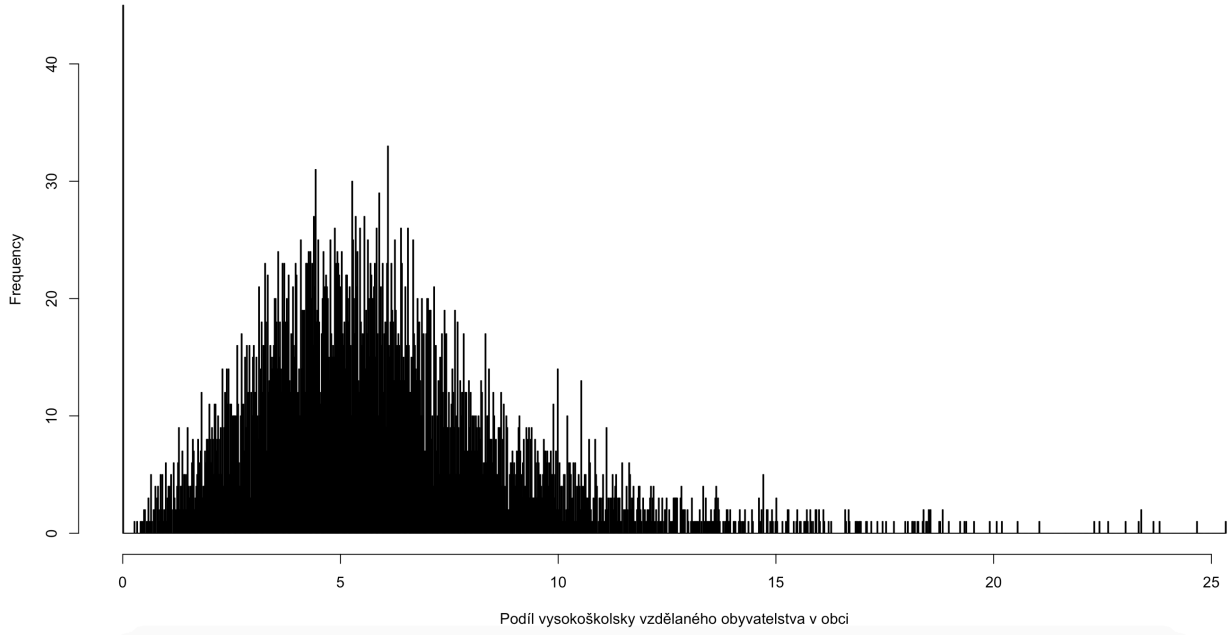


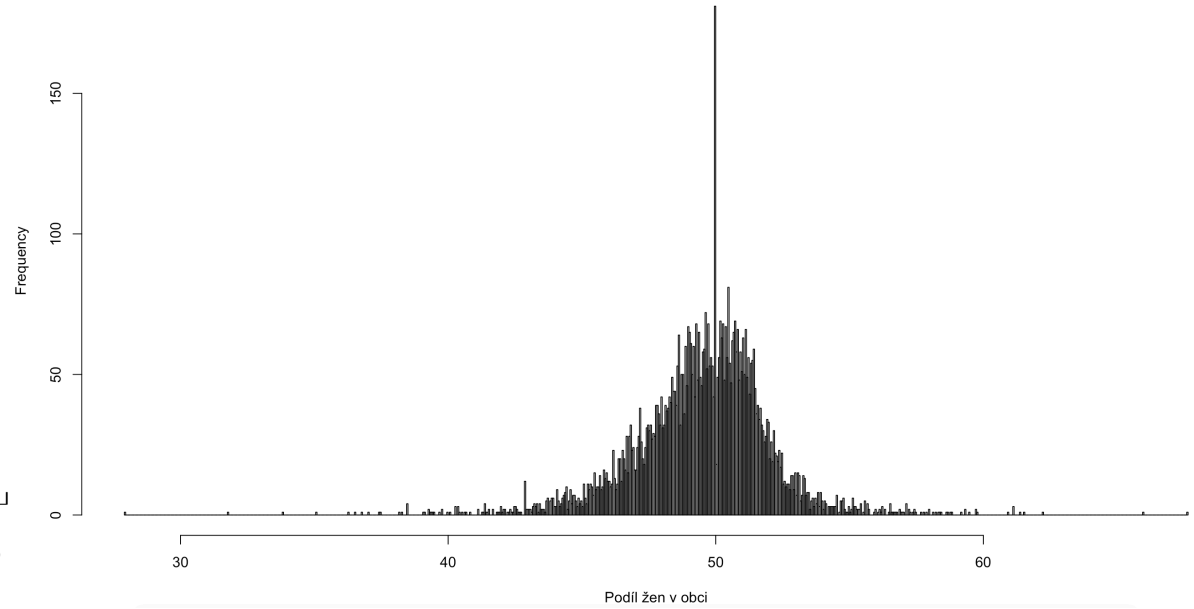
quadrantgroup.co.uk

Galtonova deska

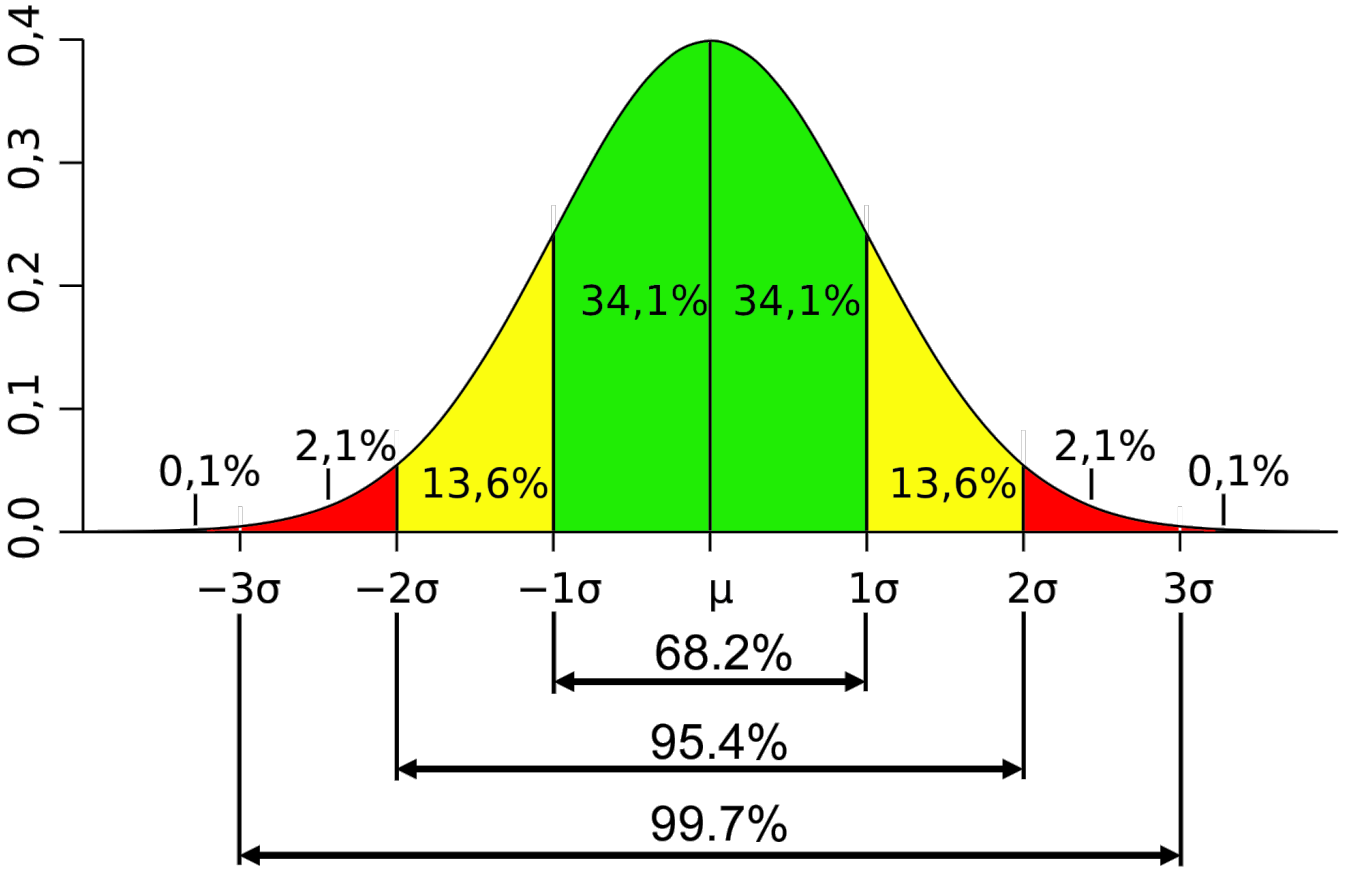
Normální rozložení kolem nás

- IQ, výška lidí, výkon v testu, míra nezaměstnanosti v obcích a mnoho dalších





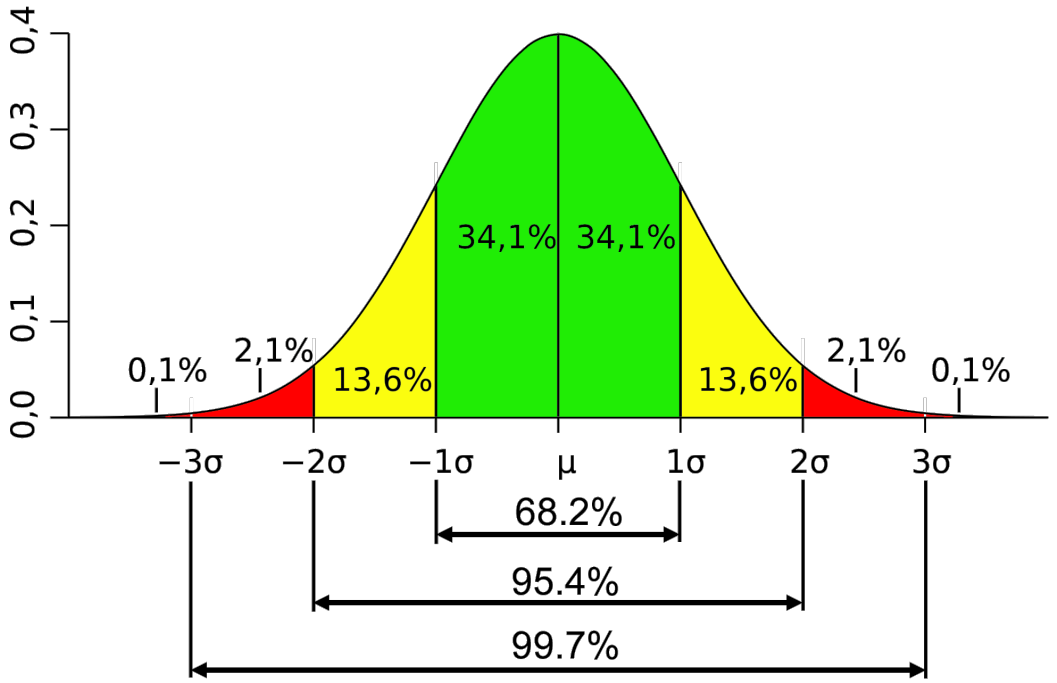
Přínosy normálního rozložení



kanbanize.com

Přínosy normálního rozložení

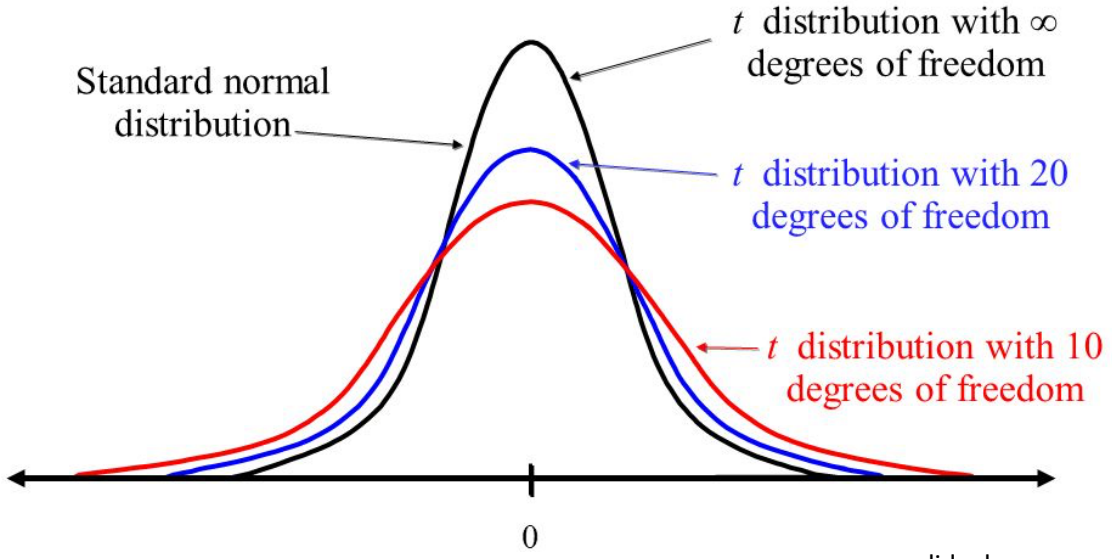
- výhodou je, že máme popsáno rozložení dat
- víme, s jakou pravděpodobností nastanou určité jevy (objeví se určité hodnoty)
- jde o to, s jakou pravděpodobností může pozorovaný vzorek (s určitým průměrem) pocházet z nepozorované populace



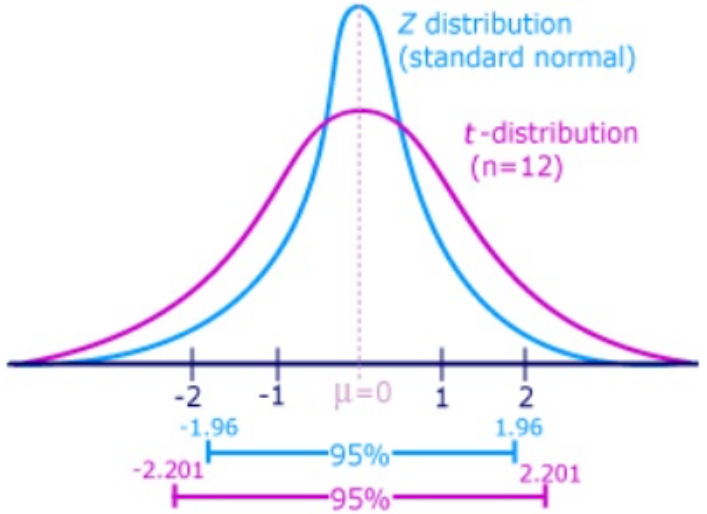
kanbanize.com

t distribuce

- je velmi podobná normální distribuci, ale zohledňuje velikost vzorku (používáme pro práci s malými vzorky dat)
- stupně volnosti (degrees of freedom) jsou počet pozorování, která mohou variovat aniž by byla změněna hodnota průměru (df = n – 1)



slideplayer.com



slideshare.net

Centrální limitní teorém

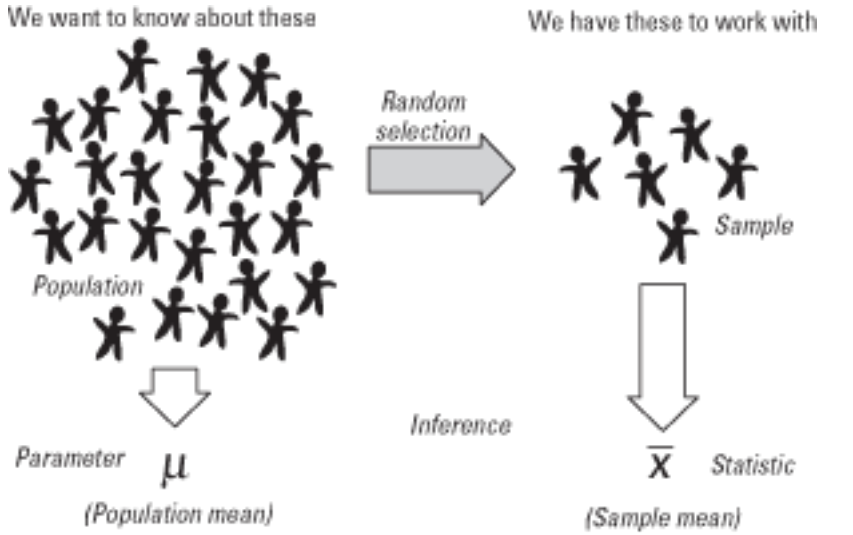
- existuje populace s průměrem μ a rozptylem σ^2
- z této populace vybereme vzorek o velikosti *n*
- to uděláme opakovaně a získáme tak výběrový vzorek průměru
- distribuce tohoto výběrové vzorku se blíží normálnímu rozložení s průměrem μ a rozptylem σ^2/n, jak se velikost vybíraného vzorku zvětšuje
- **toto platí nehledě na to, jaké je původní rozložení dat!**
- tento jev je základem řady statistických postupů

- praktická ukázka:

http://students.brown.edu/seeing-theory/probability-distributions/index.html

Usuzování ze vzorku

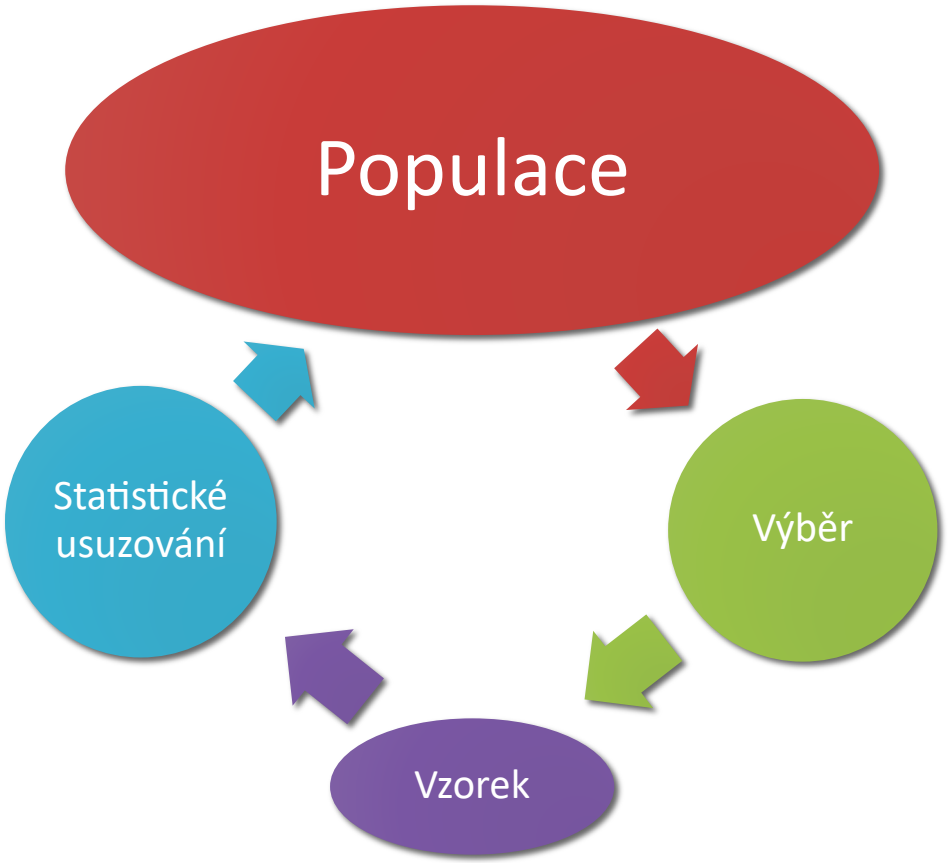
- důvodem je praktická nutnost
- nemáme dostatek prostředků (finančních, personálních, časových atd.) k tomu, abychom zkoumali celou populaci
- proto si vybereme pouze vzorek a pro něj shromáždíme data
- důležitá je metoda výběru (náhodný, kvótní apod. – viz zadaná literatura)
- klíčová je reprezentativnost



cliffsnotes.com

Logika usuzování

- vzorek je vždy horší než populace, ale nic jiného často nezbývá
- čím větší vzorek, tím jsme si u jednotlivých parametrů jistější v jejich hodnotě
- problémem je to, že v rámci výběru vždy existuje výběrová chyba
- průměr a standardní odchylka našeho vzorku se velmi vzácně rovná hodnotám v populaci
- navíc každý výběr bude trochu jiný v těchto hodnotách



Statistická chyba

- ze zmíněných důvodů počítáme statistickou chybu
- pro odhalení parametrů je pro nás většinou zásadní průměr určité hodnoty jako nejdůležitější centrální hodnota
- pro odhad průměru populace využíváme jako nejlepší dostupnou hodnotu průměr vzorku (nic lepšího nemáme)
- využíváme výhod centrálního limitního teorému (průměr a směrodatná odchylka opakovaného výběru populace mají normální rozložení, i když původní populace normálně rozložená není)
- to znamená, že pokud vezmeme z populace určitý vzorek, je velmi pravděpodobné, že průměr vzorku bude podobný průměru populace a méně pravděpodobné, že průměr vzorku bude výrazně odlišný od průměru populace (viz centrální limitní teorém)

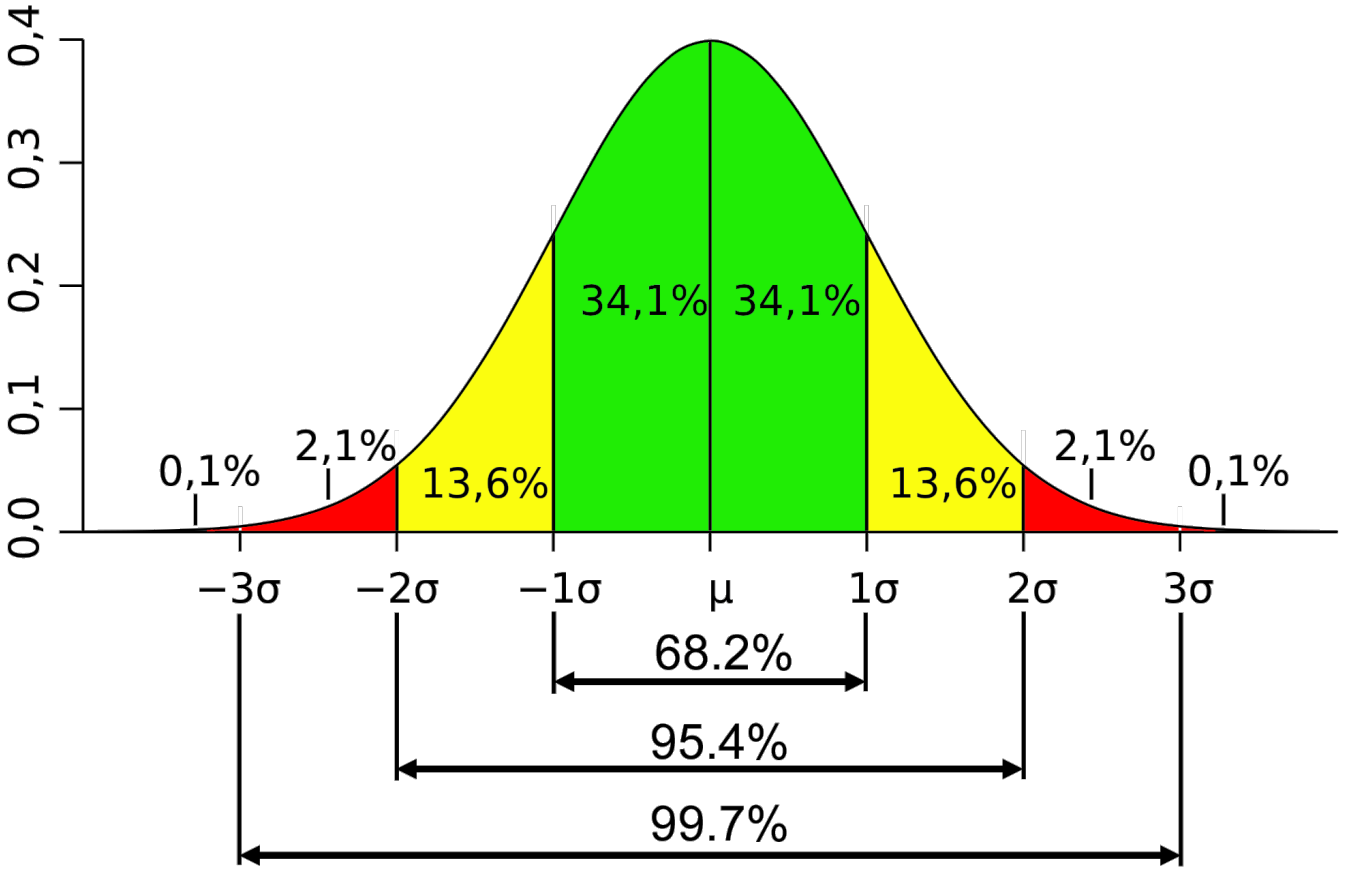
Statistická chyba

- jak ale tuto chybu spočítat?
- mohli bychom vzít více vzorků a popsat rozptyl jejich průměrů
  - např. vezmeme 10 vzorků a řekneme, že jejich průměry jsou od 12 do 34
- my ale máme z důvodu nedostatku prostředků jen jeden vzorek
- proto spočítáme standardní chybu průměru
- $SE = \frac{směrodatná\ odchylka\ vzorku}{\sqrt{velikost\ vzorku}} = \frac{s}{\sqrt{n}}$
- ze vzorce je zřejmé, že velikost standardní chyby je ovlivněna velikostí vzorku
  - čím větší je vzorek, tím menší je statistická chyba

Příklad

- v dotazníkovém šetření mezi 50 lidmi jsme získali informace o podpoře prezidenta (škála 0-100)
  - 35,67,45,23,66,45,58,89,34,78,61,65,34,85,62 atd.
- vzorek má průměr $\bar{x}$ = 60 a směrodatnou odchylku s = 10
- standardní chyba odhadu průměru celé populace: $SE(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{10}{\sqrt{50}} \cong 1{,}41$
- chceme mít jistotu 95 %, že jsme odhalili skutečný průměr populace
  - 95 % je běžná úroveň jistoty používaná v sociálních vědách (v poslední době se prosazuje použití 99 %)
- z normálního rozložení víme, že 95 % rozložení dat spadá mezi 1,96 směrodatné odchylky na jednu i druhou stranu od průměru

Úrovně jistoty



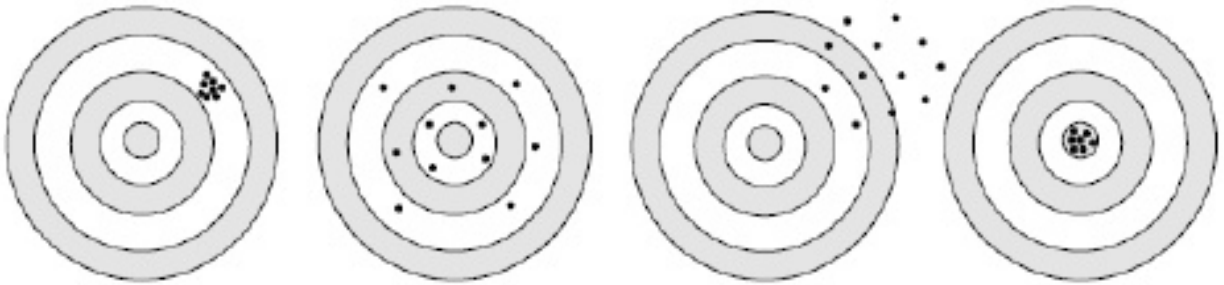
kanbanize.com

Příklad

- v dotazníkovém šetření mezi 50 lidmi jsme získali informace o podpoře prezidenta (škála 0-100)
  - 35,67,45,23,66,45,58,89,34,78,61,65,34,85,62 atd.
- vzorek má průměr $\bar{x}$ = 60 a směrodatnou odchylku s = 10
- standardní chyba odhadu průměru celé populace: $SE(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{10}{\sqrt{50}} \cong 1{,}41$
- chceme mít jistotu 95 %, že jsme odhalili skutečný průměr populace
  - 95 % je běžná úroveň jistoty používaná v sociálních vědách (v poslední době se prosazuje použití 99 %)
- z normálního rozložení víme, že 95 % rozložení dat spadá mezi 1,96 směrodatné odchylky na jednu i druhou stranu od průměru
- $60 \pm 1{,}96 * 1{,}41 \cong [57{,}24; 62{,}76]$
  - interval spolehlivosti obsahuje skutečný parametr nejméně v 95 % případů, ve kterých zopakujeme výběr
  - jde o tzv. interval spolehlivosti (*confidence interval*) - http://rpsychologist.com/d3/CI/

Statistická inference

- usuzování o celé populaci na základě vzorku dat
- aby mohlo proběhnout statistické uvažování, musí být výběr dat skutečně reprezentativní
- čím více dat máme, tím lépe
  - pořád ale musí být reprezentativní! – když budeme mít mnoho nereprezentativních informací, je to horší než málo reprezentativních



↑ zkreslení
↓ rozptyl
A

↓ zkreslení
↑ rozptyl
B

↑ zkreslení
↑ rozptyl
C

↓ zkreslení
↓ rozptyl
D

slideshare.net

Ze vzorku zpět k populaci

giveitanudge.com

Shrnutí

- pro statistické uvažování je podstatný koncept pravděpodobností
- existuje několik podob rozložení dat
- nejcennějším rozložením je distribuce normální
- pro statistické uvažování je zásadní centrální limitní teorém