Matematika IV – 9. přednáška
Náhodné veličiny a jejich transformace

Michal Bulant

Masarykova univerzita
Fakulta informatiky

17. 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.

Typy spojitých náhodných veličin

Rovnoměrné spojité rozdělení $\mathrm{Rs}(a, b)$ 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

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

Exponenciální rozdělení $\mathrm{Ex}(\lambda)$ 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

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

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 (je vidět analogie s geometrickým rozdělením?). Zřejmě tedy je distribuční funkce rozdělení pro X dána

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

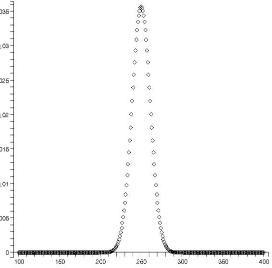
Je vidět, že skutečně jde rostoucí funkci s hodnotami mezi nulou a jedničkou a správnými limitami v $\pm\infty$.
Hustotu tohoto rozdělení dostaneme derivováním distribuční funkce, tj.

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

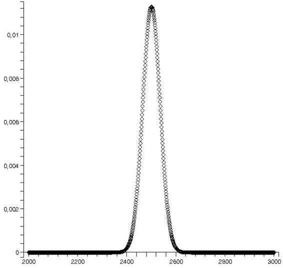
Normální rozdělení

Jde o nejdůležitější rozdělení. Uveďme nejprve motivaci pro jeho zavedení.

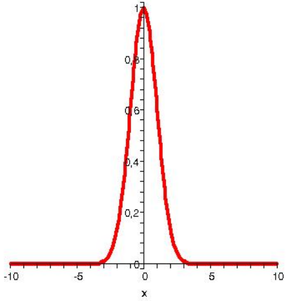
Pokud budeme v **binomickém rozdělení** $\mathrm{Bi}(n, p)$ zvyšovat n při zachování úspěšnosti p, bude mít pravděpodobnostní funkce pořád přibližně stejný tvar.



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



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



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 $\mathrm{e}^{-x^2/2}$ vždy kladná funkce, potřebovali bychom spočíst $\int_a^b \mathrm{e}^{-x^2/2}\,dx$ což není pomocí elementárních funkcí možné. Je však možné (i když ne úplně snadné) ověřit, že příslušný nevlastní integrál konverguje k hodnotě

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

Odtud vyplývá, že hustota rozdělení náhodné veličiny může být

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

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

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 přesněji zformulovat asymptotickou blízkost normálního a binomického rozdělení pro $n \to \infty$, musíme si vytvořit další nástroje pro práci s náhodnými veličinami. Budeme k tomu používat funkce dvojím různým způsobem.

Příklad

Nechť má náhodná veličina X 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

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

Derivováním pak obdržíme hustotu pravděpodobnosti.

Příklad (rozdělení $\chi^2(1)$)

Nechť Z má normované normální rozdělení. Určete hustotu transformované náhodné veličiny $X = Z^2$.

Řešení

Zřejmě je pro $x \leq 0$ distribuční funkce nulová, pro $x > 0$ dostáváme: $F_X(x) = P[Z^2 < x] = P[-\sqrt{x} < Z < \sqrt{x}] =$

$$= \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{z^2}{2}} \,\mathrm{d}z = 2\int_0^x \frac{1}{2\sqrt{2\pi}} t^{-\frac{1}{2}} \mathrm{e}^{-\frac{t}{2}} \,\mathrm{d}t$$

a derivací podle x dostaneme hustotu $f_X(x) = \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} \mathrm{e}^{-\frac{x}{2}}$.
Rozdělení náhodné veličiny s touto hustotou se nazývá (Pearsonovo) χ^2 rozdělení s jedním stupněm volnosti a značí se $X \sim \chi^2(1)$.

Transformace náhodných veličin

Místo náhodné veličiny X, např. „roční plat zaměstnance“, budeme vyčíslovat jinou závislou hodnotu $\psi(X)$, např. „roční čistý příjem zaměstnance po zdanění a včetně sociálních dávek“.
V systému se značnou sociální solidaritou je první veličina hodně variabilní, zatímco druhá může být skoro konstantní. Statisticky se proto budou značně odlišovat.
Připomeňme si přechod od binomického k Poissonovu rozdělení:

Věta (Poissonova)

Je-li $X_n \sim \mathrm{Bi}(n, p_n)$ *taková, že* $\lim_{n\to\infty} np_n = \lambda$ *a* $X \sim \mathrm{Po}(\lambda)$, *pak*

$$\lim_{n\to\infty} P[X_n = k] = P[X = k]$$

pro $k = 0, 1, \ldots$.

Nejjednodušší funkcí, po konstantách, je afinní závislost $\psi(x) = a + bx$.

V případě afinní závislosti $x = \frac{1}{b}(y - a)$ je proto pravděpodobnostní funkce nenulová právě v bodech $y_i = ax_i + b$.

Ukážeme si, že v případě rozdělení X_n typu $\mathrm{Bi}(n, p)$ převádí transformace $x = y\sqrt{np(1-p)} + np$ náhodnou veličinu X_n na rozdělení Y_n s distribuční funkcí blízkou distribuční funkci spojitého rozdělení $N(0, 1)$.

Dříve uvedená Poissonova věta popisuje asymptotické chování binomického rozdělení pří $n \to \infty$ a $p \to 0$, následující věta pak chování v případě konstantní pravděpodobnosti zdaru p.

Věta (de Moivre-Laplaceova)

Pro náhodné veličiny X_n s rozdělením $\mathrm{Bi}(n, p)$ platí

$$\lim_{n\to\infty} P\left[a < \frac{X_n - np}{\sqrt{np(1-p)}} < b\right] = \Phi(b) - \Phi(a),$$

kde Φ je distribuční funkce normovaného normálního rozdělení.

Příklad

Hodíme kostkou celkem 12 000 krát. Určete pravděpodobnost toho, že počet hozených šestek je mezi 1 800 a 2 100.

Řešení

Přesná pravděpodobnost je dána výrazem $\sum_{k=1800}^{2100} \binom{12000}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{12000-k}$, což je obtížně vyčíslitelné. Využijeme tvrzení Moivre-Laplaceovy věty, přepsaného do tvaru

$$P\left[A < X_n < B\right] - \left(\Phi\left(\frac{B - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{A - np}{\sqrt{np(1-p)}}\right)\right) \to 0$$

pro $n \to 0$.

Řešení (pokr.)

Volbou $p = 1/6$, $A = 1800$, $B = 2100$, $n = 12000$ dostáváme odhad

$$P \approx \Phi\left(\frac{2100 - 2000}{\sqrt{12000 \cdot \frac{1}{6}\frac{5}{6}}}\right) - \Phi\left(\frac{1800 - 2000}{\sqrt{12000 \cdot \frac{1}{6}\frac{5}{6}}}\right) =$$
$$= \Phi(\sqrt{6}) - \Phi(-2\sqrt{6}) \approx 0,992.$$

Poznámka

Statistické tabulky – viz např. `https://is.muni.cz/auth/el/1433/jaro2013/MB104/um/StatTab.pdf` nebo sbírka příkladů [BMO].

Příklad

Pravděpodobnost narození chlapce je 0,515. Jaká je pravděpodobnost, že mezi tisíci novorozenci bude alespoň tolik děvčat jako chlapců?

Příklad

Nezávisle opakujeme pokus s výsledky 1 a 0, které mají **neznámé** pravděpodobnosti p a $1-p$. Parametr p chceme odhadnout pomocí *relativních četností* X_n/n (X_n je počet jedniček při n pokusech). Víme, že je $X_n \sim \mathrm{Bi}(n,p)$, proto nám Moivre-Laplaceova věta umožní určit počet pokusů n potřebný k zajištění požadované přesnosti odhadu δ se spolehlivostí $1-\beta$.

Řešení

Využijeme Moivre-Laplaceovu větu zapsanou ve tvaru

$$0 = \lim_{n\to\infty}\left| P\left[\left|\frac{X_n}{n} - p\right| < \delta\right] - \left(\Phi\left(\frac{n\delta}{\sqrt{np(1-p)}}\right) - \Phi\left(-\frac{n\delta}{\sqrt{np(1-p)}}\right)\right)\right|$$

Řešení

Hledáme nejmenší n, splňující nerovnost $P[|X_n/n - p| < \delta] \geq 1 - \beta$, kterou můžeme podle věty aproximovat nerovností

$$\Phi\left(\frac{n\delta}{\sqrt{np(1-p)}}\right) - \Phi\left(-\frac{n\delta}{\sqrt{np(1-p)}}\right) =$$
$$= 2\Phi\left(\frac{n\delta}{\sqrt{np(1-p)}}\right) - 1 \geq 1 - \beta.$$

Ta je ekvivalentní s podmínkou $n\delta/\sqrt{np(1-p)} \geq z(\beta/2)$, kde $z(p)$ je řešení rovnice $\Phi(z(p)) = 1 - p$ (tzv. *kritická hodnota* normovaného normálního rozdělení). Pro $\delta = 0{,}05$ a $1 - \beta = 0{,}9$ máme z tabulek $z(\beta/2) \approx 1{,}645$ a s využitím zřejmého odhadu $p(1-p) \leq 1/4$ dostáváme $n \geq (z(\beta/2)/2\delta)^2 \approx 270{,}6$.

Transformace normálně rozdělené veličiny

Podobně zkusme opačnou transformaci provést na veličinu Y s normálním rozdělením $N(0,1)$. Pro pevně zvolená čísla $\mu, \sigma \in \mathbb{R}$, $\sigma > 0$ spočtěme rozdělení náhodné veličiny $Z = \mu + \sigma Y$. Dostáváme distribuční funkci

$$\begin{aligned} F_Z(z) &= P(Z < z) = P(\mu + \sigma Y < z) \\ &= F_Y\left(\frac{z-\mu}{\sigma}\right) = \int_{-\infty}^{\frac{z-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-t^2/2}\, dt \\ &= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}\sigma}\, \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\, \mathrm{d}x, \end{aligned}$$

kde poslední úprava vychází ze substituce $x = \mu + \sigma t$. Hustota naší nové náhodné veličiny Z je proto

$$f_Z = \frac{1}{\sqrt{2\pi}\sigma}\, \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

a takovému rozdělení se říká normální typu $\mathrm{N}(\mu, \sigma)$.

Střední hodnota

Při statistickém zkoumání hodnot náhodných veličin (např. zpracování výsledků nějakého měření) hledáme výpovědi o náhodné veličině pomocí různých z ní odvozených čísel.
Jako nejjednodušší příklad může sloužit **střední hodnota**[1] $E(X)$ náhodné veličiny X, která je definována

$$E(X) = \begin{cases} \sum_i x_i \cdot f_X(x_i) & \text{pro diskrétní veličinu} \\ \int_{-\infty}^{\infty} x \cdot f_X(x)\,\mathrm{d}x & \text{pro spojitou veličinu.} \end{cases}$$

Obecně střední hodnota náhodných veličin nemusí existovat, protože příslušné sumy či integrály nemusí konvergovat.

[1] Často se místo $E(X)$ píše EX.

Střední hodnota transformované náhodné veličiny

Střední hodnotu můžeme přímo vyjádřit také pro funkce $Y = \psi(X)$ náhodné veličiny X. V diskrétním případě můžeme přímo spočíst

$$
\begin{aligned}
E(Y) &= \sum_j y_j P(Y = y_j) \\
&= \sum_j y_j \sum_{\psi(x_i)=y_j} P(X = x_i) \\
&= \sum_i \psi(x_i) P(X = x_i) = \sum_i \psi(x_i) f_X(x_i).
\end{aligned}
$$

Je tedy $E(\psi(X))$ přímo spočítatelná pomocí pravděpodobnostní funkce f_X.
Podobně vyjadřujeme střední hodnotu funkce ze spojité náhodné veličiny:

$$
E(\psi(X)) = \int_{-\infty}^{\infty} \psi(x) f_X(x)\,\mathrm{d}x,
$$

pokud tento integrál absolutně konverguje.

Příklad

Spočtěme střední hodnotu binomického rozdělení.

Řešení

Pro $X \sim \mathrm{Bi}(n, p)$ je

$$
\begin{aligned}
E(X) &= \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} = \\
&= np \sum_{k=1}^{n} \frac{(n-1)!}{(n-k)!(k-1)!} p^{k-1} (1-p)^{n-k} = \\
&= np \sum_{j=0}^{n-1} \frac{(n-1)!}{(n-1-j)!j!} p^{j} (1-p)^{n-1-j} = \\
&= np(p + (1-p))^{n-1} = np.
\end{aligned}
$$

Základní vlastnosti střední hodnoty

Věta

Nechť $a, b \in \mathbb{R}$ a X, Y jsou náhodné veličiny s existující střední hodnotou. Pak

- $E(a) = a$,
- $E(a + bX) = a + bE(X)$,
- $E(X + Y) = E(X) + E(Y)$,
- *jsou-li X a Y* **nezávislé**, *pak $E(XY) = E(X) \cdot E(Y)$.*

Důkazy těchto tvrzení jsou přímočaré, zkuste si je udělat!
Analogická tvrzení platí i pro náhodné vektory.

Příklad

Spočtěme ještě jednou střední hodnotu binomického rozdělení, tentokrát s využitím vlastností střední hodnoty.

Řešení

Vyjádříme počet zdarů v n pokusech jako počet zdarů v jednotlivých pokusech

$$X = \sum_{k=1}^{n} Y_i,$$

přičemž náhodné veličiny Y_k mají všechny alternativní rozdělení $A(p)$. Snadno spočítáme $E(Y_k) = 1 \cdot p + 0 \cdot (1 - p) = p$. Dále víme, že střední hodnota součtu je součtem středních hodnot, proto

$$E(X) = \sum_{k=1}^{n} E(Y_k) = np.$$