Náhodná veličina [1]

[1]Tyto materiály byly vytvořeny za pomoci grantu FRVŠ číslo 1145/2004.

Náhodná veličina

Většina náhodných pokusů má jako výsledky reálná čísla. Budeme tedy dále náhodnou veličinou rozumět proměnnou, která může nabývat různých reálných čísel v závislosti na náhodě. Na tomto místě si připomeňme, že lze rozlišovat náhodnou veličinu diskrétní a spojitou. Je-li náhodná veličina X spojitá, pak může nabývat všech hodnot z konečného nebo nekonečného intervalu. Naopak, náhodnou veličinu X považujeme za diskrétní, nabývá-li konečného nebo spočetného počtu hodnot.

Náhodnou veličinu X lze tedy chápat jako reálnou funkci prvků prostoru elementárních jevů.

Pravděpodobnost toho, že náhodná veličina X dosáhla hodnoty x_i značíme takto

$$P(X = x_i) \ . \tag{1}$$

Umíme-li pro každé reálné x určit pravděpodobnost, že náhodná veličina X nabyde hodnoty menší nebo rovné x, pak známe tzv. **rozdělení pravděpodobnosti** náhodné veličiny X.

Způsob popisu náhodných veličin

Prvním způsobem je zápis prostřednictvím tabulky rozdělení pravděpodobností viz tabulka 1.

Tabulka 1: Tabulka rozdělení pravděpodobností – hod ideální kostkou

x_i	1	2	3	4	5	6
$P(X = x_i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Druhou možností je pomocí polygonu rozdělení pravděpodobností. Další možností popisu rozdělení pravděpodobnosti je prostřednictvím **distribuční funkce**. Ta je definována vztahem:

$$F(x) = P(X \leq x_i) \ . \tag{2}$$

Distribuční funkce je definována na předem daném intervalu. Její základní vlastnosti jsou:

$$\begin{gathered} 0 \leq F(x) \leq 1 \\ F(x_i) \leq F(x_j) \ \ \text{pro každou dvojici čísel} \ \ x_i < x_j \\ \lim_{x \to -\infty} F(x) = F(-\infty) = 0 \\ \lim_{x \to +\infty} F(x) = F(+\infty) = 1 \\ P(a < X \leq b) = F(b) - F(a) \end{gathered} \tag{3}$$

Distribuční funkce $F(x)$ je zprava spojitá a má nejvýš spočetně bodů nespojitosti. Grafu distribuční funkce odpovídá v popisné statistice graf kumulativních relativních četností. Distribuční funkce diskrétní náhodné veličiny je nespojitá. Pro diskrétní náhodnou veličinu platí:

$$F(x_i) = P(X \leq x_i) = \sum_{j \leq i} p_j \tag{4}$$

Pro spojitou náhodnou veličinu, nabývající všech hodnot z intervalu $x \in [a; b]$

$$F(x) = P(X \leq x_i) = \int_a^x f(t)dt \tag{5}$$

Dalším důležitým pojmem je **hustota pravděpodobnosti** (někdy také frekvenční funkce). Jde o funkci která je definována vztahem

$$f(x) = \frac{dF(x)}{dx} = F^{'}(x) \tag{6}$$

Základní vlastnosti hustoty pravděpodobnosti jsou:

$$\begin{gathered} f(x) \geq 0 \\ \lim_{x \to -\infty} f(x)dx = 0 \\ \lim_{x \to +\infty} f(x)dx = 0 \\ \int_a^b f(x)dx = 1 \quad \text{pro} \quad x \in [a; b] \\ P(a < X \leq b) = \int_a^b f(x)dx \end{gathered} \tag{7}$$

Velmi důležitým pojmem ve spojitosti s popisem náhodných veličin je pojem **kvantilu**. α-kvantilem nebo $\alpha \cdot 100\%$-ním kvantilem náhodné veličiny X, která má jisté spojité rozdělení náhodné veličiny s distribuční funkcí $F(x)$ a hustotu pravděpodobnosti f(x), je číslo x_α pro které platí

$$F(x_\alpha) = P(X \leq x_\alpha) = \int_{-\infty}^{x_\alpha} f(x)dx = \alpha \tag{8}$$

Některé kvantily mají speciální názvy např.: medián, dolní kvartil, horní kvartil, první decil, osmý percentil, atd...

Některá rozdělení diskrétních a spojitých náhodných veličin

Bernoulliho rozdělení

Někdy také Alternativní rozdělení. Pomocí tohoto rozdělení lze popsat ty situace, ve kterých může náhodná proměnná nabývat pouze dvou možných hodnot.

Příkladem může být hod ideální mincí. Dalším možným příkladem může být hlasování jedné ze dvou stran bez možnosti zdržet se hlasování. Bez ztráty obecnosti lze uvažovat o dvou možných výsledcích 0 nezdar a 1 zdar.

Bernoulliho rozdělení je definováno pomocí parametru p. Tento parametr lze interpretovat jako pravděpodobnost zdaru. Pravděpodobnostní funkce Bernoulliho rozdělení je definována takto

$$f(x;p) = \begin{cases} (1-p) & \text{pokud} \quad x = 0 \\ p & \text{pokud} \quad x = 1 \end{cases} . \tag{9}$$

Pravděpodobnostní funkci pro Bernoulliho rozdělení lze zapsat ekvivalentně jako:

$$P(X = x) = p^x (1-p)^{(1-x)} . \tag{10}$$

Distribuční funkci tohoto rozdělení pak zapíšeme jako

$$F(x;p) = \begin{cases} (1-p) & \text{pokud} \quad x = 0 \\ 1 & \text{pokud} \quad x = 1 \end{cases} . \tag{11}$$

Střední hodnota Bernoulliho rozdělení je dána hodnotou p, rozptyl pak hodnotou $p(1-p)$. Symbolickým zápisem $X \sim Bern(p)$ nebo $A(p)$.

Binomické rozdělení

Pokud budeme opakovat n-krát určitý pokus při dodržení stejných podmínek, přičemž v každém pokusu bude moci nastat náhodný jev A, se stejnou pravděpodobností p a naopak nenastat s pravděpodobností $1-p$, pak takové schéma pokusů nazýváme Bernoulliho schéma.

Počet realizací jevu A v n nezávislých pokusech Bernoulliho schématu je zřejmě diskrétní náhodnou veličinou s definičním oborem $\{0, 1, \ldots, n\}$. Vzhledem k tomu, že jsou tyto pokusy navzájem nezávislé lze psát:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} . \tag{12}$$

Náhodnou veličinu X mající binomické rozdělení lze vyjádřit jako součet n nezávislých náhodných veličin, které mají alternativní rozdělení se stejným parametrem p :

$$X = X_1 + X_2 + \ldots + X_n . \tag{13}$$

Střední hodnotu lze pak určit jako:

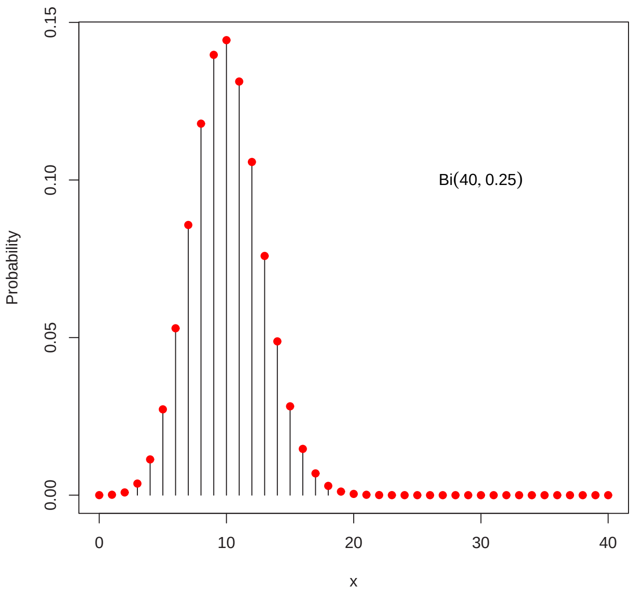
$$E(X) = E(X_1) + E(X_2) + \ldots + E(X_n) = np .$$

Pro rozptyl pak

$$D(X) = D(X_1) + D(X_2) + \ldots + D(X_n) = np(1-p) .$$

Například předpokládejme, že počet x vadných výrobků mezi n nezávisle vyrobenými výrobky má binomické rozdělení $Bi(n, p)$, kde p udává pravděpodobnost, že při výrobním procesu bude vyroben zmetek. Pro $n = 40$ a $p = 0,05$ získáme následující graf pravděpodobnostní funkce viz graf 2.

Obrázek 1: Pravděpodobnostní funkce pro $X \sim Bi(40; 0.05)$



Poissonovo rozdělení

V některých případech není počet dat výsledkem předem stanoveného počtu zkoušek. Například pokud y představuje počet úmrtí při automobilových nehodách v ČR během následujícího týdne, pak teoreticky není stanovena horní hranice n pro y. Vhodný pravděpodobnostní model pak představuje Poissonovo rozdělení.

Poissonovo rozdělení má pouze jeden jediný parametr. Tím je λ. Tento parametr udává jak střední hodnotu tak rozptyl. Skutečnost, že se střední hodnota Poissonova rozdělení musí být shodná s rozptylem je velice důležitá, zvláště při modelování některých typů dat. Poissonova pravděpodobnostní funkce je definována takto

$$f(x; \lambda) = \frac{e^{-\lambda} \lambda^x}{x!} \ . \tag{14}$$

Distribuční funkce pak jako

$$F(x;\lambda) = \sum_{z=0}^{x} \frac{e^{-\lambda}\lambda^z}{z!} \ . \tag{15}$$

Pokud náhodná veličina X sleduje Poissonovo rozdělení s parametrem λ pak píšeme $X \sim Po(\lambda)$.

Jednou z cest jak definovat Poissonovo rozdělení je pomocí aproximace binomickým rozdělením, a to za předpokladu, že n je extrémně vysoké a π je blízko nule. Přesněji pokud $n \to \infty$, $\pi \to 0$, a $n\pi \to \lambda$, pak binomické rozdělení s parametry n a π aproximuje Poissonovo rozdělení s parametrem λ. Neboť lze binomickou pravděpodobnostní funkci zapsat jako

$$\begin{aligned}
\tbinom{n}{x}\pi^x(1-\pi)^{n-x} &= \tfrac{n!}{(n-x)!}\pi^x(1-\pi)^{n-x} \\
&= \tfrac{n(n-1)\cdots(n-x+1)}{x!}\pi^x(1-\pi)^{n-x} \\
&= \tfrac{n(n-1)\cdots(n-x+1)}{x!}\left(\tfrac{n\pi}{n}\right)^x\left(1-\tfrac{n\pi}{n}\right)^{n-x} \\
&= n(n-1)\cdots(n-x+1)\tfrac{1}{x!}\left(\tfrac{n\pi}{n}\right)^x\left(1-\tfrac{n\pi}{n}\right)^n\left(1-\tfrac{n\pi}{n}\right)^{-x} \\
&= n(n-1)\cdots(n-x+1)\tfrac{1}{x!}(n\pi)^x\left(\tfrac{1}{n}\right)^x\left(1-\tfrac{n\pi}{n}\right)^n\left(1-\tfrac{n\pi}{n}\right)^{-x} \\
&= \tfrac{n(n-1)\cdots(n-x+1)}{n^x}(1-\pi)^{-x}\tfrac{(n\pi)^x}{x!}\left(1-\tfrac{n\pi}{n}\right)^n \ .
\end{aligned} \tag{16}$$

A platí

$$\lim_{n\to\infty}\frac{n(n-1)\cdots(n-x+1)}{n^x} = 1 \tag{17}$$

$$\lim_{\pi\to 0}(1-\pi)^{-x} = 1 \ . \tag{18}$$

Dále platí

$$\lim_{\substack{\pi\to\infty \\ n\pi\to\lambda}}\left(1-\frac{n\pi}{n}\right)^n = e^{-\lambda}, \tag{19}$$

a tedy

$$\lim_{n\pi\to\lambda}\frac{(n\pi)^x}{x!} = \frac{\lambda^x}{x!} \ . \tag{20}$$

Čímž je dokázáno to, že pokud $n \to \infty$, $\pi \to 0$ a $n\pi \to \lambda$, pak je pravděpodobnostní funkce binomického rozdělení rovna

$$\frac{e^{-\lambda}\lambda^x}{x!} \ ,$$

což je právě pravděpodobnostní funkce Poissonova rozdělení.

Hypergeometrické rozdělení

Náhodná veličina X má hypergeometrické rozdělení s parametry N, M, n, jestliže má definovanou pravděpodobnostní funkci následujícím způsobem:

$$P(X = x) = \begin{cases} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} & \text{pokud} \quad x \in \langle max(0, M - N + n); min(M, n) \rangle \\ \\ 0 & \text{jinak} . \end{cases} \tag{21}$$

Přičemž N, M, n a x jsou přirozená čísla, pro která platí $\leq M \leq N$ a $1 \leq n \leq N$. Uvědomte si, že faktoriál je definován pouze pro nulu a přirozená čísla.

Význam jednotlivých symbolů lze vysvětlit takto: Mějme N objektů, z nichž M má jistou sledovanou vlastnost. Z takového souboru vybereme náhodně n objektů, přičemž každá n-tice vytvořená z těchto N objektů má stejnou pravděpodobnost že bude vybrána.

Pro malá n/N přibližně pro $n/N \leq 0,1$ lze hypergeometrické rozdělení aproximovat binomickým rozdělením s parametrem $p = M/N$. V případě, že je n/N a M/N malé a n velké, řekněme $n/N \leq 0,1$, $M/N \leq 0,1$ a $n > 30$, lze hypergeometrické rozdělení aproximovat tzv. Poissonovým rozdělením s parametrem $\lambda = nM/N$.

Uveďme si jeden příklad. Jaká je pravděpodobnost, že správně zaškrtneme na lístku SAZKY 3 čísla ze šesti, tj. výhry pátého místa ve sportce?

Postupujme selským rozumem. Jev uhodnutí 3 ze 6 vyhrávajících nastane tehdy, pokud se budou shodovat 3 čísla z 6 vyhrávajících. Takových možných trojic je $\binom{6}{3}$. Ostatní čísla, pak musí být čísla které nevyhrávají, těch je $49 - 6 = 43$. Takových ”nevýtězných” trojic je tedy $\binom{43}{3}$. Neboť každá výtězná trojice může být zkombinována s nevyhrávající, pak počet všech vyhovujících výsledků je $\binom{6}{3} \cdot \binom{43}{3}$. Celkový počet šestic které lze vytvořit z 49 čísel která jsou v osudí je $\binom{49}{6}$. Pravděpodobnost hledané výhry je tedy:

$$P(X = 3) = \frac{\binom{6}{3} \cdot \binom{43}{3}}{\binom{49}{6}} .$$

Připomíná Vám něco tento výsledek? Měl by, neboť:

$$P(X = 3) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = \frac{\binom{6}{3} \cdot \binom{43}{3}}{\binom{49}{6}} .$$

Normální rozdělení

V souvislosti s tímto rozdělením se lze setkat i s názvem Laplace – Gaussovo rozdělení. Patří mezi nejdůležitější spojitá rozdělení náhodných veličin a má zásadní význam jak v statistické teorii, tak i v aplikacích. Lze říci, že tímto rozdělením lze popsat jevy, na jejichž kolísání má vliv velký počet nepatrných a vzájemně nezávislých vlivů.

Normální rozdělení, jak bylo již výše zmíněno patří mezi nejužívanější pravděpodobnostní modely. Tvar distribuční funkce odvodil na základě velkého počtu pokusů hodu mincí francouzský matematik Moivre již v roce 1733. Znovu byla tato křivka objevena na základě chyb měření v astronomii na začátku 19. století, a byla pojmenována po známém německém matematikovi Carlu Friedrichovi Gaussovi (1777-1855). Pojmenování „Normální rozdělení“, pak poprvé zavedl francouzský matematik Quételet.

Hustota pravděpodobnosti tohoto rozdělení je dána funkcí

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \quad \text{pro} \quad x_i \in (-\infty,\ \infty)\ . \tag{22}$$

Jak je z výrazu patrné, má toto rozdělení dva parametry μ a σ^2 Normální rozdělení s těmito parametry se zpravidla značí $N(\mu, \sigma^2)$, kde první parametr je střední hodnotou a druhý je rozptylem náhodné veličiny. Normální rozdělení je symetrické kolem své střední hodnoty, která je současně mediánem i modem.

Pokud bychom hodnoty náhodné veličiny X s normálním rozdělením vhodně transformovali resp. znormovali, pak bychom získali náhodnou veličinu jejíž rozdělení bylo opět normální s jednotkovým rozptylem a nulovou střední hodnotu. Tomuto rozdělení se říká normované normální rozdělení a značíme jej $N(0, 1)$.
Distribuční funkce je stejně jako hustota pravděpodobnosti tabelována právě pro normované normální rozdělení, neboť každé normální rozdělení lze transformovat na normální normované rozdělení. Tabulky hustoty pravděpodobnosti spolu s distribuční funkcí jsou sestaveny většinou pro nezáporné hodnoty normované veličiny U. Kde hodnotu u_i normované veličiny U získáme transformací

$$u_i = \frac{x_i - \mu}{\sigma}\ . \tag{23}$$

Hustotu normovaného normálního rozdělení důsledně označujeme symbolem $\varphi(x)$. Distribuční funkci rozdělení $N(0, 1)$ důsledně označujeme prostřednictvím symbolu $\phi(x)$. Hodnoty pro $x \prec 0$ plynou ze vztahů

$$\varphi(-x) = \varphi(x)\ , \tag{24}$$

$$\phi(-x) = 1 - \phi(x)\ . \tag{25}$$

Dalším velmi důležitým vztahem je předpis

$$u_\alpha = -u_{1-\alpha}\ . \tag{26}$$

Chi–kvadrát rozdělení

Uvažujme navzájem n nezávislých náhodných veličin $U_1, U_2, \cdots, U_n$, z nichž každá má normované normální rozdělení. Potom rozdělení součtu čtverců těchto náhodných veličin má χ^2 rozdělení. Tedy

$$\chi^2 = \sum_{i=1}^{n} U_i^2\ . \tag{27}$$

Součet čtverců n vzájemně nezávislých normovaných normálních náhodných veličin má hustotu pravděpodobnosti danou předpisem

$$f(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n}{2}-1}, & \chi^2 > 0 \\ 0, & \chi^2 \leq 0 \end{cases} \tag{28}$$

Kde funkce $\Gamma(\frac{n}{2})$ se nazývá gama funkce, která je definována jako

$$\begin{aligned} &\Gamma(\tfrac{1}{2}) = 1 \\ &\Gamma(\tfrac{n}{2}) = (\tfrac{n-2}{2})! && \text{pro} \quad n = 2, 4, 6, \cdots \\ &\Gamma(\tfrac{n}{2}) = \tfrac{n-2}{2}\tfrac{n-4}{2}\cdots\tfrac{3}{2}\tfrac{1}{2}\sqrt{\pi} && \text{pro} \quad n = 3, 5, 7, \cdots \end{aligned} \tag{29}$$

Parametr n nazýváme počtem stupňů volnosti. V našem případě mluvíme o χ^2 rozdělení s n stupni volnosti, které značíme $\chi^2(n)$. Distribuční funkce tohoto rozdělení je definována rovnicí

$$F(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} \int_0^{\chi^2} e^{-\frac{t}{2}} t^{-\frac{n}{2}-1} dt, & \chi^2 > 0 \\ 0, & \chi^2 \leq 0 \ . \end{cases} \tag{30}$$

Charakteristiky tohoto rozdělení jsou

$$E(\chi^2) = n \ ,$$

$$D(\chi^2) = 2n \ .$$

Frekvenční funkce χ^2 rozdělení je asymetrická. Její průběh závisí na počtu stupňů volnosti. S rostoucím n se χ^2 rozdělení blíží normálnímu rozdělení $N(n, 2n)$. Pokud $n > 30$ lze toto rozdělení aproximovat normovaným normálním rozdělením.

Studentovo nebo také t-rozdělení

Jedním z nejčastěji využívaným rozdělením je tzv. t-studentovo rozdělení. Lze jej definovat pomocí dvou nezávislých náhodných veličin U a χ^2, které mají po řadě $N(0,1)$ a $\chi^2(n)$ rozdělení. Náhodná veličina t kde ta je definována jako

$$t = \frac{U}{\sqrt{\frac{\chi^2}{n}}} \ , \tag{31}$$

má hustotu pravděpodobnosti

$$f(x; n) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n} \cdot \Gamma(\frac{n}{2})} (1 + \frac{x^2}{n})^{-\frac{n+1}{2}} \quad x \in (-\infty; \infty) \ . \tag{32}$$

Rozdělení s touto hustotou pravděpodobnosti se nazývá t rozdělení, též Studentovo rozdělení o n stupních volnosti. Počet stupňů volnosti veličiny χ^2 ve jmenovateli veličiny t určuje počet stupňů volnosti Studentova rozdělení.

Rozdělení t při rostoucím počtu stupňů volnosti rychle konverguje k normálnímu rozdělení. Pro $n > 30$ lze nahradit Studentovo rozdělení normálním normovaným rozdělením. Studentovo rozdělení je symetrické jednovrcholové. Vzhledem k symetrii platí:

$$t_\alpha(n) = -t_{1-\alpha}(n) \ . \tag{33}$$

Fisherovo–Snedecorovo rozdělení

Dalším hojně využívaným rozdělením je Fisherovo–Snedecorovo rozdělení, známé rovněž jako F-rozdělení. Lze jej definovat prostřednictvím dvou nezávislých náhodných veličin které pocházejí z Chi-kvadrát rozdělení s m resp. n stupni volnosti. Náhodná veličina F je definována takto:

$$F = \frac{\frac{\chi_1^2}{m}}{\frac{\chi_2^2}{n}} \ . \tag{34}$$

Rozdělení s touto hustotou pravděpodobnosti se symbolicky zapisuje jako $F(m,n)$. Uvědomte si, že zde záleží na pořadí stupňů volnosti m, n. Nicméně platí vztah

$$F_\alpha(m,n) = \frac{1}{F_{1-\alpha}(n,m)} \ . \tag{35}$$

Rozdělení F se při velkých počtech stupňů volnosti blíží k rozdělení normálnímu, ale dosti pomalu. Toto rozdělení je asymetrické.