Matematika IV – 8. přednáška
Náhodné veličiny – základní vlastnosti a typy

Michal Bulant

Masarykova univerzita
Fakulta informatiky

10. 4. 2013

Obsah přednášky

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, **Statistika**, Masarykova univerzita, 2004, distanční studijní opora ESF, `http://www.math.muni.cz/~budikova/esf/Statistika.zip`.
- Marie Budíková, Tomáš Lerch, Štěpán Mikoláš, **Základní statistické metody**, Masarykova univerzita, 2005, 170 stran, ISBN 80-210-3886-1.

Vraťme se k jednoduchému a názornému příkladu statistik kolem výsledků studentů v daném předmětu, který je a není podobný klasické pravděpodobnosti a s ní související statistice při házení kostkou.

Na jedné straně jsme připustili pouze konečný počet možných bodových hodnocení (v tomto případě celá čísla od 0 do 40), zároveň ale není patrně vhodné představovat si výsledky jednotlivých studentů jako analogii nezávislého házení kostkou (to by byla skutečně divně vedená přednáška).

Místo toho máme na základním prostoru Ω všech studentů definovánu funkci bodového ohodnocení $X : \Omega \to \mathbb{R}$. Je to typický příklad **náhodné veličiny**.

U každé náhodné veličiny potřebujeme umět pracovat s vhodnou množinou jevů. Zpravidla požadujeme, abychom mohli pracovat s pravděpodobnostmi příslušnosti hodnoty X do předem zadaného intervalu.

Přirozenější interpretací výsledku pokusu je totiž často spíše než zjištění, zda náhodný jev *nastal či nenastal*, nějaká hodnota:

- součet bodů na dvou kostkách,
- počet bakterií v daném množství roztoku nebo
- počet studentů, kteří uspěli u zkoušky nebo kteří získali alespoň 5 bodů z konkrétního příkladu.

Od pravděpodobnostního prostoru $(\Omega, \mathcal{A}, P)$ tedy potřebujeme přejít k obdobné dvojici $(\mathbb{R}, \mathcal{B})$ tak, abychom podmnožinám $\mathbb{R}$, ležícím v σ-algebře $\mathcal{B}$ byli schopni přiřadit pravděpodobnost odvozenou z $(\Omega, \mathcal{A}, P)$.

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 $(\Omega, \mathcal{A}, P)$ je taková funkce $X : \Omega \to \mathbb{R}$, že vzor $X^{-1}(B)$ patří do $\mathcal{A}$ pro každou Borelovskou množinu $B \in \mathcal{B}$ na $\mathbb{R}$ (tj. $X : \Omega \to \mathbb{R}$ 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 $(\Omega, \mathcal{A}, P)$ je k–tice náhodných veličin.

Příklad

Hodíme jedenkrát kostkou, množina elementárních jevů je $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$. Jevovým polem nechť je $\mathcal{A} = \{\emptyset, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4, \omega_5, \omega_6\}, \Omega\}$.
Zjistěte jestli zobrazení $X : \Omega \to \mathbb{R}$ dané předpisem

a) $X(\omega_i) = i$ pro každé $i \in \{1, 2, 3, 4, 5, 6\}$,

b) $X(\omega_1) = X(\omega_2) = -2, X(\omega_3) = X(\omega_4) = X(\omega_5) = X(\omega_6) = 3$

je náhodnou veličinou vzhledem k $\mathcal{A}$.

Příklad

Je dáno jevové pole $(\Omega, \mathcal{A})$, kde $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$ a

$$\mathcal{A} = \{\emptyset, \{\omega_1, \omega_2\}, \{\omega_3\}, \{\omega_4, \omega_5\}, \{\omega_1, \omega_2, \omega_3\},$$
$$\{\omega_1, \omega_2, \omega_4, \omega_5\}, \{\omega_3, \omega_4, \omega_5\}, \Omega\}.$$

Najděte nějaké (co nejobecnější) zobrazení $X : \Omega \to \mathbb{R}$, které bude náhodnou veličinou vzhledem k $\mathcal{A}$.

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 $F : \mathbb{R} \to \mathbb{R}$ definovaná pro všechny $x \in \mathbb{R}$ vztahem

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

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

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

Diskrétní náhodné veličiny

Předpokládejme, že náhodná veličina X na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$ nabývá jen konečně mnoha hodnot $x_1, x_2, \ldots, x_n \in \mathbb{R}$. 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 $\sum_{i=1}^{n} f(x_i) = 1$.
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 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 $-\infty \leq a \leq b \leq \infty$

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

Definice

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

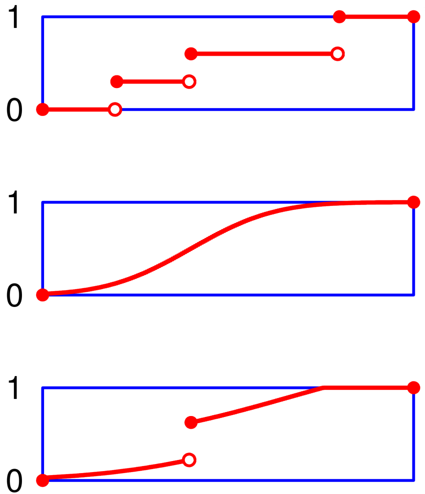
Vlastnosti distribuční funkce

Věta

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

1. *F je neklesající.*
2. *F je zprava spojitá, $\lim_{x\to-\infty} F(x) = 0$ a $\lim_{x\to\infty} F(x) = 1$.*
3. *Je-li X diskrétní s hodnotami $x_1, \ldots, x_n$, 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 - příklady



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):

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

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

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

Marginální rozložení pro jednu z proměnných obdržíme tak, že přes ostatní posčítáme nebo zintegrujeme.
Náhodné veličiny X a Y jsou **stochasticky nezávislé**, jestliže je jejich simultánní distribuční funkce splňují

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

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

Rovnoměrné (diskrétní) rozdělení popisuje náhodnou veličinu, která může nabývat konečně mnoha hodnot se stejnou pravděpodobností, značíme $X \sim Rd(n)$ (např. $X \sim Rd(6)$ odpovídá hodu kostkou).

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

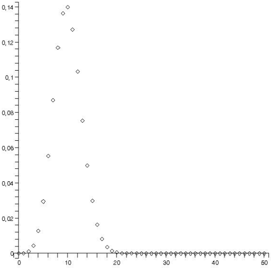
$$F_X(t) = \begin{cases} 0 & t < 0 \\ 1 - p & 0 \leq t < 1 \\ 1 & t > 1 \end{cases} \qquad f_X(t) = \begin{cases} p & t = 1 \\ 1 - p & t = 0 \\ 0 & \text{jinak} \end{cases}.$$

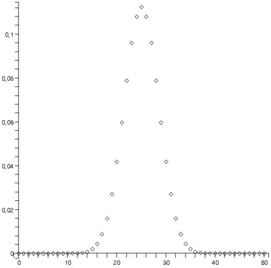
Binomické rozdělení Bi(n, p) odpovídá n–krát nezávisle opakovanému pokusu popsanému alternativním rozdělením, přičemž naše náhodná veličina měří počet zdarů. Je tedy

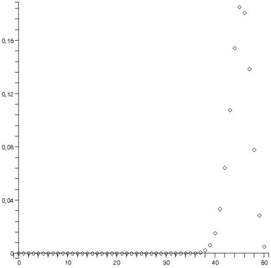
$$f_X(t) = \begin{cases} \binom{n}{t} p^t (1 - p)^{n-t} & t \in \{0, 1, \dots, n\} \\ 0 & \text{jinak} \end{cases}.$$

Binomické rozdělení

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







Binomické rozdělení

S binomickým rozdělením se potkáváme 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 $\mathrm{Bi}(r, 1/n)$.

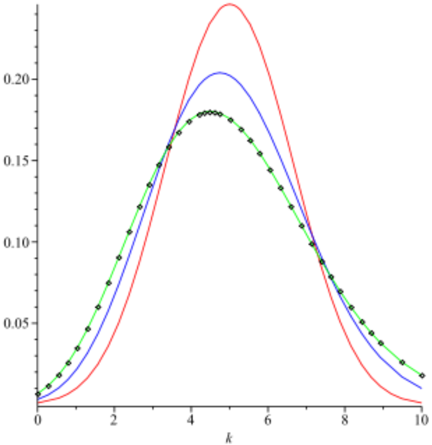
Binomické $\to$ Poissonovo rozdělení

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 $n \to \infty$. Takovéto chování popisuje např. fyzikální soustavy s velikým počtem molekul plynu. Standardní úpravy vedou při $\lim_{n\to\infty} r_n/n = \lambda$ k výsledku:

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

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

Binomické → Poissonovo rozdělení



Tečky znázorňují Poissonovo rozdělení, červeně Bi$(10, \frac{1}{2})$, modře Bi$(20, \frac{1}{4})$ a zeleně Bi$(1000, \frac{1}{200})$

Poissonovo rozdělení Po(λ)

Poissonovo rozdělení popisuje náhodné veličiny s pravděpodobnostní funkcí

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

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

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

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 strojů a zařízení

Příklad

- počet branek ve fotbalovém zápase (za 90 minut)
- počet telefonních hovorů za minutu na call centru
- počet aut přijíždějících na křižovatku
- ...

Geometrické rozdělení

Geometrické rozdělení má náhodná veličina $X \sim \mathrm{Ge}(p)$, 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 $X \sim \mathrm{Hg}(N, M, n)$ udává počet vybraných prvků s danou vlastností. Zřejmě tato náhodná velišina může nabývat pouze celočíselných hodnot z intervalu $[\max\{0, M-N+n\}, \min\{n, M\}]$. Pro t z tohoto intervalu pak

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