Všechno, co jste chtěli vědět z teorie pravděpodobnosti, z teorie informace a ...

báli jste se zeptat

(2. část)

Jedinečnou funkcí statistiky je, že umožňuje vědci číselně vyjádřit nejistotu v jeho závěrech. (G. W. Snedecor)

(pro potřeby přednášky Úvod do strojového učení, PFL054)

Náhodná veličina

- náhodný jev $\omega \in \Omega$ chceme popsat prostřednictvím některé jeho číselné charakteristiky $X(\omega)$, kterou nazveme **náhodná veličina**; $\boldsymbol{X} : \Omega \to \mathcal{R}$
- **diskrétní** (nabývá konečného nebo spočetného počtu hodnot), **spojitá** (nabývá všech hodnot z daného intervalu)
- základní charakteristiky: průměr, rozptyl

Diskrétní pravděpodobnostní rozdělení

- $\Sigma_{(i=1 \ldots \infty)} P[X=x_i] = 1$
- seznam hodnot, kterých nabývá diskrétní náhodná veličina, a seznam pravděpodobností, s nimiž těchto hodnot náhodná veličina nabývá, udává **diskrétní pravděpodobnostní rozdělení**

Střední hodnota (průměr) diskrétní náhodné veličiny

- $E[X] \equiv \Sigma_{i=1\ldots n} x_i \, P(X=x_i)$ (μ)
- $E[X] \equiv \Sigma_{i=1\ldots \infty} x_i \, P(X=x_i)$

Rozptyl (variance)

- popisuje velikost kolísání náhodné veličiny kolem střední hodnoty
- **var [*X*] = E (*X*-E[*X*])^2 (σ^2)**

Směrodatná odchylka

- $\sigma = \sqrt{\mathrm{var}[X]}$

Spojitá náhodná veličina

- pravděpodobnostní rozdělení je popsáno hustotou (frekvenční fcí) f(x)

Binomické rozdělení - motivace

- hod mincí: panna? orel?
- Jaká je pravděpodobnost p, že padne panna?
- Házejme n-krát, z toho r-krát padla panna
- $\mathbf{p = r/n}$
- opakujme n hodů mincí; $\mathbf{r' \neq r,\ p' \neq p}$

Binomické rozdělení – motivace (pokračování)

- binomické rozdělení popisuje, pro libovolnou hodnotu **r**, pravděpodobnost jevu, že při **n** nezávislých hodech mincí právě **r**-krát padne **panna** za předpokladu, že pravděpodobnost **panny** v jednotlivých hodech je **p**

Kdy binomické rozdělení?

1. výsledky pokusu se dají popsat náhodnou veličinou ***X***, která má dvě možné hodnoty {0,1}
2. **P(*X*=1)** je dáno konstantou **p**, nezávislou na výsledku jakéhokoli pokusu; většinou je **p** neznámé – JAK ODHADNOUT?

Binomické rozdělení *Bin(n,p)*

- n nezávislých pokusů, zdar/nezdar - prostor elementárních jevů $\Omega = \{0,1\}^n$
- náhodná veličina $X(\omega) = \Sigma_{(i=1 \ldots n)}\omega_i$ vyjadřuje počet (0,1,…*n*) úspěchů v *n* nezávislých pokusech, kdy v každém z jednotlivých pokusů je pravděpodobnost úspěchu rovna *p*
- $\omega \in \Omega$, $\omega=(\omega_1,\omega_2,\ldots,\omega_n)$, ω_i je počet zdarů v i-tém pokusu, $p(\omega_i) = p^{\omega_i} (1-p)^{(1-\omega i)}$
- nezávislost pokusů: $p(\omega) = \Pi_{(i=1..n)}p(\omega_i) = p^{\Sigma\, \omega i}(1-p)^{(n-\, \Sigma\, \omega i)}$
- pro $k=\Sigma_{(i=1 \ldots n)}\omega_i$, je počet elem. jevů = $n!/k!(n-k)!$
- $P(X=k) = n!/k!(n-k)!\ p^k(1-p)^{(n-k)}$

Binomické rozdělení: střední hodnota, rozptyl, směrodatná odchylka

- $E[X] = np$
- $var[X] = np(1-p)$
- $\sigma = \sqrt{np(1-p)}$

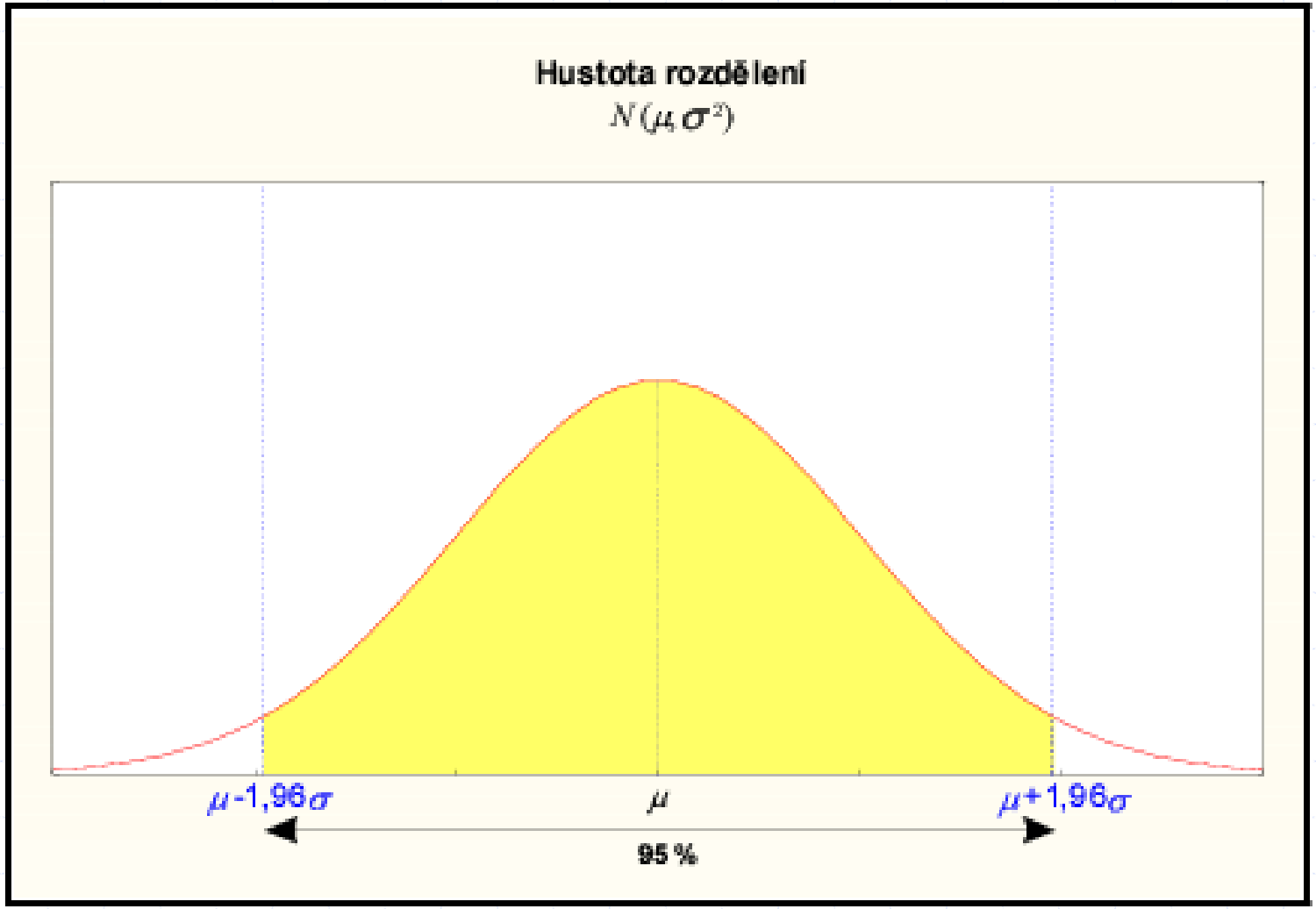
Normální rozdělení (spojité) $N(\mu, \sigma^2)$

- $f(x) = 1/(\sqrt{2\pi\sigma^2})e^{-1/2((x-\mu)/\sigma)^2}$
- normální rozdělení je určeno parametry μ (střední hodnotou) a σ (sm. odchylkou)
- μ a σ jsou konstanty, které určují polohu křivky na ose x (μ) a její roztažení podél osy x (σ)

Normální rozdělení - pokračování

- Jestliže náhodná veličina X vyhovuje normálnímu rozdělení, potom:
  - $P(X \in (a,b)) = p(x)dx$
  - $E[X] = \mu$, $var(X) = \sigma^2$, $\sigma_X = \sigma$

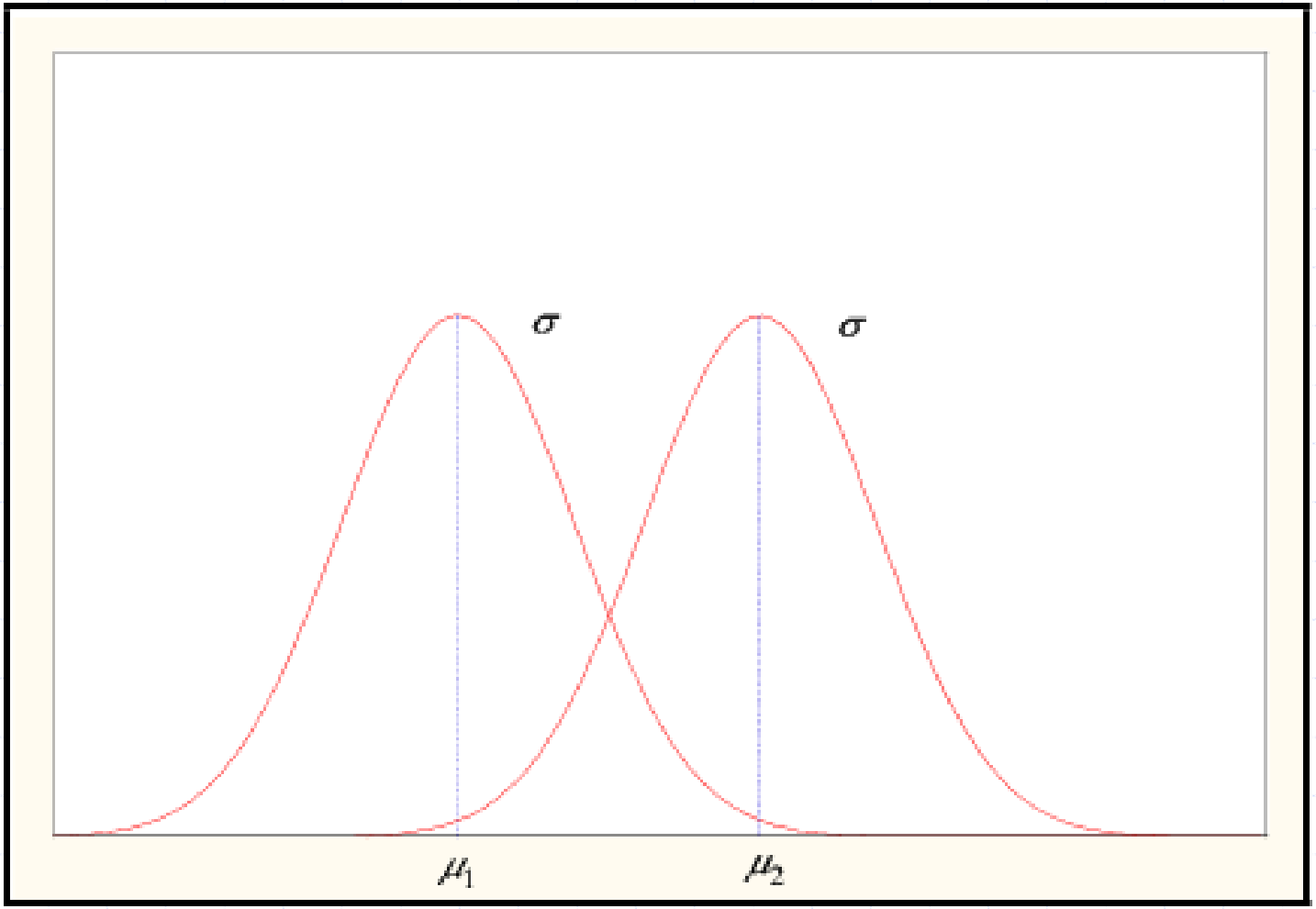
Normální rozdělení graficky



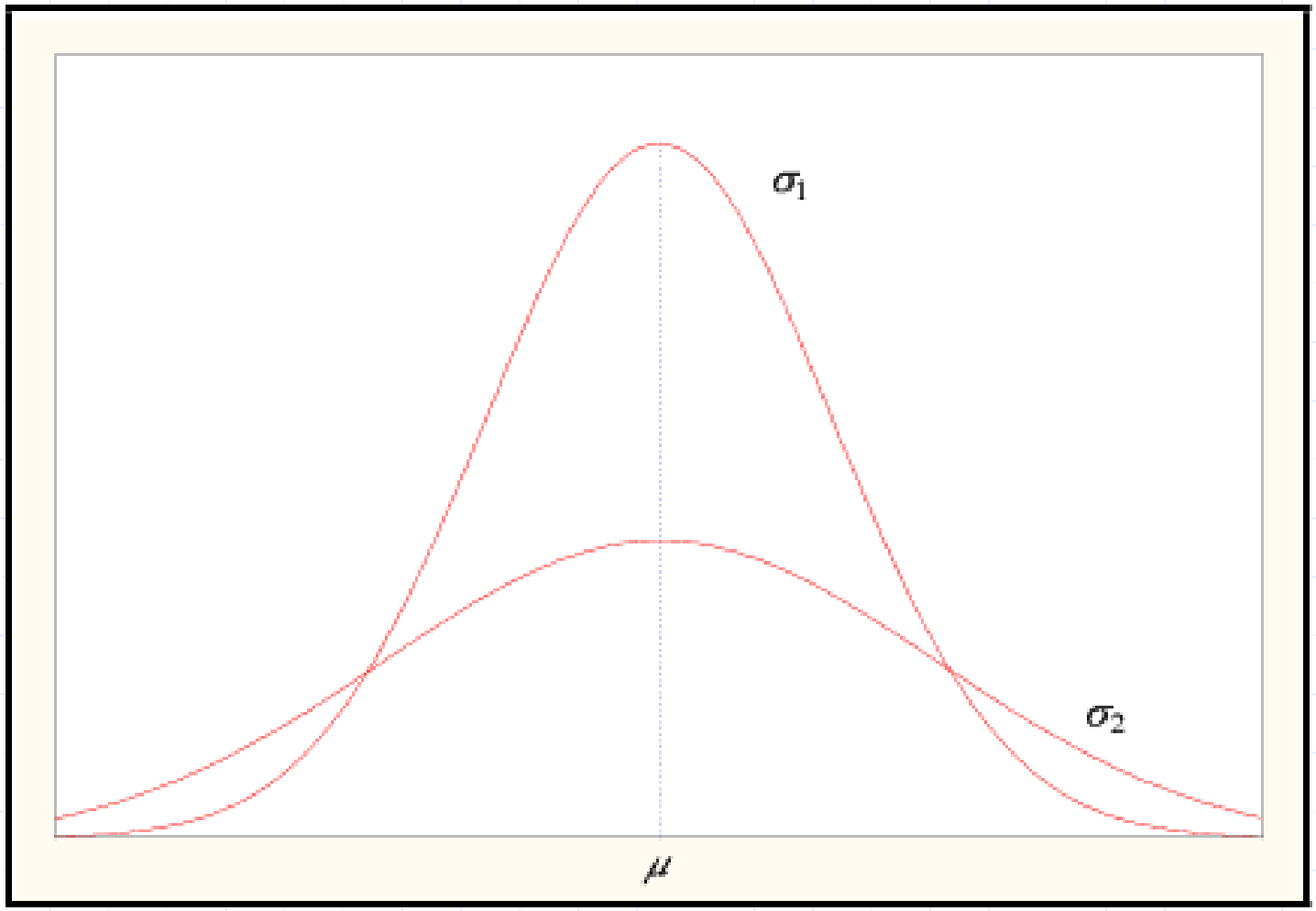
Normální rozdělení graficky - vysvětlení

- jednovrcholové, symetrické okolo střední hodnoty
- plocha pod křivkou hustoty je rovna jedné
- pravděpodobnost, že náhodná veličina nabude hodnot z určitého intervalu, je rovna ploše pod hustotou nad tímto intervalem
- např. pro interval s hranicí –1,96 a 1,96 má tato plocha velikost 0,95. Náhodná veličina nabývá hodnot z tohoto intervalu s 95% pravděpodobností a pouze s 5% pravděpodobností leží její hodnoty mimo uvedený interval

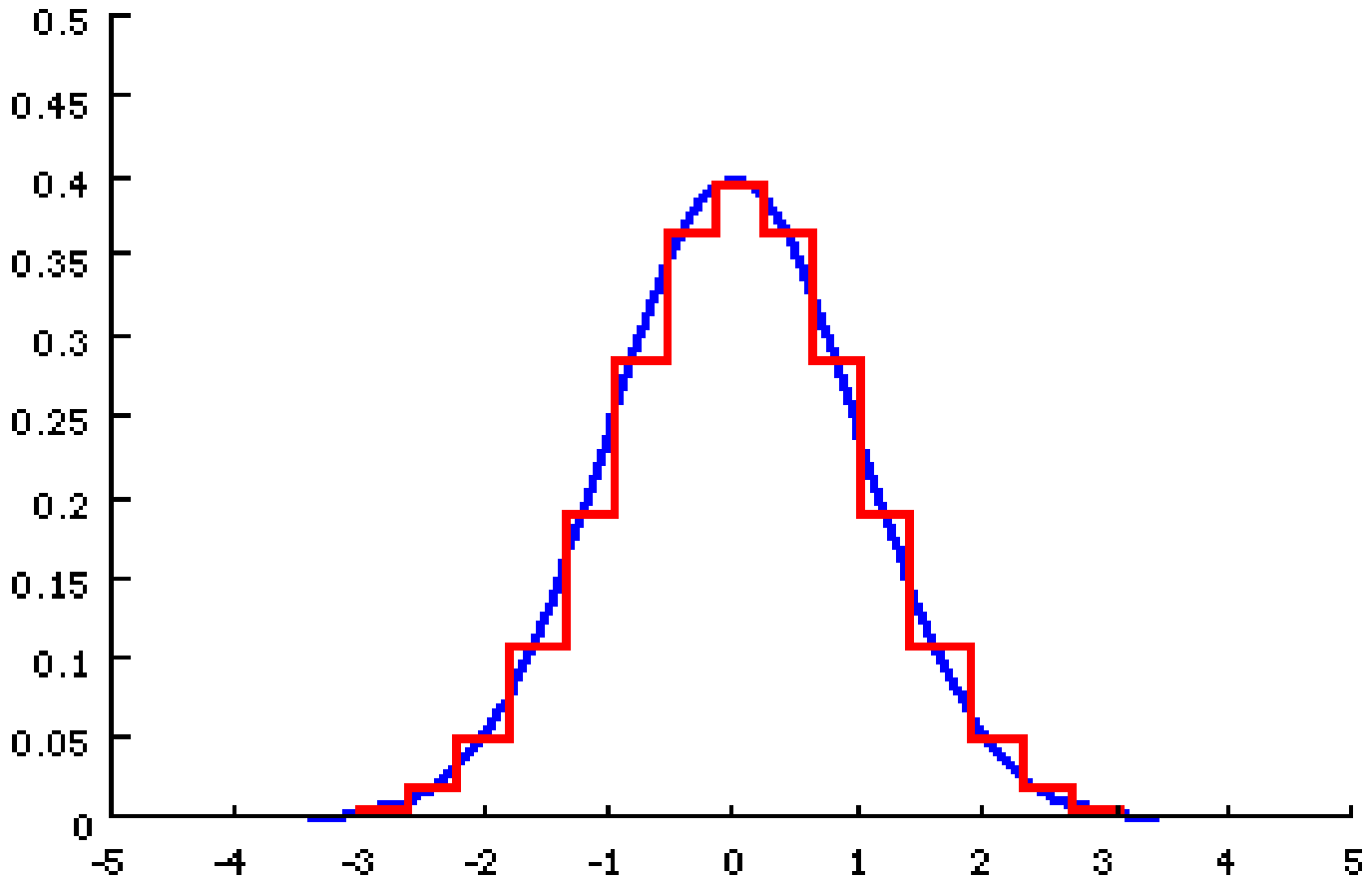
Průměr náhodné veličiny určuje polohu rozdělení na na číselné ose ($\mu_1 < \mu_2$)



Směrodatná odchylka určuje tvar hustoty ($\sigma_1 < \sigma_2$)



Centrální limitní věta

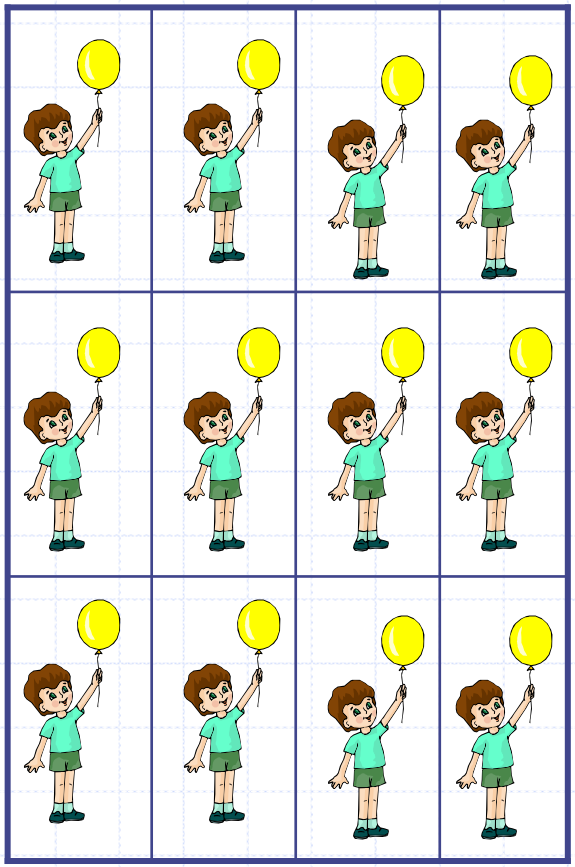


Statistická metodologie

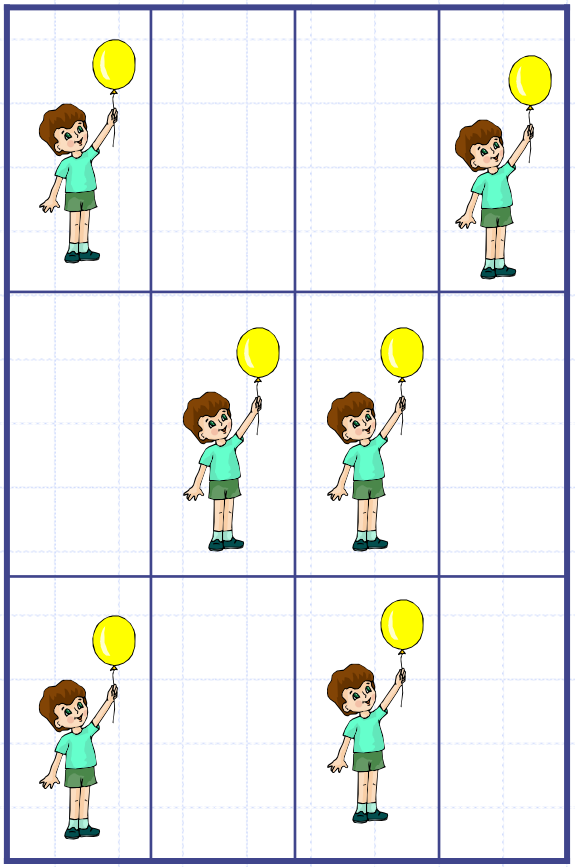
Nemusíte sníst celého vola na to, abyste poznali, že maso je tuhé. (S. Johnson)

- **induktivní statistika** – zobecňování závěrů s udáním stupně jejich nejistoty; schopnost učit se ze zkušenosti
- **populace**: základní soubor (výčtem/vymezením některých společných vlastností)
  - **parametr**: číselná charakteristika populace (např. průměrná výška osmiletých dětí v ČR)
- **výběr**: požadované vlastnosti se zjišťují pouze u některých prvků populace; reprezentativnost výběru; za určitých předpokladů se dají závěry z výběrů pomocí statistické indukce zobecnit na celou populaci s vyjádřením míry nejistoty zobecňovaných závěrů

populace
12 osmiletých dětí



výběr 6 dětí



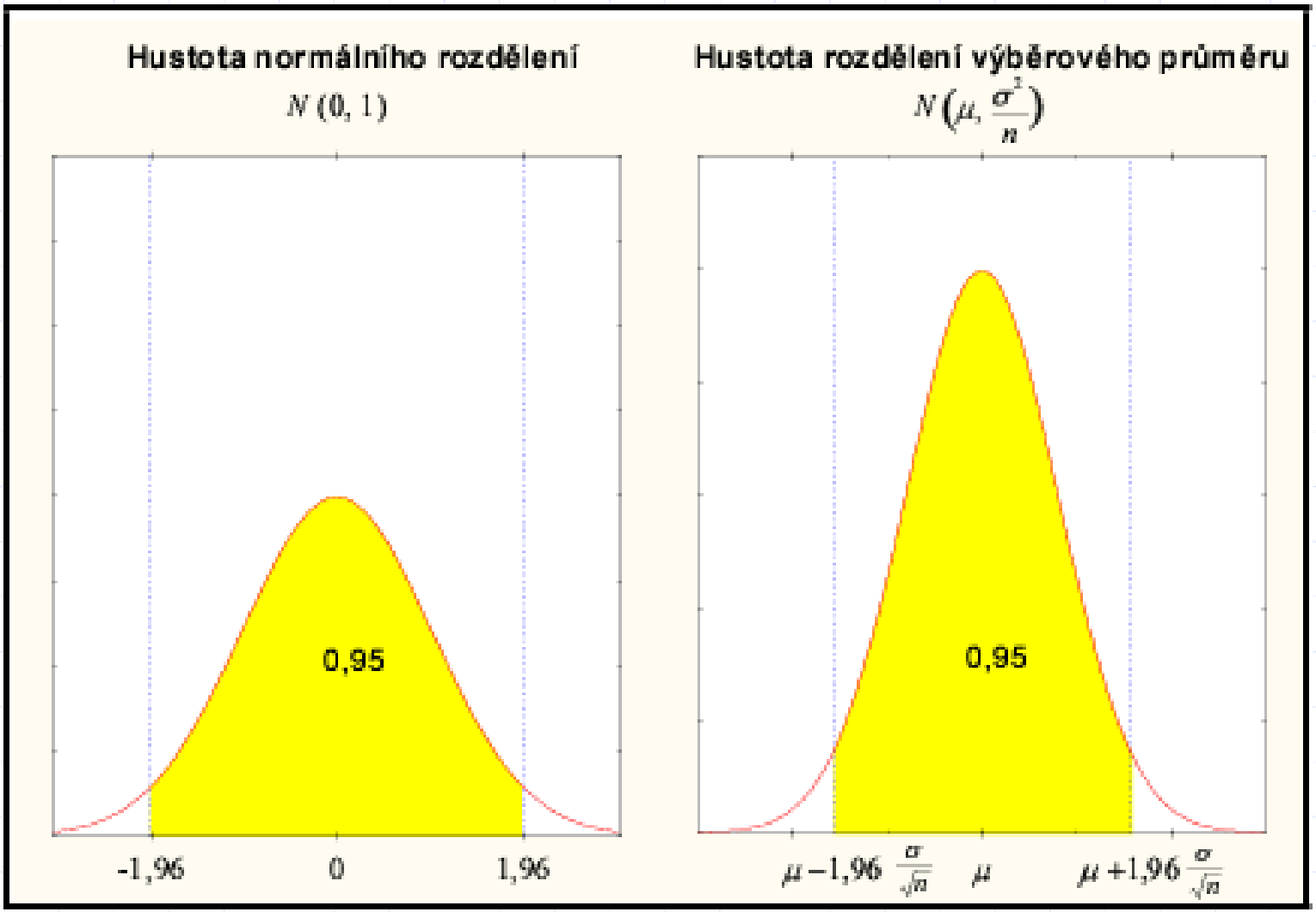
Zkreslení odhadu

- odhad: je náhodná veličina použitá pro odhad parametru populace, z které je daný vzorek vybírán
- zkreslení odhadu libovolného parametru p : $\mathbf{E[X] - p}$
- nestranný odhad: $E[X] - p = 0$

Jak odhadnou populační průměr z výběru pomocí tzv. intervalu spolehlivosti?

- populační (μ) vs. výběrový ($\hat{x}$) průměr
- provedeme-li opakované výběr a spočítáme průměry, pak se tyto výběry budou obvykle chovat tak, jako kdyby pocházely z normálního rozdělení
- (bez důkazu) $\sigma_{výběr} = \sigma_{populace} / \sqrt{n}$, kde n je rozsah výběru, $\sigma_{výběr}$ je směrodatná odchylka rozdělení výběrových průměrů, $\sigma_{populace}$ je směrodatná odchylka původního rozdělení
- interval místo jednoduchého bodového odhadu

Vlastnosti rozdělení výběrového průměru



Interval spolehlivosti

- N% interval spolehlivosti pokrývá parametr p s pravděpodobností N

Interval spolehlivosti - pokračování

hranice spolehlivosti N%	50	68	80	90	95	98	99
konstanta z_n	0,67	1,00	1,28	1,64	1,96	2,33	2,58

- konstanta z_n určuje šířku nejmenšího intervalu kolem střední hodnoty, který pokrývá N% pravděpodobností v rámci normálního rozdělení
- čím vyšší je koeficient spolehlivosti, tím delší – a tedy méně přesný – je výsledný interval; je potřeba najít kompromis mezi požadovanou spolehlivostí a přesností odhadu, tj. délkou intervalu

Pro dané N - jak určit velikost intervalu, který obsahuje N% pstí?

- pro binomické rozdělení značně obtížné
- ALE – máme štěstí: pro dostatečně velkou množinu instancí je možné binomické rozdělení aproximovat rozdělením normálním se stejnou střední hodnotou a se stejným rozptylem (Centrální limitní věta)

Interval spolehlivosti

- jestliže náhodná veličina X vyhovuje normálnímu rozdělení se střední hodnotu μ a směrodatnou odchylkou σ, potom hodnota x veličiny X padne do intervalu $\mu \pm z_N \sigma$ v N% případů
- střední hodnota μ padne do intervalu $x \pm z_N \sigma$ v N% případů