Statistika

(MD360P03Z, MD360P03U)
ak. rok 2007/2008

Karel Zvára

karel.zvara@mff.cuni.cz
http://www.karlin.mff.cuni.cz/∼zvara

5. listopadu 2007

alternativní rozdělení

- diskrétní, s jediným parametrem π (nikoliv Ludolfovo číslo)
- $\mathrm{P}(X=1)=\pi, \qquad \mathrm{P}(X=0)=1-\pi \qquad (0<\pi<1)$
- X – kolikrát v jednom pokusu došlo k události, která má pravděpodobnost π (jen dvě možné hodnoty: 0 nebo 1)
- **střední hodnota** (populační průměr)

$$\mu_X = 1\cdot\mathrm{P}(X=1)+0\cdot\mathrm{P}(X=0)=\pi$$

- (populační) **rozptyl**

$$\begin{aligned}\sigma_X^2 &= (1-\mu_X)^2\mathrm{P}(X=1)+(0-\mu_X)^2\mathrm{P}(X=0)\\ &= (1-\pi)^2\cdot\pi+(0-\pi)^2\cdot(1-\pi)\\ &= (1-\pi)^2\pi+\pi^2(1-\pi)=\pi(1-\pi)\end{aligned}$$

binomické rozdělení bi(n,π) (1)

- diskrétní rozdělení s parametry $n,\pi \qquad (0<\pi<1)$
- n **nezávislých** pokusů
- v každém zdar s pravděpodobností π, nezdar s pstí $1-\pi$
- **celk. počet zdarů** X má binomické rozdělení s parametry n, π
- zapisujeme $X\sim\mathrm{bi}(n,\pi)$
- X je součet n nezávislých náhodných veličin X_i
  (X_i = počet zdarů v i-tém pokusu)
  každé X_i má alternativní rozdělení s parametrem π
- z vlastnosti střední hodnoty součtu náh. veličin: $\boxed{\mu_X=n\pi}$
- z vlastnosti rozptylu součtu nezávislých náhodných veličin

$$\boxed{\sigma_X^2=n\pi(1-\pi)}$$

binomické rozdělení bi(n,π) (2)

- pravděpodobnosti možných hodnot

$$\mathrm{P}(X=k)=\binom{n}{k}\pi^k(1-\pi)^{n-k}, \qquad k=0,1,,\ldots,n$$

- pst, že v **daných** k pokusech zdar Z, v ostatních nezdar N

$$\underbrace{ZZ\ldots Z}_{k}\underbrace{NN\cdots N}_{n-k}\text{s pstí }\pi^k(1-\pi)^{n-k}$$

- zvolíme k míst pro zdar Z, na ostatních místech nezdar N, počet možností:

$$\binom{n}{k}=\frac{n!}{k!(n-k)!}=\frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots2\cdot1}$$

příklad: zkoušky

- C – zdar = udělat zkoušku, $\mathrm{P}(C) = 0{,}8$
- zkoušku dělá $n = 10$ studentů stejně připravených (u všech stejná pravděpodobnost π), studenti neopisují (nezávislost)
- pst, že zkoušku udělá nějakých 9 studentů

$$\mathrm{P}(X = 9) = \binom{10}{9} \cdot 0{,}8^9 \cdot 0{,}2^1 = 10 \cdot 0{,}8^9 \cdot 0{,}2^1 = 0{,}268$$

- pst, že právě jeden student (nějaký) zkoušku neudělá

$$\mathrm{P}(Y = 1) = \binom{10}{1} \cdot 0{,}2^1 \cdot 0{,}8^9 = 10 \cdot 0{,}2^1 \cdot 0{,}8^9 = 0{,}268$$

- pst, že zkoušku udělá **daných** 9 studentů: 0,0268

příklad: kouření

- víme, že mezi dvacetiletými muži je (řekněme) 35 % kuřáků (např. je-li 70 tisíc dvacetiletých, pak je mezi nimi asi 24 500 kuřáků, ale nevíme, kteří to jsou)
- vybereme náhodně 60 dvacetiletých mužů, X – počet kuřáků mezi nimi, tedy $X \sim \mathrm{bi}(60, 0{,}35)$
-

$$\mu_X = 60 \cdot 0{,}35 = 21 \qquad \sigma_X^2 = 60 \cdot 0{,}35 \cdot 0{,}65 = 13{,}65 = (3{,}7)^2$$

- ukázky pravděpodobností možných hodnot

[BINOMDIST(15;60;0,35;0)] [dbinom(15,60,0.35)]

k	15	17	19	21	23	25
$\mathrm{P}(X = k)$	0,029	0,062	0,095	0,107	0,091	0,059

Poissonovo rozdělení Po(λ) (1)

- diskrétní rozdělení (zákon vzácných jevů), $Y \sim \mathrm{Po}(\lambda)$
- Y – počet výskytů jevu ve zvolené časové (prostorové, plošné ...) jednotce
- $\lambda > 0$ – jediný parametr, intenzita výskytu jevu (jak často se v průměru vyskytuje ve zvolené jednotce)

$$\mathrm{P}(Y = k) = \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda}, \qquad k = 0, 1, \ldots$$

- střední hodnota, (populační) rozptyl

$$\mu_Y = \lambda, \qquad \sigma_Y^2 = \lambda$$

- u binomického rozdělení bylo $\mu_X > \sigma_X^2$, zde rovnost

Poissonovo rozdělení Po(λ) (2)

- parametr λ znamená hustotu na jednotku plochy (populační průměr počtu případů na jednotku)
- změníme-li jednotku plochy, změní se parametr: při počítání pravděpodobností toho, kolikrát najdeme případ na trojnásobku původní jednotky (trojnásobné ploše, ve trojnásobném čase ...), bude novým parametrem 3λ
- analogicky pro jiné kladné násobky
- aproximace: $X \sim \mathrm{bi}(n, \pi)$, n velké, π malé ($\mu_X = n \cdot \pi$) pak pravděpodobnosti hodnot X lze aproximovat (přibližně vyjádřit) pomocí pravděpodobností hodnot $Y \sim \mathrm{Po}(n \cdot \pi)$
- Poissonovo rozdělení Po($n \cdot \lambda$) aproximuje binomické bi(n, π)

příklady Poissonova rozdělení

- do pasti padá za noc v průměru 8 brouků ($\lambda = 8$)
- s jakou pravděpodobností jich tam ráno najdeme 10?
  [POISSON(10;8;0)] [dpois(10,8)]

$$P(Y = 10) = \frac{8^{10}}{10!}e^{-8} = 0{,}099$$

- vezmeme-li past s poloviční obvodem, očekáváme poloviční průměr za noc ($\lambda = 4$)

$$P(Y = 10) = \frac{4^{10}}{10!}e^{-4} = 0{,}005$$

$$P(Y = 5) = \frac{4^{5}}{5!}e^{-4} = 0{,}156$$

souvislost binomického a Poissonova rozdělení

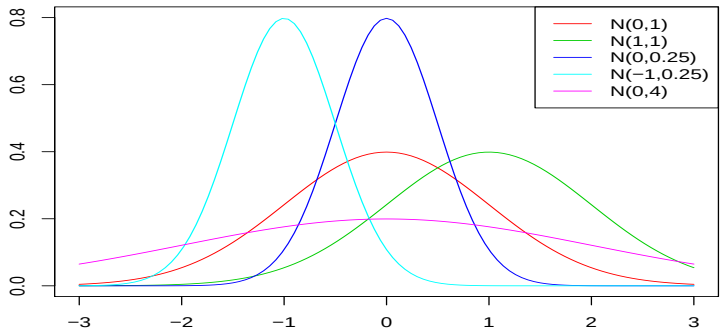
- s jakou pravděpodobností **neudělá** 12 z 50 stejně připravených studentů zkoušku? (pst neúspěchu = 0,2)
- binomické rozdělení bi(50, 0,2)
  [BINOMDIST(12;50;0,2)] [dbinom(12,50,0.2)]

$$P(X = 12) = \binom{50}{12} \cdot 0{,}2^{12} \cdot 0{,}8^{38} = 0{,}103$$

- Poissonovo rozdělení Po(50 · 0,2)=Po(10)
  [POISSON(12;10;0)] [dpois(12,10)]

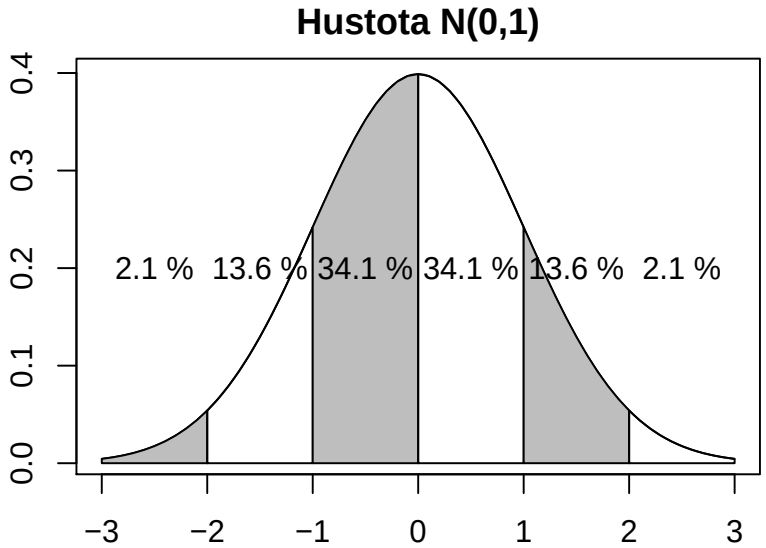
$$P(Y = 12) = \frac{10^{12}}{12!}e^{-10} = 0{,}095$$

normální (Gaussovo) rozdělení $N(\mu, \sigma^2)$



- spojité rozdělení, symetrické okolo střední hodnoty μ
- maximální hodnota hustoty je úměrná $1/\sigma$ ($\frac{1}{\sqrt{2\pi\sigma^2}} \doteq \frac{0{,}4}{\sigma}$)
- model vzniku: součet velkého počtu nepatrných příspěvků

normované normální rozdělení $Z \sim N(0, 1)$



příklady pravděpodobností o normálním rozdělení

- pro $X \sim \mathrm{N}(\mu, \sigma^2)$ platí

  $\boxed{\mu_X = \mathrm{E}\, X = \mu}$ $\qquad$ $\boxed{\sigma_X^2 = \mathrm{E}\,(X - \mu_X)^2 = \sigma^2}$

- $\boxed{X \sim \mathrm{N}(\mu, \sigma^2) \qquad \Rightarrow \qquad Z = \dfrac{X - \mu}{\sigma} \sim \mathrm{N}(0,1)}$

- $$\mathrm{P}\,(|Z| < c) = \mathrm{P}\left(\left|\frac{X-\mu}{\sigma}\right| < c\right) = \mathrm{P}\,(|X - \mu| < c \cdot \sigma)$$

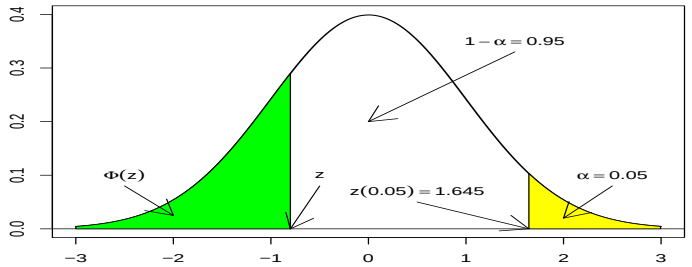
- tedy

$$\mathrm{P}(|X - \mu| < 1{,}00\,\sigma) = 0{,}68,\ \text{tj. } 68\ \%$$
$$\mathrm{P}(|X - \mu| < 1{,}96\,\sigma) = 0{,}95,\ \text{tj. } 95\ \%$$
$$\mathrm{P}(|X - \mu| < 2{,}00\,\sigma) = 0{,}9545,\ \text{tj. } 95{,}45\ \%$$
$$\mathrm{P}(|X - \mu| < 3{,}00\,\sigma) = 0{,}9973,\ \text{tj. } 99{,}73\ \%$$

normované normální rozdělení $Z \sim \mathrm{N}(0,1)$

tabelováno:

- hustota $\varphi(z)$
  [NORMDIST(z;0;1)] [dnorm(z)]
- distribuční funkce $\Phi(z) = \mathrm{P}(Z \le z)$
  [NORMSDIST(z)] [pnorm(z)]
- **kritické hodnoty** $z(\alpha)$: $\mathrm{P}(Z \le z(\alpha)) = \Phi(z(\alpha)) = 1 - \alpha$
  [NORMSINV(z)] [qnorm(z)]



zajímavé kritické hodnoty

$$z(0{,}025) = 1{,}96 \text{ tj. } \mathrm{P}(Z > 1{,}96) = 2{,}5\ \%$$
$$z(0{,}025) = 1{,}96 \text{ tj. } \mathrm{P}(Z < -1{,}96) = 2{,}5\ \%$$
$$z(0{,}025) = 1{,}96 \text{ tj. } \mathrm{P}(|Z| > 1{,}96) = 5\ \%$$
$$z(0{,}005) = 2{,}58 \text{ tj. } \mathrm{P}(Z > 2{,}58) = 0{,}5\ \%$$
$$z(0{,}005) = 2{,}58 \text{ tj. } \mathrm{P}(Z < -2{,}58) = 0{,}5\ \%$$
$$z(0{,}005) = 2{,}58 \text{ tj. } \mathrm{P}(|Z| > 2{,}58) = 1\ \%$$
$$z(0{,}050) = 1{,}64 \text{ tj. } \mathrm{P}(Z > 1{,}64) = 5\ \%$$
$$z(0{,}050) = 1{,}64 \text{ tj. } \mathrm{P}(Z < -1{,}64) = 5\ \%$$
$$z(0{,}050) = 1{,}64 \text{ tj. } \mathrm{P}(|Z| > 1{,}64) = 10\ \%$$

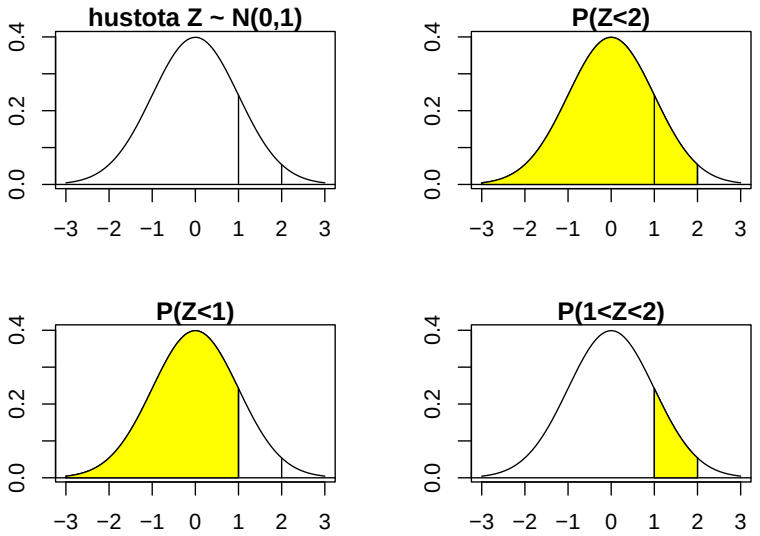
výpočet pravděpodobností pro $Z \sim \mathrm{N}(0,1)$

- u spojitého rozdělení je $\mathrm{P}(X < x) = \mathrm{P}(X \le x)$, tedy i u Z
- $Z \sim \mathrm{N}(0,1), a < b$, pak $\boxed{\mathrm{P}(a < Z < b) = \Phi(b) - \Phi(a)}$
- odvození: jevy $(Z \le a)$ a $(a < Z \le b)$ jsou neslučitelné (tvrzení nemohou platit současně)
  jejich sjednocením je jev $(Z \le b)$, proto

$$\mathrm{P}(Z \le b) = \mathrm{P}(Z \le a) + \mathrm{P}(a < Z \le b)$$
$$\Phi(b) = \Phi(a) + \mathrm{P}(a < Z \le b)$$

- příklad: $\mathrm{P}(1 < Z < 2) = \Phi(2) - \Phi(1) = 0{,}977 - 0{,}841 = 0{,}136$, jak bylo na obrázku
  [NORMSDIST(2)-NORMSDIST(1)] [pnorm(2)–pnorm(1)]

Postup výpočtu $\mathrm{P}(1 < Z < 2)$ $(Z \sim \mathrm{N}(0,1))$ pomocí tabelované funkce $\Phi(z) = F_Z(z) = \mathrm{P}(Z \leq z)$



výpočet pro $X \sim \mathrm{N}(\mu, \sigma^2)$

$$X \sim \mathrm{N}(\mu, \sigma^2) \Rightarrow Z = \frac{X-\mu}{\sigma} \sim \mathrm{N}(0,1)$$

$$\mathrm{P}(X \leq x) = \mathrm{P}\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = \mathrm{P}\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

$$\boxed{\mathrm{P}(a < X < b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

příklad: $X \sim \mathrm{N}(136{,}1, 6{,}4^2)$ (výšky 10letých hochů v roce 1951)

$$\mathrm{P}(134{,}5 < X < 140{,}5) = \Phi\left(\frac{140{,}5-136{,}1}{6{,}4}\right) - \Phi\left(\frac{134{,}5-136{,}1}{6{,}4}\right)$$
$$= 0{,}754 - 0{,}401 = 0{,}353$$

tedy v rozmezí 135 cm až 140 cm bylo asi 35,3 % hochů

pohodlnější možnost

- $X \sim \mathrm{N}(136{,}1, 6{,}4^2)$
- počítáme $\mathrm{P}(134{,}5 < X < 140{,}5)$
- Excel i R nabízejí možnost dosadit skutečné parametry normálního rozdělení
- druhým parametrem je **směrodatná odchylka**
- Excel (nepřehlédněte, že nejde o NORMSDIST!): [NORMDIST(140,5;136,1;6,4;1)-NORMDIST(134,5;136,1;6,4;1)]
- R: [pnorm(140.5,136.1,6.4)-pnorm(134.5,136.1,6.4)]