Pravděpodobnost, logika usuzování, velikost vzorku

METODOLOGICKÝ PROSEMINÁŘ II

TÝDEN 6 | 28. BŘEZNA 2018

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

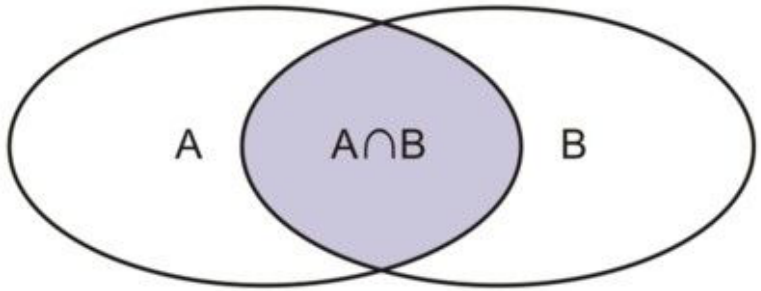
- pravděpodobnost jevu A označíme P(A) a definujeme ji jako
  - $P(A) = \frac{\text{počet příznivých jevů A}}{\text{počet možných událostí}} = \frac{|A|}{|\Omega|}$
- intuitivně využíváme pravděpodobnost jako připisování reálných čísel ke každému jevu způsobem, aby byl součet těchto čísel roven číslu 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

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

Skládání pravděpodobností - průnik

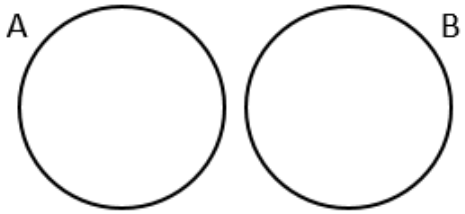
- pravděpodobnost existence více jevů
  - nezávislost dvou náhodných jevů znamená, že nastání jednoho jevu neovlivňuje pravděpodobnost nastání či nenastání druhého jevu
  - například opakované házení kostkou jsou nezávislé jevy; naopak opakované tahání karet z balíčku bez nahrazení vytažených karet nejsou nezávislé jevy
- $P(A \cap B)$ - průnik pravděpodobností (značíme ∩; objevuje se jev A i B)
- $P(A \cap B) = P(A) * P(B)$
  - pozor, tento vzorec platí pouze pro vzájemně nezávislé jevy!
  - pro závislé jevy je výpočet složitější



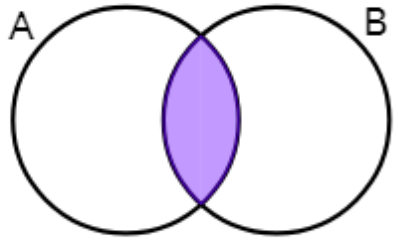
testbook.com

Skládání pravděpodobností - sjednocení

- pravděpodobnost existence jednoho nebo druhého jevu
- důležité je, zda jsou nebo nejsou jevy vzájemně výlučné
  - jevy A a B jsou vzájemně výlučné, jestliže nemůže dojít k tomu, že by se odehrály současně
  - např. jevy hlavy a panny jsou vzájemně výlučné; v kartách jevy srdce a spodka ale nejsou vzájemně výlučné
- $P(A \cup B)$ - sjednocení obou událostí (značíme ∪; objevuje se jev A nebo B nebo oba)
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - pokud nejsou jevy vzájemně výlučné, bylo by $P(A \cap B)$ započítáno dvakrát, proto jednou odečítáme
  - tento vzorec můžeme použít pro jevy, které se vzájemně vylučují i nevylučují (v případě na sobě výlučných jevů bude $P(A \cap B)$ rovno 0)



vzájemně výlučné jevy; onlinemathlearning.com



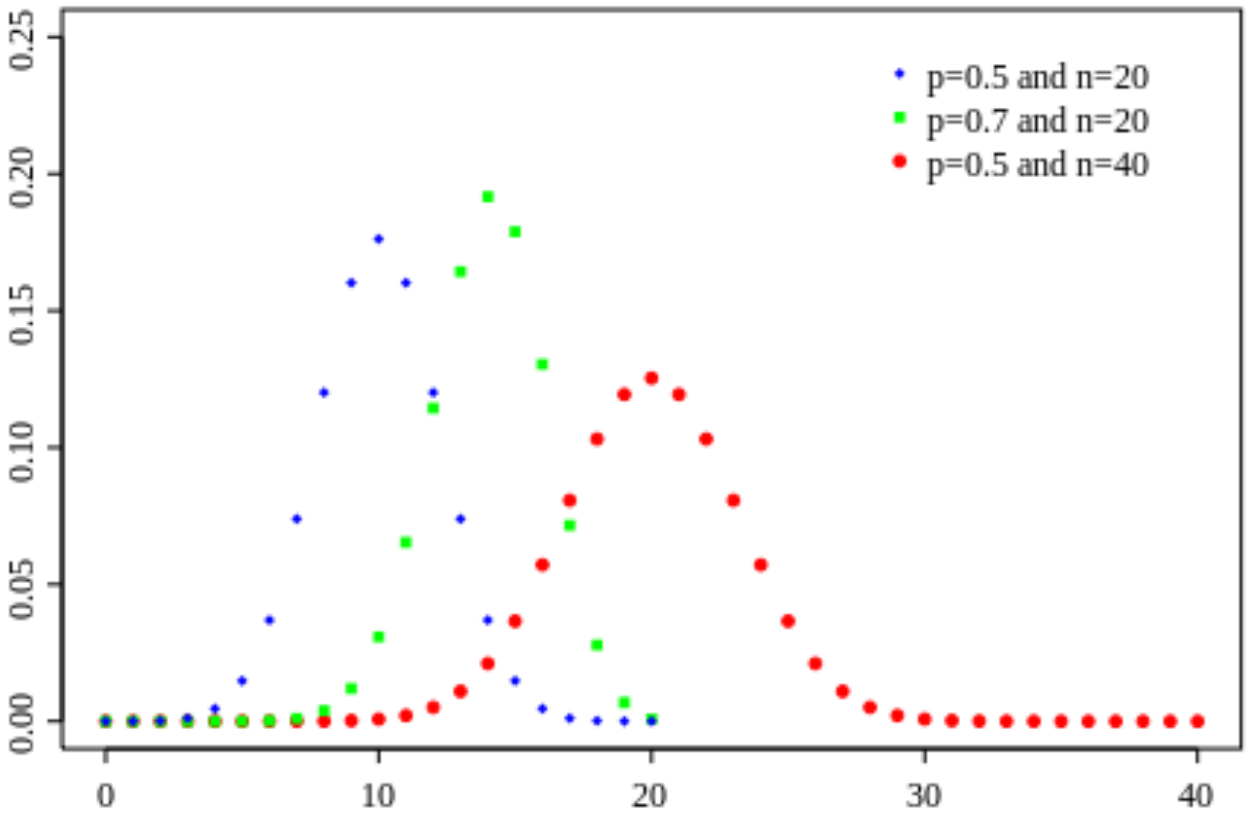
jevy nejsou vzájemně výlučné; onlinemathlearning.com

Příklad

- v krabici jsou 2 černé koule a 8 bílých – jaká je pravděpodobnost vytažení černé koule?
  - $P(\check{c}ern\acute{a}) = \frac{2}{10} = 0{,}2$
- jaká je pravděpodobnost, že na hrací kostce padne 3x po sobě šestka?
  - $P(3x\ \check{s}estka) = \frac{1}{6} * \frac{1}{6} * \frac{1}{6} = \frac{1}{216} \cong 0{,}0046$
- jaká je pravděpodobnost, že při hodu kostkou padne liché číslo nebo číslo dělitelné čtyřmi?
  - jevy se vzájemně vylučují (nemůže padnout číslo, které by bylo liché a zároveň dělitelně čtyřmi)
  - $P(lich\acute{e}\ \check{c}\acute{i}slo\ \cup \check{c}\acute{i}slo\ d\check{e}liteln\acute{e}\ 4) = \frac{1}{2} + \frac{1}{6} = \frac{3}{6} + \frac{1}{6} = \frac{2}{3} \cong 0{,}67$
- jaká je pravděpodobnost, že při hodu kostkou padne liché číslo nebo číslo větší než 3?
  - POZOR, jde o jevy, které se navzájem nevylučují
  - oba jevy se vzájemně překrývají (číslo 5 splňuje podmínku lichého čísla i čísla většího než 3), a proto je nutné odečíst jejich překryv $P(A \cap B)$
  - $P(lich\acute{e}\ \check{c}\acute{i}slo \cup \check{c}\acute{i}slo > 3) = \frac{1}{2} + \frac{1}{2} - \frac{1}{6} = \frac{5}{6} \cong 0{,}83$

Binomické rozdělení

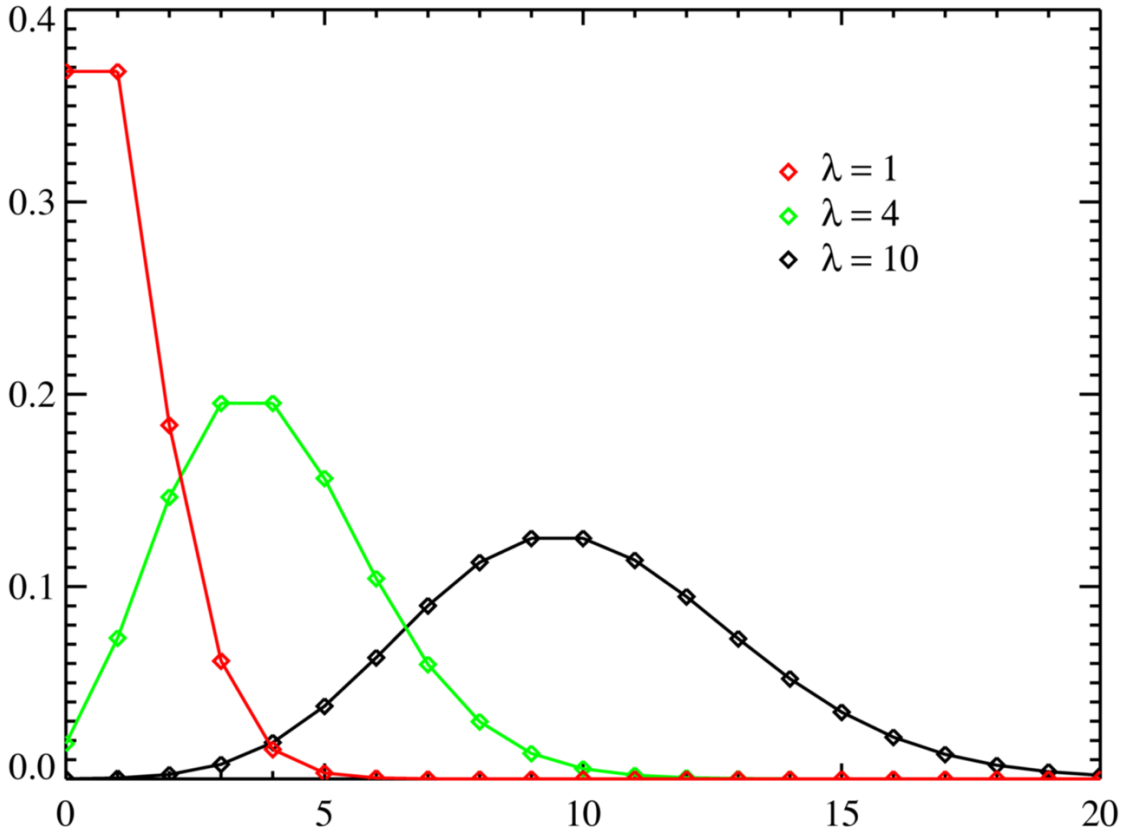
- nespojitá distribuce pravděpodobností úspěchu na sobě nezávislých experimentů s dichotomickým výsledkem (úspěch vs. neúspěch)
- určujícími parametry jsou počet pokusů (n) a pravděpodobnost úspěchu (p)



en.wikipedia.org

Poissonovo rozdělení

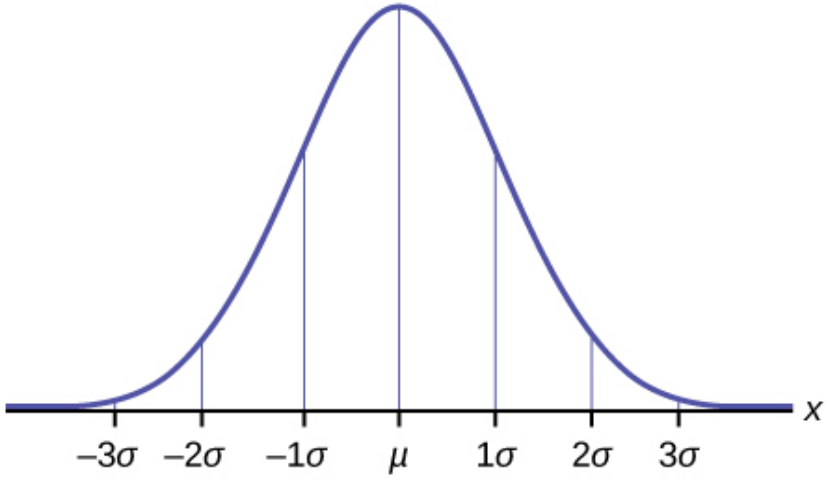
- rozdělení pravděpodobností vyjadřující počet výskytů jevu v určitém intervalu času nebo prostoru
- příkladem může být počet přednesených poslaneckých projevů v jednom dni
- konkrétní podoba rozdělení je charakterizována hodnotou λ (lambda)



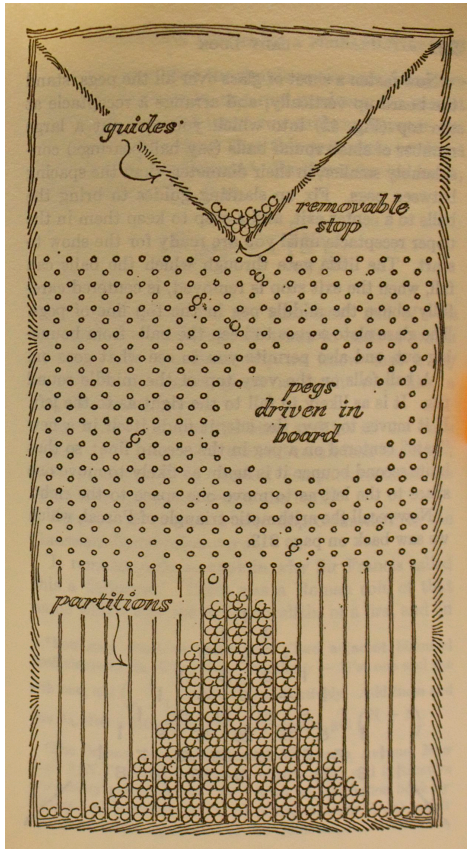
cs.wikipedia.org

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

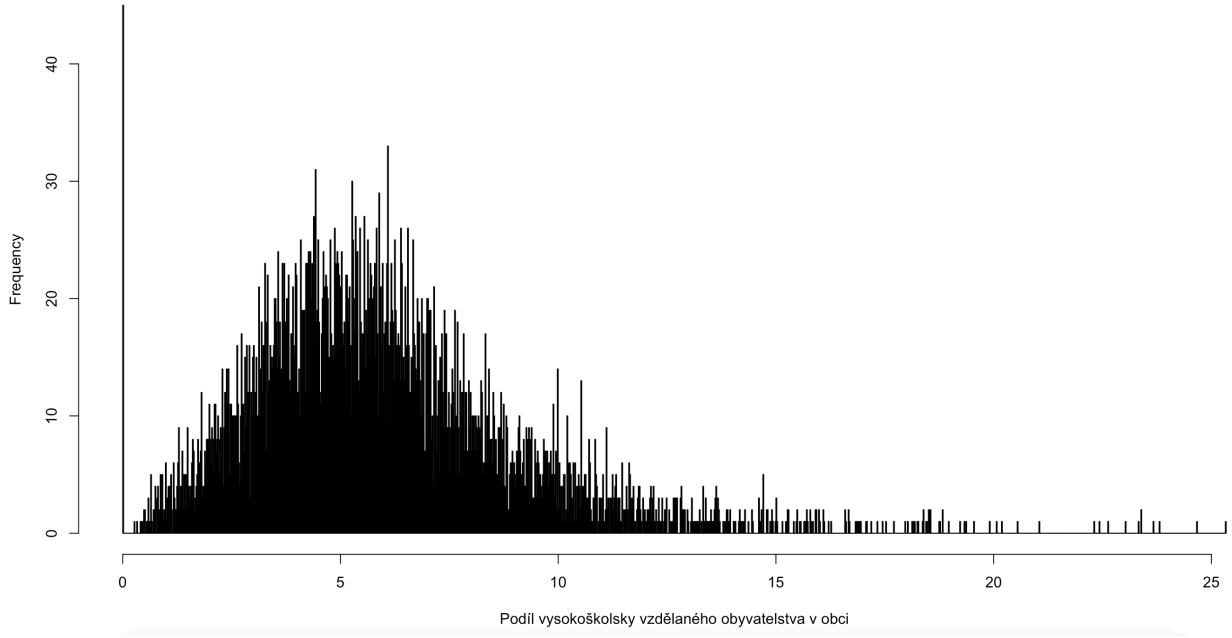


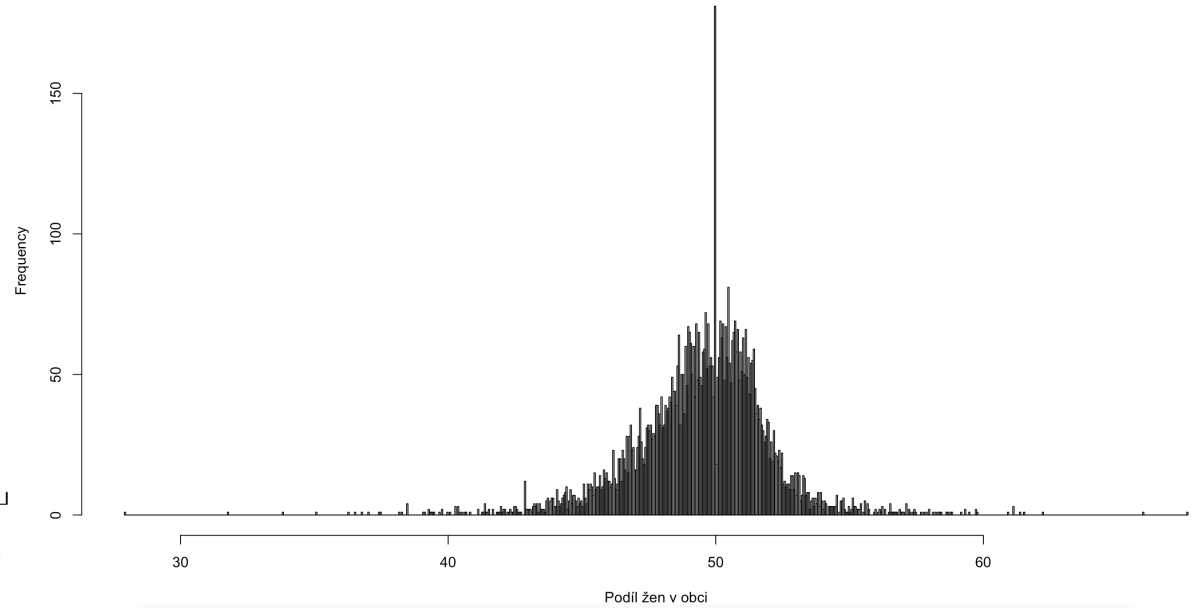
quadrantgroup.co.uk

Galtonova deska

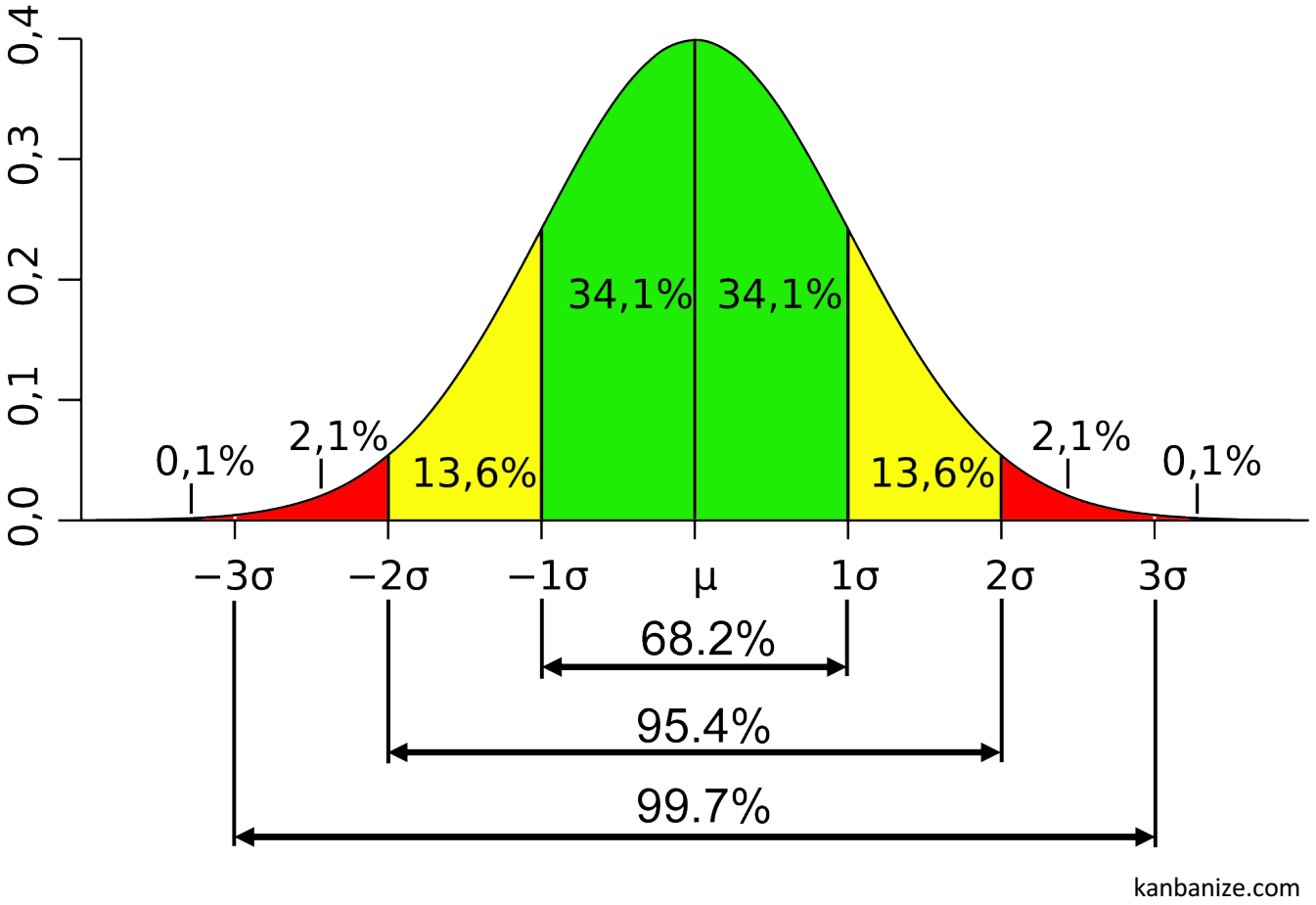
Normální rozložení kolem nás

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





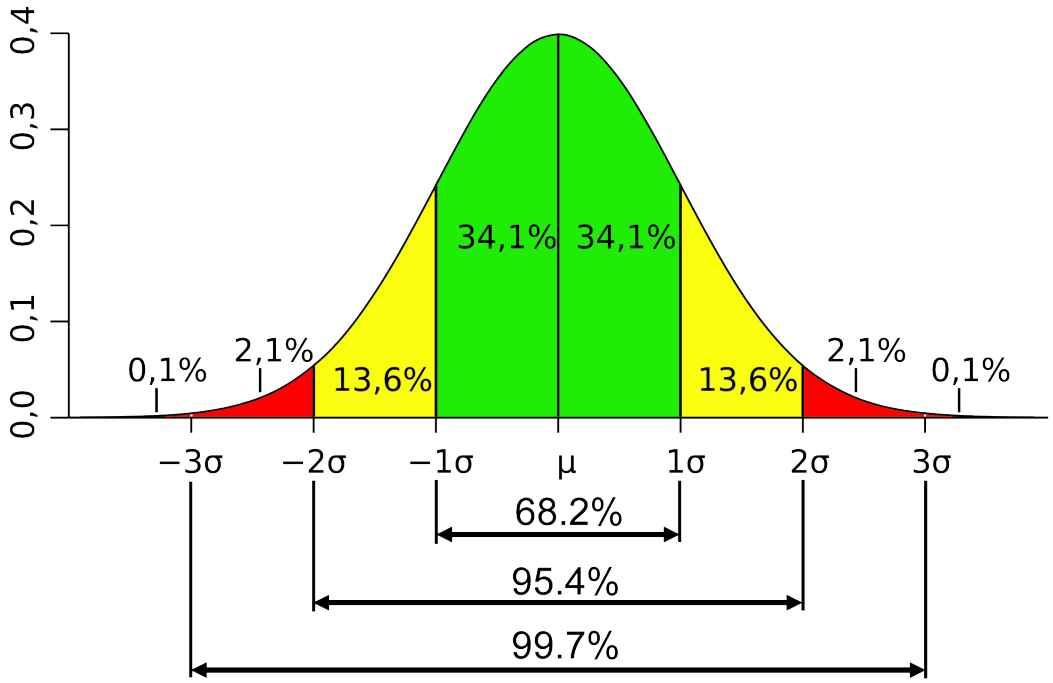
Přínosy normálního rozložení



kanbanize.com

Přínosy normálního rozložení

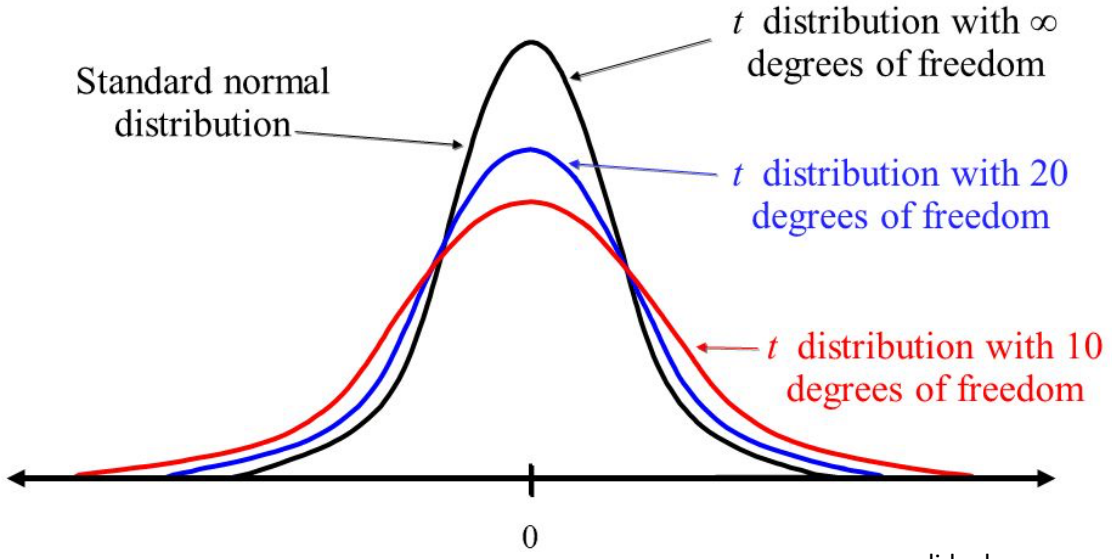
- výhodou je, že máme popsáno rozložení dat
- víme, s jakou pravděpodobností nastanou určité jevy (objeví se určité hodnoty)
- jde o to, s jakou pravděpodobností může pozorovaný vzorek (s určitým průměrem) pocházet z nepozorované populace



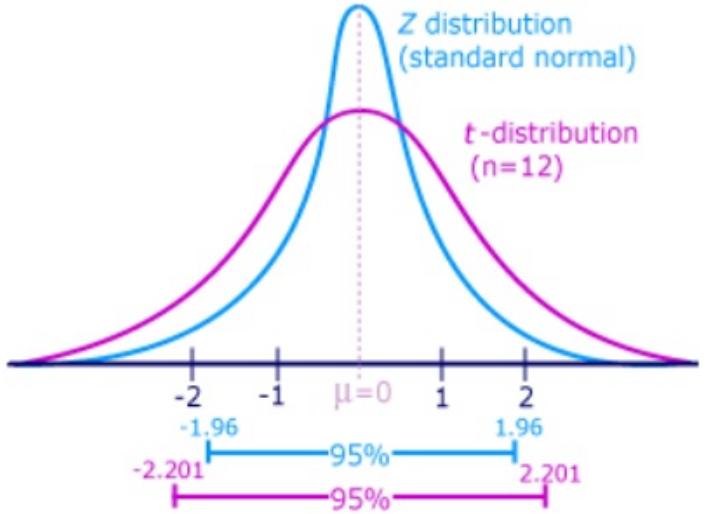
kanbanize.com

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, které mohou variovat aniž by byla změněna hodnota průměru (df = n – 1)



slideplayer.com



slideshare.net

Centrální limitní teorém

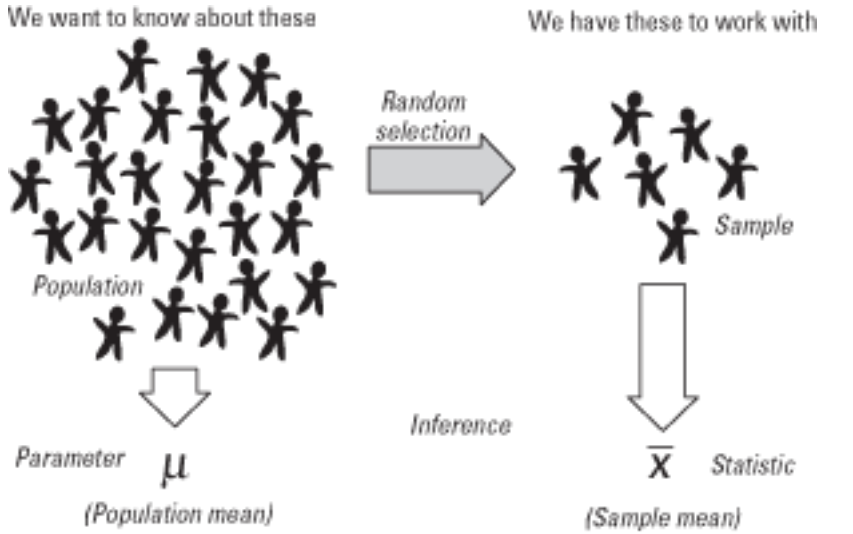
- existuje populace s průměrem μ a rozptylem σ^2
- z této populace vybereme vzorek o velikosti *n*
- to uděláme opakovaně a získáme tak výběrový vzorek průměru
- distribuce tohoto výběrové vzorku se blíží normálnímu rozložení s průměrem μ a rozptylem σ^2/n, jak se velikost vybíraného vzorku zvětšuje
- **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

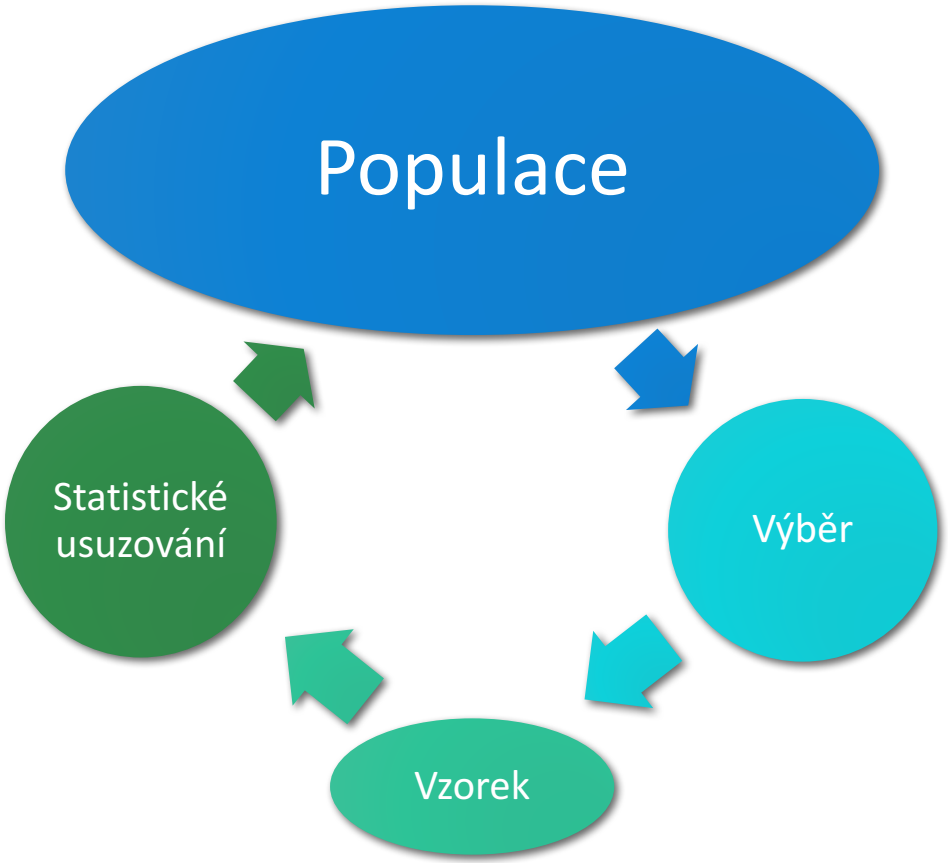
- důvodem je praktická nutnost
- nemáme dostatek prostředků (finančních, personálních, časových atd.) k tomu, abychom zkoumali celou populaci
- proto si vybereme pouze vzorek a pro něj shromáždíme data
- důležitá je metoda výběru (náhodný, kvótní apod. – viz předchozí semestr a zadaná literatura)
- klíčová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

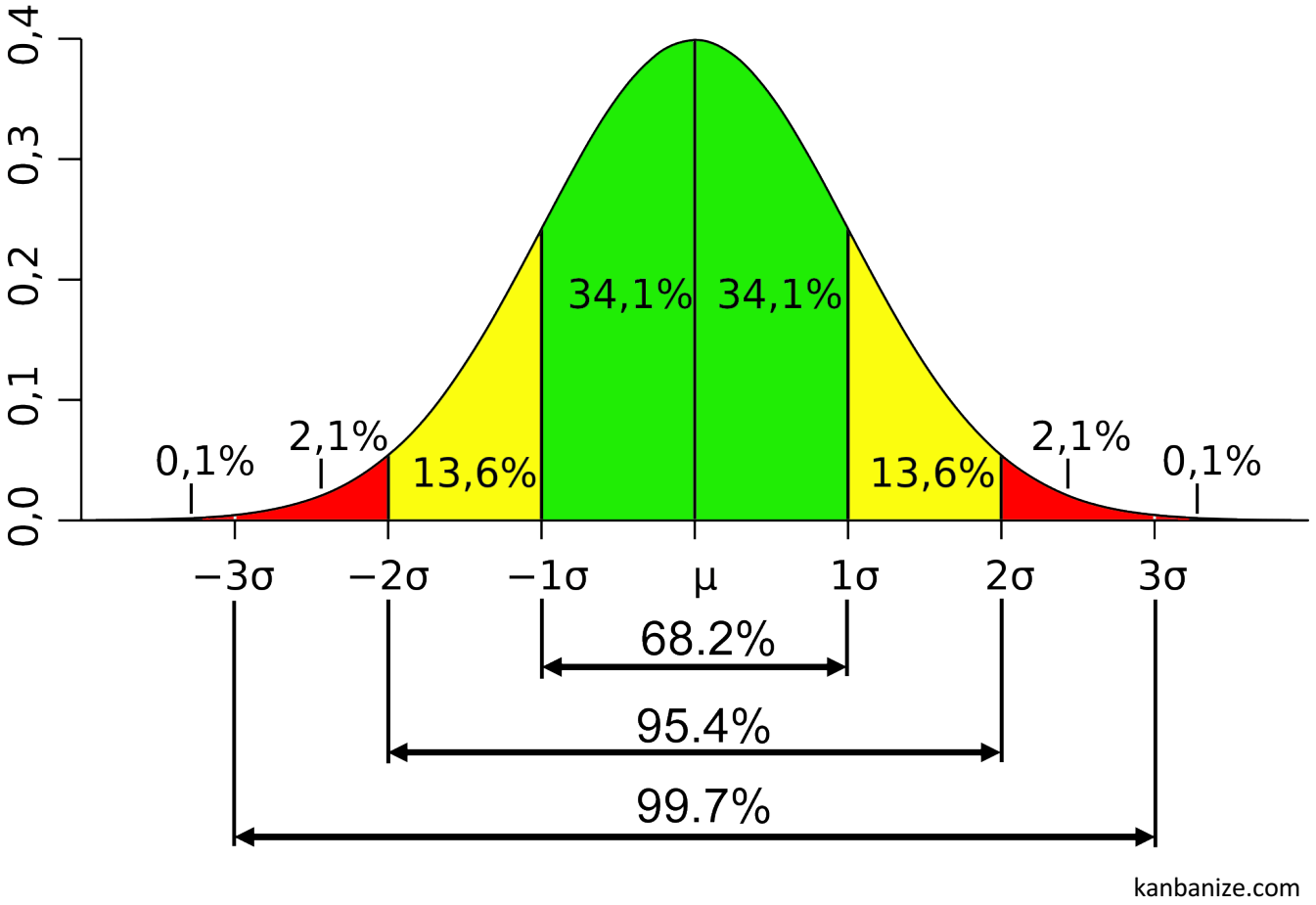
Statistická chyba

- 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 politologii (v poslední době se prosazuje použití 99 %)
- z normálního rozložení víme, že 95 % rozložení dat spadá mezi 1,96 standardní 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)
  - 45,71,58,68,87,54,60,42,48,60,51,44,56,61,74,58,58,61,68,40,52,56,68,54,48,71,68,54 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 politologii (v poslední době se prosazuje použití 99 %)
- z normálního rozložení víme, že 95 % rozložení dat spadá mezi 1,96 standardní odchylky na jednu i druhou stranu od průměru
- $60 \pm 1{,}96 * 1{,}41 \cong [57{,}24; 62{,}76]$
  - jsme si na 95 % jistí, že průměr populace se nalézá v tomto intervalu
  - jde o tzv. interval spolehlivosti (*confidence interval*) - http://rpsychologist.com/d3/CI/

t distribuce

- v případě malého vzorku (n < 40) využíváme t distribuci (raději než normální rozložení)
- je „přísnější“ – v případě malých vzorků je v krajních částech distribuce více prostoru
- srovnání t distribuce s různými stupni volnosti a normálním rozložením
  - http://rpsychologist.com/d3/tdist/
- stejně jako v předchozím případě má vzorek průměr $\bar{x}$ = 60 a směrodatnou odchylku s = 10
- ovšem nyní je vzorek jen o velikosti 15 respondentů
- opět chceme mít jistotu 95 %, že jsme odhalili skutečný průměr populace
- z t tabulky zjistíme, že pro 14 stupňů volnosti (df = n – 1) a 95 % úroveň jistoty je nutné standardní chybu vynásobit hodnotou 2,145
- $60 \pm 2{,}145 * \frac{10}{\sqrt{15}} \cong [51{,}70; 68{,}30]$ (všimněte si, že interval je větší než v předchozím případě)
  - interpretace je opět stejná - jsme si na 95 % jistí, že průměr populace se nalézá v tomto intervalu

Ze vzorku zpět k populaci

giveitanudge.com

Shrnutí

- koncept pravděpodobnosti
  - logika užití, základní výpočty (sčítání a násobení)
- různé druhy rozložení dat
  - nejdůležitějším rozložením je to normální
- základem pro statistické uvažování je centrální limitní teorém
- logika usuzování je založena na práci se vzorkem a zobecněním pro celou populaci
- podstatné je vždy uvádět statistickou chybu
  - vždy existuje!