

Matematika IV – 8. přednáška

Náhodné veličiny – základní vlastnosti a typy

Michal Bulant

Masarykova univerzita
Fakulta informatiky

14. 4. 2008

Obsah přednášky

- 1 Náhodné veličiny
- 2 Typy diskrétních náhodných veličin
- 3 Typy spojitých náhodných veličin

Doporučené zdroje

- Martin Panák, Jan Slovák, **Drsná matematika**, e-text.
- Karel Zvára, Josef Štěpán, **Pravděpodobnost a matematická statistika**, Matfyzpress, 4. vydání, 2006, 230 stran, ISBN 80-867-3271-1.
- Marie Budíková, Štěpán Mikoláš, Pavel Osecký, **Teorie pravděpodobnosti a matematická statistika (sbírka příkladů)**, Masarykova univerzita, 3. vydání, 2004, 117 stran, ISBN 80-210-3313-4.
- Marie Budíková, Štěpán Mikoláš, Pavel Osecký, **Popisná statistika**, Masarykova univerzita, 3. vydání, 2002, 48 stran, ISBN 80-210-1831-3.
- Marie Budíková, Tomáš Lerch, Štěpán Mikoláš, **Základní statistické metody**, Masarykova univerzita, 2005, 170 stran, ISBN 80-210-3886-1.

Na prostoru \mathbb{R}^k uvažujme nejmenší jevové pole \mathcal{B} obsahující všechny k -rozměrné intervaly. Množinám v \mathcal{B} říkáme **borelovské množiny** (nebo také měřitelné množiny) na \mathbb{R}^k .

Speciálně pro $k = 1$ jde o množiny, které obdržíme z **intervalů konečnými průniky a nejvýše spočetnými sjednoceními**.

Definice

Náhodná veličina X na pravděpodobnostním prostoru (Ω, \mathcal{A}, P) je taková funkce $X : \Omega \rightarrow \mathbb{R}$, že vzor $X^{-1}(B)$ patří do \mathcal{A} pro každou Borelovskou množinu $B \in \mathcal{B}$ na \mathbb{R} (tj. $X : \Omega \rightarrow \mathbb{R}$ je tzv. borelovsky měřitelná).

Množinová funkce

$$P_X(B) = P(X^{-1}(B))$$

se nazývá **rozdělení pravděpodobnosti** náhodné veličiny X .

Náhodný vektor (X_1, \dots, X_k) na (Ω, \mathcal{A}, P) je k -tice náhodných veličin.

Definice náhodné veličiny zajišťuje, že pro všechny $-\infty \leq a \leq b \leq \infty$ existuje pravděpodobnost $P(a < X \leq b)$, kde používáme stručné značení pro jev $A = (\omega \in \Omega; a < X(\omega) \leq b)$.

Definice

Distribuční funkcí (*distribution, cumulative density function*) náhodné veličiny X je funkce $F : \mathbb{R} \rightarrow \mathbb{R}$ definovaná pro všechny $x \in \mathbb{R}$ vztahem

$$F(x) = P(X \leq x).$$

Distribuční funkcí náhodného vektoru (X_1, \dots, X_k) je funkce $F : \mathbb{R}^k \rightarrow \mathbb{R}$ definovaná pro všechny $(x_1, \dots, x_k) \in \mathbb{R}^k$ vztahem

$$F(x) = P(X_1 \leq x_1 \wedge \dots \wedge X_k \leq x_k).$$

Diskrétní náhodné veličiny

Předpokládejme, že náhodná veličina X na pravděpodobnostním prostoru (Ω, \mathcal{A}, P) nabývá jen konečně mnoha hodnot $x_1, x_2, \dots, x_n \in \mathbb{R}$. Pak existuje tzv. **pravděpodobnostní funkce** $f(x)$ taková, že

$$f(x) = \begin{cases} P(X = x_i) & \text{pro } x = x_i \\ 0 & \text{jinak.} \end{cases}$$

Evidentně $\sum_1^n f(x_i) = 1$.

Takové náhodné veličině se říká **diskrétní**.

Každá náhodná veličina definovaná pro klasickou pravděpodobnost je diskrétní. Obdobně lze definici pravděpodobnostní funkce rozšířit na veličiny se spočetně mnoha hodnotami (pracujeme pak s nekonečnými řadami)

Spojité náhodné veličiny

I když hodnoty náhodné veličiny X nejsou diskrétní, můžeme postupovat podobně s užitím ideí diferenciálního a integrálního počtu. Intuitivně lze uvažovat takto: **hustotu** $f(x)$ **pravděpodobnosti** pro X si představíme jako

$$P(x < X \leq x + dx) = f(x)dx.$$

To znamená, že chceme pro $-\infty \leq a \leq b \leq \infty$

$$P(a < X \leq b) = \int_a^b f(x)dx. \quad (*)$$

Definice

Náhodná veličina X , pro kterou existuje její **hustota pravděpodobnosti** splňující (*), se nazývá **spojitá**.

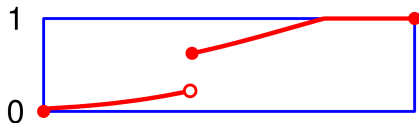
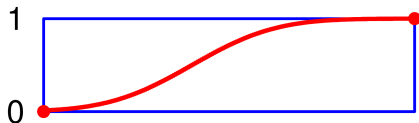
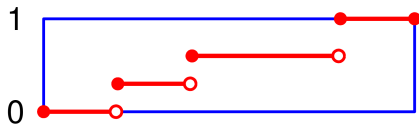
Vlastnosti distribuční funkce

Věta

Nechť X je náhodná veličina, $F(x)$ je její distribuční funkce.

- 1 *F je neklesající.*
- 2 *F je zprava spojitá, $\lim_{x \rightarrow -\infty} F(x) = 0$ a $\lim_{x \rightarrow \infty} F(x) = 1$.*
- 3 *Je-li X diskrétní s hodnotami x_1, \dots, x_n , pak je $F(x)$ po částech konstantní, $F(x) = \sum_{x_i \leq x} P(X = x_i)$ a $F(x) = 1$ kdykoliv $x \geq x_n$.*
- 4 *Je-li X spojitá, pak je $F(x)$ diferencovatelná a její derivace se rovná hustotě X , tj. platí $F'(x) = f(x)$.*

Distribuční funkce



Obdobně definujeme distribuční funkce a hustotu a pravděpodobnostní funkci pro spojité a diskrétní náhodné **vektory**. Hovoříme také o **simultánních pravděpodobnostních funkcích a hustotách**.

Pro dvě proměnné (vektor (X, Y) náhodných veličin):

$$f(x, y) = \begin{cases} P(X = x_i \wedge Y = y_j) & x = x_i \wedge y = y_j \\ 0 & \text{jinak.} \end{cases}$$

u diskrétních a pro všechny $a, b \in \mathbb{R}$ pro spojité:

$$P(-\infty < X \leq b, -\infty < Y \leq a) = \int_{-\infty}^a \int_{-\infty}^b f(x, y) dx dy.$$

Marginální rozložení pro jednu z proměnných obdržíme tak, že přes ostatní počítáme nebo zintegrujeme.

Náhodné veličiny X a Y jsou **stochasticky nezávislé**, jestliže je jejich simultánní distribuční funkce

$$F(x, y) = G(x) \cdot H(y)$$

kde F a G jsou distribuční funkce veličin X a Y .

Alternativní rozdělení popisuje pokus se dvěma možnými výsledky, často nazývanými *zdar*, resp. *nezdar*. Náhodná veličina $X \sim A(p)$ nabývá hodnoty 1 (*zdar*) s pravděpodobností p . Distribuční a pravděpodobnostní funkce jsou tedy tvaru:

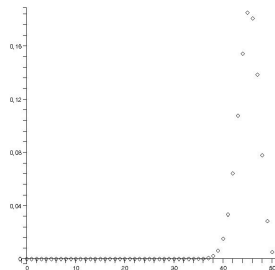
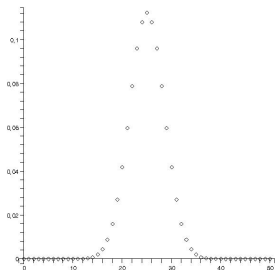
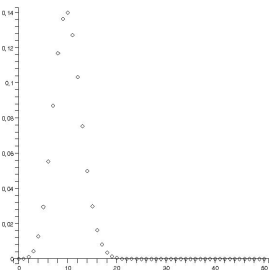
$$F_X(t) = \begin{cases} 0 & t < 0 \\ 1 - p & 0 \leq t < 1 \\ 1 & t > 1 \end{cases} \quad f_X(t) = \begin{cases} p & t = 1 \\ 1 - p & t = 0 \\ 0 & \text{jinak} \end{cases}.$$

Binomické rozdělení $\text{Bi}(n, p)$ odpovídá n -krát nezávisle opakovanému pokusu popsanému alternativním rozdělením, přičemž naše náhodná veličina měří počet zdarů. Je tedy

$$f_X(t) = \begin{cases} \binom{n}{t} p^t (1-p)^{1-t} & t \in \{0, 1, \dots, n\} \\ 0 & \text{jinak} \end{cases}.$$

Binomické rozdělení

Na obrázku jsou pravděpodobnostní funkce pro $\text{Bi}(50, 0.2)$, $\text{Bi}(50, 0.5)$ a $\text{Bi}(50, 0.9)$. Rozdělení pravděpodobnosti dobře odpovídá intuici, že nejvíce výsledků bude blízko u hodnoty np :



Binomické rozdělení

S binomickým rozdělením se setkáváme velice často v praktických úlohách. Jednou z nich je popis náhodné veličiny, která popisuje počet X předmětů v jedné zvolené přihrádce z n možných, do nichž jsme náhodně rozdělili r předmětů. Umístění kteréhokoliv předmětu do pevně zvolené přihrádky má pravděpodobnost $1/n$ (každá z nich je stejně pravděpodobná). Zjevně tedy bude pro jakýkoliv počet $k = 0, \dots, r$

$$P(X = k) = \binom{r}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{r-k} = \binom{r}{k} \frac{(n-1)^{r-k}}{n^r},$$

jde proto o rozložení X typu $\text{Bi}(r, 1/n)$.

Binomické → Poissonovo rozdělení

Jestliže nám bude vzrůstat počet přihrádek n společně s počtem předmětů r_n tak, že v průměru nám na každou přihrádku bude připadat (přibližně) stejný počet prvků λ , můžeme dobře vyjádřit chování našeho rozdělení veličin X_n při limitním přechodu $n \rightarrow \infty$. Takovéto chování popisuje např. fyzikální soustavy s velikým počtem molekul plynu. Standardní úpravy vedou při $\lim_{n \rightarrow \infty} r_n/n = \lambda$ k výsledku:

$$\begin{aligned}\lim_{n \rightarrow \infty} P(X_n = k) &= \lim_{n \rightarrow \infty} \binom{r_n}{k} \frac{(n-1)^{r_n-k}}{n^{r_n}} \\ &= \lim_{n \rightarrow \infty} \frac{r_n(r_n-1)\dots(r_n-k+1)}{(n-1)^k} \frac{1}{k!} \left(1 - \frac{1}{n}\right)^{r_n} \\ &= \frac{\lambda^k}{k!} \lim_{n \rightarrow \infty} \left(1 + \frac{-\frac{r_n}{n}}{r_n}\right)^{r_n} = \frac{\lambda^k}{k!} e^{-\lambda}\end{aligned}$$

protože obecně funkce $(1 + x/n)^n$ konvergují stejnoměrně k funkci e^x na každém omezeném intervalu v \mathbb{R} .

Poissonovo rozdělení $Po(\lambda)$

Poissonovo rozdělení popisuje náhodné veličiny s pravděpodobnostní funkcí

$$f_X(t) = \begin{cases} \frac{\lambda^k}{k!} e^{-\lambda} & t \in \mathbb{N} \\ 0 & \text{jinak.} \end{cases}$$

Jak jsme odvodili výše, toto diskrétní rozdělení (rozložené do nekonečně mnoha bodů) dobře aproximuje binomická rozdělení $Bi(n, \lambda/n)$ pro konstantní $\lambda > 0$ a velká n .
Snadno ověříme

$$\sum_{k=0}^{\infty} f_X(k) = \sum_k \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_k \frac{\lambda^k}{k!} = e^{-\lambda + \lambda} = 1.$$

Poissonovo rozdělení

Dobře modeluje výskyt jevů:

- s očekávanou konstantní hustotou na jednotku objemu – např. bakterie ve vzorku (popis očekávaného výskytu k bakterií při rozdělení vzorku na n stejných částí)
- rozdělení událostí, které se vyskytují náhodně v čase a bez závislosti na předchozí historii – v praxi jsou takové procesy často spojeny s poruchovostí strujů a zařízení

Geometrické rozdělení má náhodná veličina $X \sim \text{Ge}(p)$, která udává celkový počet *nezdarů*, které v posloupnosti opakovaných pokusů předcházejí prvnímu *zdaru*, přičemž pravděpodobnost úspěchu v každém pokusu je rovna p .

$$f_X(t) = \begin{cases} (1-p)^t \cdot p & \text{pro } t = 0, 1, \dots \\ 0 & \text{jinak.} \end{cases}$$

Hypergeometrické rozdělení. Mějme N předmětů, z nichž právě M má danou vlastnost. Z těchto N předmětů náhodně vybereme n předmětů bez vracení. Náhodná veličina $X \sim \text{Hg}(N, M, n)$ udává počet vybraných prvků s danou vlastností. Zřejmě tato náhodná veličina může nabývat pouze celočíselných hodnot z intervalu $[\max\{0, M - N + n\}, \min\{n, M\}]$. Pro t z tohoto intervalu pak

$$f_X(t) = \frac{\binom{M}{t} \binom{N-M}{n-t}}{\binom{N}{n}}.$$

Rovnoměrné spojité rozdělení $R_s(a, b)$ je nejjednodušším příkladem spojitého rozdělení. Ilustruje, že při jednoduše formulovaném požadavku na chování rozdělení nám nezbude moc prostoru pro jeho definici. Nyní chceme, aby pravděpodobnost každé hodnoty v předem daném intervalu $(a, b) \subset \mathbb{R}$ byla stejná, tj. hustota f_X našeho rozdělení náhodné veličiny X má být konstantní. Pak ovšem jsou pro libovolná reálná čísla $-\infty < a < b < \infty$ jen jediné možné hodnoty

$$f_X(t) = \begin{cases} 0 & t \leq a \\ \frac{1}{b-a} & t \in (a, b) \\ 0 & t \geq b, \end{cases} \quad F_X(t) = \begin{cases} 0 & t \leq a \\ \frac{t-a}{b-a} & t \in (a, b) \\ 1 & t \geq b. \end{cases}$$

Exponenciální rozdělení $\text{ex}(\lambda)$ je dalším rozdělením, které je snadno určeno požadovanými vlastnostmi náhodné veličiny. Předpokládejme, že sledujeme náhodný jev, jehož výskyty v nepřekrývajících se časových intervalech jsou nezávislé. Je-li tedy $P(t)$ pravděpodobnost, že jev nenastane během intervalu délky t , pak nutně $P(t + s) = P(t)P(s)$ pro všechna $t, s > 0$. Předpokládejme navíc diferencovatelnost funkce P a $P(0) = 1$. Pak jistě $\ln P(t + s) = \ln P(t) + \ln P(s)$, takže limitním přechodem

$$\lim_{s \rightarrow 0_+} \frac{\ln P(t + s) - \ln P(t)}{s} = (\ln P)'_+(0).$$

Označme si spočtenou derivaci zprava v nule jako $-\lambda \in \mathbb{R}$. Pak tedy pro $P(t)$ platí $\ln P(t) = -\lambda t + C$ a počáteční podmínka dává jediné řešení

$$P(t) = e^{-\lambda t}.$$

Všimněme si, že z definice našich objektů vyplývá, že $\lambda > 0$.

Nyní uvažme náhodnou veličinu X udávající (náhodný) okamžik, kdy náš jev poprvé nastane. Zřejmě tedy je distribuční funkce rozdělení pro X dána

$$F_X(t) = 1 - P(t) = \begin{cases} 1 - e^{-\lambda t} & t > 0 \\ 0 & t \leq 0. \end{cases}$$

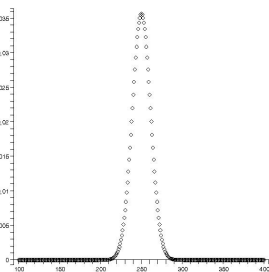
Je vidět, že skutečně jde rostoucí funkci s hodnotami mezi nulou a jedničkou a správnými limitami v $\pm\infty$.
Hustotu tohoto rozdělení dostaneme derivováním distribuční funkce, tj.

$$f_X = \begin{cases} \lambda e^{-\lambda t} & t > 0 \\ 0 & t \leq 0. \end{cases}$$

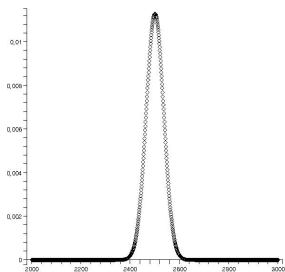
Normální rozdělení

Jde o nejdůležitější rozdělení. Uved' me nejprve motivaci pro jeho zavedení.

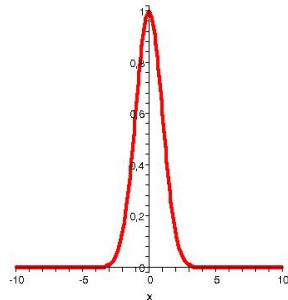
Pokud budeme v **binomickém rozdělení** $\text{Bi}(n, p)$ zvyšovat n při zachování úspěšnosti p , bude mít pravděpodobnostní funkce pořád přibližně stejný tvar.



$\text{Bi}(500, 0.5)$



$\text{Bi}(5000, 0.5)$



graf funkce $e^{-x^2/2}$

Normální rozdělení $N(0, 1)$

Vzhledem k uvedené motivaci se nabízí hledat vhodné spojité rozdělení, které by mělo hustotu danou nějakou obdobnou funkcí. Protože je $e^{-x^2/2}$ vždy kladná funkce, potřebovali bychom spočítat $\int_a^b e^{-x^2/2} dx$ což není pomocí elementárních funkcí možné. Je však možné (i když ne úplně snadné) ověřit, že příslušný nevlastní integrál konverguje k hodnotě

$$\int_{-\infty}^{\infty} e^{-x^2/2} dx = \sqrt{2\pi}.$$

Odtud vyplývá, že možná hustota rozdělení náhodného rozdělení může být

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Rozdělení s touto hustotou se nazývá **normální rozdělení** $N(0, 1)$.

Normální rozdělení $N(0, 1)$

Příslušnou distribuční funkci

$$F_X(x) = \int_{-\infty}^x e^{-x^2/2} dx$$

nelze vyjádřit pomocí elementárních funkcí, přesto se s ní numericky běžně počítá (pomocí tabulek nebo softwarových aplikací).

Hustotě f_X se také často říká **Gaussova křivka**.

Abychom uměli pořádněji sformulovat asymptotickou blízkost normálního a binomického rozdělení pro $n \rightarrow \infty$, musíme si vytvořit další nástroje pro práci s náhodnými veličinami. Budeme k tomu používat funkce dvojným způsobem.

Příklad

Nechť veličina náhodná veličina X má rovnoměrné rozdělení na intervalu $\langle 0, r \rangle$. Určete distribuční funkci a hustotu pravděpodobnosti rozdělení objemu koule o poloměru X .

Řešení

Určeme nejprve distribuční funkci F (pro $0 < d < \frac{4}{3}\pi r^3$)

$$F(d) = P \left[\frac{4}{3}\pi X^3 \leq d \right] = P \left[X \leq \sqrt[3]{\frac{3d}{4\pi}} \right] = \frac{\sqrt[3]{\frac{3d}{4\pi}}}{r},$$

celkem

$$F(x) = \begin{cases} 0 & \text{pro } x \leq 0 \\ \sqrt[3]{\frac{3}{4\pi r^3}} x^{\frac{1}{3}} & \text{pro } 0 < x < \frac{4}{3}\pi r^3 \\ 1 & \text{pro } x \geq \frac{4}{3}\pi r^3 \end{cases}$$

Derivováním pak obdržíme hustotu pravděpodobnosti.