

Náhodná veličina a její rozdělení

Marie Turčičová

18. března 2021

1 Úvod k rozdělením a diskrétní rozdělení

Formální definice pojmu **náhodná veličina** je poměrně složitá, nicméně intuitivně si pod tímto pojmem lze představit nějaký znak, který měříme, popř. pokus, který provádíme. Přívlastek náhodná odpovídá tomu, že předem nevíme, jak tento pokus dopadne. Náhodnou veličinu typicky označujeme velkými písmeny z konce abecedy (tedy X, Y, Z) a její konkrétní realizace malými písmeny (tedy x, y, z). Příklady náhodných veličin mohou být:

- X = výška náhodně vybraného člověka (pak x jsou kladná reálná čísla)
- X = výsledek hodu kostkou (x nabývá hodnot od 1 do 6)
- X = zda náhodně vybraný člověk kouří (x nabývá hodnot *ano* nebo *ne*).

Provedeme-li pro náhodnou veličinu n měření, dostaneme hodnoty x_1, \dots, x_n , které nazýváme **náhodný výběr**. Chování náhodné veličiny můžeme popsat jedním ze dvou následujících způsobů

- buď vyjmenujeme všechny hodnoty, kterých daná veličina může nabývat (tj. všechna možná x) a udáme pravděpodobnosti, s jakými se tyto hodnoty realizují, tj. udáme $P(X = x)$ pro všechna x (u spojitých rozdělení udáme hustotu pravděpodobnosti $f(x)$)
- nebo udáme hodnoty tzv. **distribuční funkce** $F(x) := P(X \leq x)$ (tj. pro jednotlivá x určíme, s jakou pravděpodobností bude hodnota náhodné veličiny nejvýše rovna danému x). Distribuční funkce se často označuje též $F_X(x)$, aby se zdůraznilo, že jde o distribuční funkci náhodné veličiny X v bodě x .

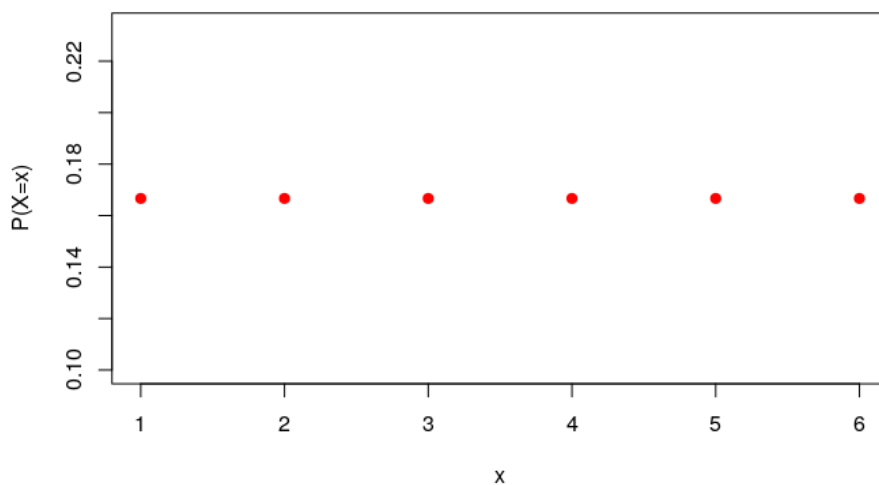
Obě tyto možnosti nějakým způsobem popisují **rozdělení pravděpodobnosti** náhodné veličiny X , tj. udávají, jak je celková pravděpodobnost (která je rovna jedné) rozdělena přes jednotlivé hodnoty x .

Příklad - hod kostkou

Pro příklad si nyní zvolme náhodnou veličinu X , která bude vyjadřovat výsledek hodu kostkou. V tomto případě bude $x \in \{1, 2, 3, 4, 5, 6\}$ (tj. výsledek našeho pokusu bude prvek z množiny 1 až 6). Pravděpodobnost je pak přes jednotlivá x rozdělena následovně:

$$\begin{aligned} P(X = 1) &= \frac{1}{6} & P(X = 2) &= \frac{1}{6} \\ P(X = 3) &= \frac{1}{6} & P(X = 4) &= \frac{1}{6} \\ P(X = 5) &= \frac{1}{6} & P(X = 6) &= \frac{1}{6}, \end{aligned}$$

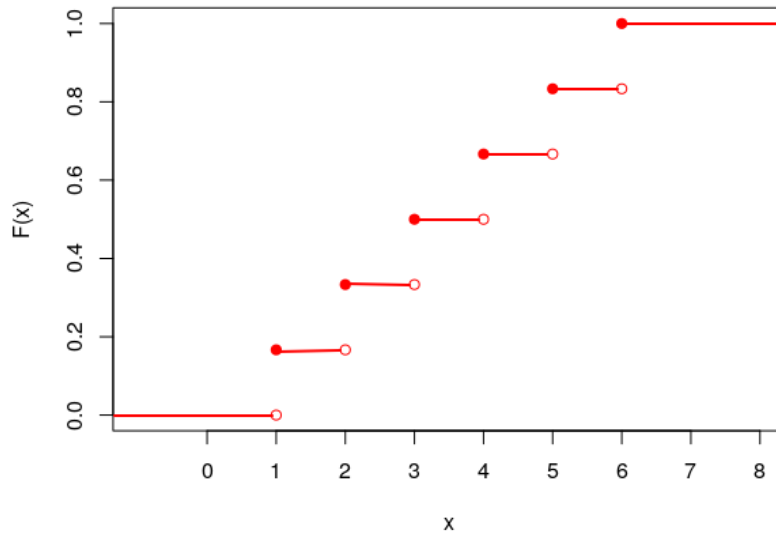
stručně řečeno $P(X = x) = \frac{1}{6}$ pro všechna $x \in \{1, 2, 3, 4, 5, 6\}$. Rozdělení si lze graficky znázornit následovně



Distribuční funkci lze definovat na celé reálné ose, akorát v mnoha úsecích bude konstantní. Pojďme se nyní na jednotlivé úseky podívat:

- je zřejmé, že pro všechna $x < 1$ bude $F(x) = P(X \leq x) = 0$, neboť na kostce nelze hodit číslo ostře menší než 1.
- v jedničce nastane skok, neboť $F(x) = P(X \leq x) = \frac{1}{6}$ pro všechna x taková, že $1 \leq x < 2$ (neboť samozřejmě $P(X \leq 1,5) = \frac{1}{6}$).
- ve dvojce nastane další skok, tudíž: $F(x) = P(X \leq x) = \frac{2}{6}$ pro všechna x taková, že $2 \leq x < 3$
- analogicky pro $3 \leq x < 4$ dostáváme $F(x) = P(X \leq x) = \frac{3}{6}$
- a tak dále pro ostatní body
- až nakonec pro $x \geq 6$ dostáváme $F(x) = P(X \leq x) = 1$

Graficky lze tuto distribuční funkci znázornit následovně:



Poznámka 1 Z grafu distribuční funkce lze učinit jedno důležité pozorování - distribuční funkce je neklesající! (Tj. roste, nebo je konstantní). To platí obecně pro distribuční funkci libovolné veličiny. Dále je zajímavé si povšimnout, že směrem do mínus nekonečna má $F(x)$ hodnotu 0 (nebo se k ní aspoň blíží), směrem do plus nekonečna má hodnotu 1 (nebo se k ní aspoň blíží). I to je obecná vlastnost distribučních funkcí.

Střední hodnota

Z rozdělení náhodné veličiny lze určit některé důležité charakteristiky. První z nich je **střední hodnota**, která je mírou polohy daného rozdělení. Zjednodušeně řečeno je to číslo, kolem kterého se hodnoty X nejčastěji pohybují, jakási neočekávanější hodnota X .

Představuje-li X nějaký znak (např. hmotnost), který zkoumáme v populaci jedinců, pak se střední hodnota tohoto znaku v populaci často nazývá **populační průměr**.

Střední hodnotu teoreticky vypočteme dle vzorce

$$E X = \sum_x x P(X = x), \quad (1)$$

tedy jako součet jednotlivých možných hodnot x převážených příslušnou pravděpodobností.

Poznámka 2 Všimněte si korespondence vzorečku (1) se vzorečkem pro výpočet aritmetického průměru. Máme-li naměřené hodnoty (náhodný výběr) $v = 1, 3, 2, 5, 5, 3, 5$, pak příslušný aritmetický průměr (většinou nazývaný **výběrový průměr**) je

$$\bar{v} = \frac{1}{7}(1 + 3 + 2 + 5 + 5 + 3 + 5) = 1 \cdot \frac{1}{7} + 2 \cdot \frac{1}{7} + 3 \cdot \frac{2}{7} + 5 \cdot \frac{3}{7}.$$

Relativní četnosti $\frac{1}{7}, \frac{1}{7}, \frac{2}{7}, \frac{3}{7}$ jsou odhadem pravděpodobností $P(V = 1), P(V = 2), P(V = 3), P(V = 5)$ a výběrový průměr \bar{v} je odhadem střední hodnoty $E V$.

Střední hodnota naší veličiny představující výsledek hodu kostkou je

$$E X = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3,5.$$

Poznámka 3 Střední hodnota může být i číslo, kterého daná veličina reálně nenabývá. To vidíme i zde, neboť na kostce nelze hodit 3,5.

Poznámka 4 Označení EX pochází z anglického termínu pro střední hodnotu - "Expected value of X ". Výběrový průměr se pak nazývá "sample mean".

Rozptyl

Další charakteristikou každého rozdělení je rozptyl. Ten vyjadřuje míru rozptýlenosti hodnot dané veličiny kolem její střední hodnoty. Je zadán vzorcem

$$\text{var } X = E(X - EX)^2, \quad (2)$$

přičemž často se hodí též vyjádření $\text{var } X = E(X^2) - (EX)^2$, které lze dostat ze vzorečku (2) jednoduchou algebraickou úpravou. Praktický výpočet můžeme provést podle vzorečku

$$\text{var } X = \sum_x (x - EX)^2 \cdot P(X = x).$$

Poznámka 5 Opět si povšimněte korespondence vzorečku (2) se vzorečkem pro tzv. výběrový rozptyl, který již znáte

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Symbol E v (2) byl vlastně jen nahrazen aritmetickým průměrem.

Rozptyl veličiny představující výsledek hodu kostkou je:

$$\begin{aligned} \text{var } X &= (1 - EX)^2 \cdot P(X = 1) + (2 - EX)^2 \cdot P(X = 2) + (3 - EX)^2 \cdot P(X = 3) + \\ &+ (4 - EX)^2 \cdot P(X = 4) + (5 - EX)^2 \cdot P(X = 5) + (6 - EX)^2 \cdot P(X = 6) \\ &\doteq 2,91 \end{aligned}$$

kde jsme využili předchozí výsledek $EX = 3,5$.

Poznámka 6 Označení $\text{var } X$ pochází z anglického názvu *Variance of X* .

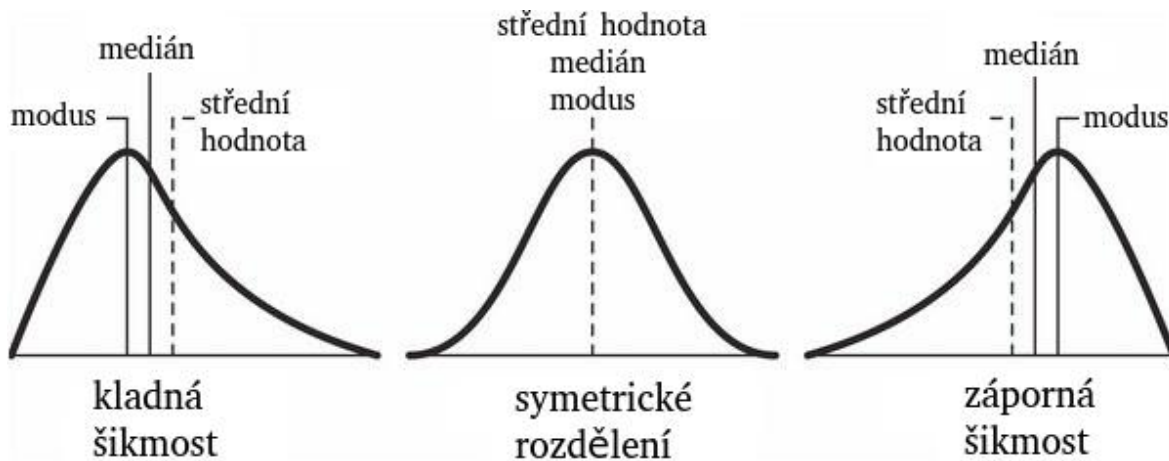
Šikmost a špičatost

Dalšími pojmy, které charakterizují každé rozdělení je šikmost a špičatost. Jsou dány vzorci

$$\begin{aligned} \gamma_3 &= \frac{E(X - EX)^3}{(\text{var } X)^{3/2}} && (\text{šikmost}) \\ \gamma_4 &= \frac{E(X - EX)^4}{(\text{var } X)^2} && (\text{špičatost}) \end{aligned}$$

Šikmost určuje, kterým směrem je rozdělení naší veličiny asymetricky rozloženo. Rozlišujeme šikmost kladnou, kdy se veličina s větší pravděpodobností nachází pod svou střední hodnotou a šikmost zápornou, kdy veličina spíše nabývá hodnot nad svou střední hodnotou. Nejlépe je to asi vidět na Obrázku 1. Symetrická rozdělení mají přirozeně nulovou šikmost.

Špičatost udává, jak rychle se chvosty rozdělení (tím se myslí hodnoty pravděpodobnosti pro $x \rightarrow \pm\infty$) přibližují k nule (tj. k ose x). Jinými slovy určuje, jak se v rozdělení dané veličiny vyskytují velmi vysoké ($x \rightarrow +\infty$) a velmi nízké ($x \rightarrow -\infty$) hodnoty.



Obrázek 1: Poloha střední hodnoty, mediánu (prostřední hodnota v uspořádaných datech) a modusu (nejčetnější hodnota) u rozdělení s různou šikmostí.

Poznámka 7 Teoretická špičatost normálního rozdělení (o němž bude řeč později) je rovna hodnotě 3. Proto se často používá alternativní definice špičatosti

$$\gamma'_4 = \frac{E(X - EX)^4}{(\text{var } X)^2} - 3,$$

která posune špičatost normálního rozdělení do nuly. Chceme-li pak porovnat špičatost nějaké náhodné veličiny se špičatostí normálního rozdělení, stačí porovnat příslušnou hodnotu γ'_4 s nulou.

Poznámka 8 Střední hodnoty jednotlivých mocnin X se nazývají **momenty rozdělení**. Tedy EX je první moment, $E(X^2)$ druhý moment, $E(X^3)$ třetí moment atd. Obecně $E(X^k)$ je k -tý moment rozdělení. Analogicky definujeme k -tý **centrální moment rozdělení** jako $E(X - EX)^k$. Tedy rozstyl $E(X - EX)^2$ je vlastně druhý centrální moment.

Poznámka 9 Ještě poznámka k chvostům rozdělení. Rozdělení má těžké chvosty, pokud velmi vysoké nebo velmi nízké hodnoty mají relativně vysokou pravděpodobnost. Pokud naopak velmi vysoké/nízké hodnoty mají malou pravděpodobnost, říkáme, že rozdělení má lehké chvosty. Tento pojem se používá převážně ve spojitosti s hustotami spojitých rozdělení.

2 Základní diskrétní rozdělení

Náhodná veličina X je označovaná jako diskrétní (neboli její rozdělení je označováno za diskrétní), pokud nabývá pouze konečně (případně spočetně = přirozená čísla) mnoha hodnot. Taková náhodná veličina většinou vyjadřuje počet něčeho (např. počet dětí náhodně vybrané matky), nebo má kvalitativní charakter (např. barva očí náhodně vybraného jedince). V prvním případě jsou její realizace x přirozená čísla, ve druhém příkladu je $x \in \{\text{hnědá, modrá, zelená}\}$.

Rozdělení diskrétní náhodné veličiny popisujeme vyjmenováním $P(X = x)$ pro všechna x , nebo distribuční funkcí. Ta je v tomto případě schodovitou funkcí se skoky v bodech x a lze ji spočítat dle vzorce

$$P(X \leq x) = \sum_{j \leq x} P(X = j), \quad x \in \mathbb{R}.$$

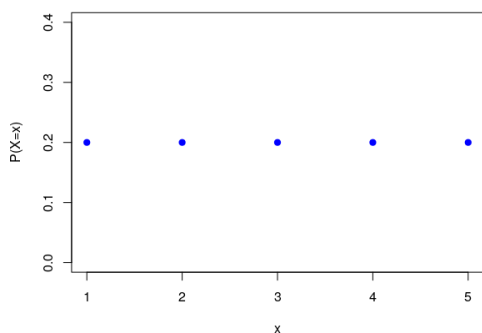
Rovnoměrné rozdělení

Náhodná veličina představující výsledek hodu (spravedlivou) kostkou je příkladem diskrétního rozdělení, které se nazývá rovnoměrné, neboť všechny možné hodnoty x mají stejnou pravděpodobnost, tedy pravděpodobnost je rovnoměrně rozdělená přes všechna možná x .

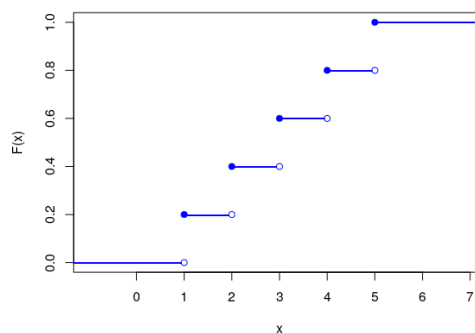
Fakt, že náhodná veličina X má rovnoměrné rozdělení na množině $1, \dots, k$ označujeme symbolem $X \sim R[1, k]$ (značení se může v různých učebnicích lišit). Pro toto rozdělení platí:

- $P(X = x) = \frac{1}{k}$, pro všechna $x \in \{1, 2, \dots, k\}$ (viz Obrázek 2a)
- distribuční funkce $F(x) = P(X \leq x) = \sum_{j=1}^x P(X = j)$ je schodovitá se skoky o velikosti vždy $1/k$ v bodech $1, 2, \dots, k$ (viz Obrázek 2b)
- $E X = \frac{1}{k}(1 + 2 + \dots + k) = \frac{1+k}{2}$ (jde tedy o aritmetický průměr minimální a maximální hodnoty) - zkuste si ověřit výpočtem dle vzorce (1)
- $\text{var } X = \frac{k^2-1}{12}$

Anglicky se toto rozdělení nazývá "discrete uniform distribution".



(a) Pravděpodobnosti jednotlivých hodnot



(b) Distribuční funkce

Obrázek 2: Rovnoměrné rozdělení $R[1, 5]$

Alternativní rozdělení

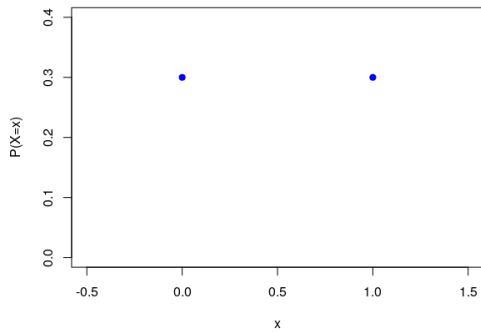
O náhodné veličině nabývající pouze dvou hodnot (typicky 0,1, ale může to být třeba ne/ano, rub/líc apod., které si do 0/1 zakódujeme) řekneme, že má alternativní rozdělení. Tento fakt označujeme symbolem $X \sim alt(p)$, kde p označuje $P(X = 1)$. Pro toto rozdělení platí:

- $P(X = 0) = 1 - p$, $P(X = 1) = p$
- $F(x) = P(X \leq x)$ (viz Obrázek 3b)
- $E X = 0 \cdot P(X = 0) + 1 \cdot P(X = 1) = 0 \cdot (1 - p) + 1 \cdot p = p$
- $\text{var } X = p(1 - p)$ (zkuste si spočítat sami)

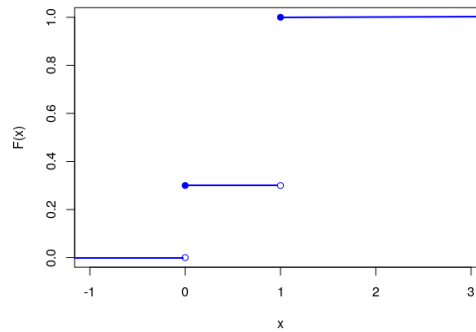
Příklady náhodných veličin s alternativním rozdělením:

- $X =$ výsledek hodu mincí, pak $X \sim alt(\frac{1}{2})$, kde $x \in \{\text{rub, líc}\}$
- $X =$ zda je náhodně vybraný člověk žena $\rightarrow X \sim alt(0, 51)$, $x \in \{\text{ano, ne}\}$ ¹

¹dle <https://data.worldbank.org/indicator/sp.pop.tot1.fe.zs> je podíl žen v ČR roven 50.8 %



(a) Pravděpodobnosti jednotlivých hodnot



(b) Distribuční funkce

Obrázek 3: Alternativní rozdělení $alt(0.3)$

Binomické rozdělení

Náhodná veličina s binomickým rozdělením vyjadřuje počet úspěchů v sérii n pokusů, kde v každém z dílčích pokusů má úspěch pravděpodobnost p . Takovým úspěchem v pokusu může být např. líc při hodu mincí, šestka při hodu kostkou, ale i třeba to, že daný jedinec má nějaké zkoumané onemocnění.

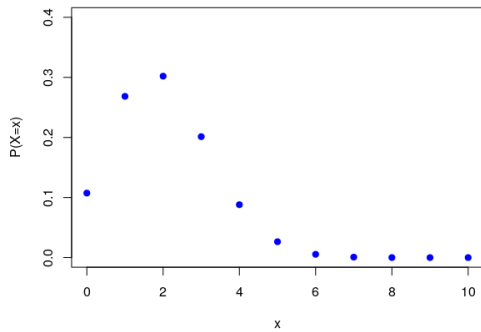
Binomické rozdělení má dva parametry - n a p a fakt, že veličina X má binomické rozdělení zapisujeme jako $X \sim Bi(n, p)$. Pro toto rozdělení platí:

- $x \in \{0, 1, 2, \dots, n\}$ (v n pokusech může nastat 0 až n úspěchů)
- $P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$ (viz Obrázek 4a)
- $F(x) = P(X \leq x)$ je nejlépe vidět na Obrázku 4b
- $E X = np$
- $\text{var } X = np(1 - p)$

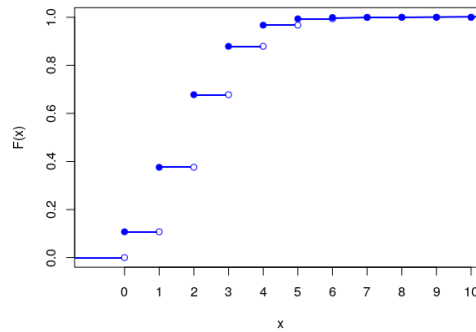
Příklady veličin s binomickým rozdělením:

- $X =$ počet líců při hodu 20 mincemi $\rightarrow X \sim Bi(20, \frac{1}{2})$
- $X =$ počet žen mezi třemi náhodně vybranými lidmi $\rightarrow X \sim Bi(3, 0.51)$

Poznámka 10 V případě jednoho pokusu (tj. $n = 1$) splývá binomické rozdělení s alternativním.



(a) Pravděpodobnosti jednotlivých hodnot



(b) Distribuční funkce

Obrázek 4: Binomické rozdělení $Bi(10, 0.2)$

Poissonovo rozdělení

Veličina s Poissonovým² rozdělením vyjadřuje počet jevů, které nastaly v daném časovém intervalu. Typicky se jedná o jevy s malou pravděpodobností, proto se také Poissonovo rozdělení někdy označuje jako "zákon řídkých jevů". Formální zápis opět zní $X \sim Po(\lambda)$, kde λ je parametr tohoto rozdělení a označuje intenzitu, s jakou se sledovaný jev vyskytuje. λ je vlastně průměrný počet událostí za jednotku času a může nabývat libovolných kladných hodnot (tj. $\lambda \in (0, \infty)$). Pro toto rozdělení platí:

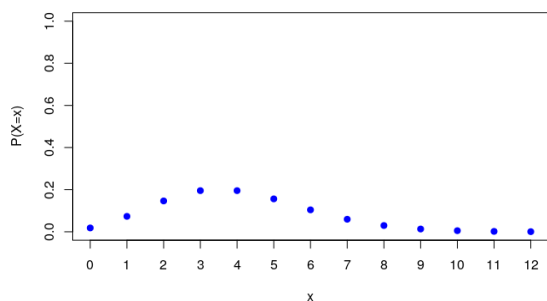
- $x \in \{0, 1, 2, 3, \dots\}$ (tedy libovolné přirozené číslo včetně nuly)
- $P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$ (viz Obrázek 5a)
- $F(x) = P(X \leq x)$ je nejlépe vidět na Obrázku 5b
- $E X = \lambda$
- $\text{var } X = \lambda$

Příklady veličin s Poissonovým rozdělením:

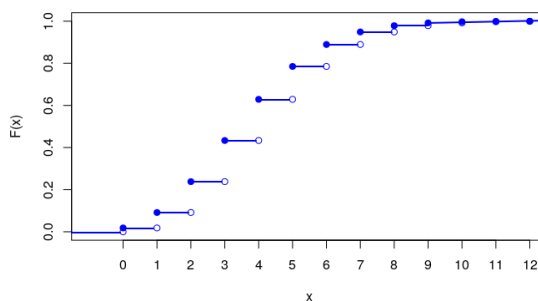
- počet lidí, kteří během jedné minuty vešli na poštu
- počet částic jaderného záření, které během 1 sekundy dopadly na olověnou desku

Poznámka 11 Binomické rozdělení pro velké n a malé p se také týká řídkých jevů a lze ho tedy dobře aproximovat Poissonovým rozdělením (viz Obrázek 6).

²Siméon Denis Poisson (francouzský matematik)

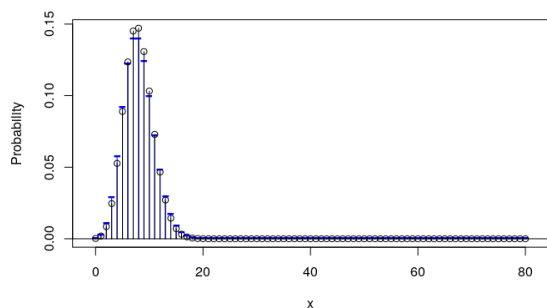


(a) Pravděpodobnosti jednotlivých hodnot

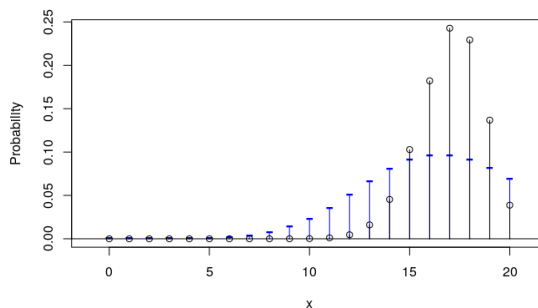


(b) Distribuční funkce

Obrázek 5: Poissonovo rozdělení $Po(4)$



(a) Aproximace $Bi(80, 0.1)$



(b) Aproximace $Bi(20, 0.85)$

Obrázek 6: Aproximace binomického rozdělení $Bi(n, p)$ Poissonovým rozdělením je dobrá při velkém n a malém p .

3 Spojitá rozdělení

Náhodná veličina je spojitá (resp. má spojitě rozdělení), pokud počet hodnot, kterých může nabývat je nekonečný. Typicky je její obor hodnot roven reálným číslům, nebo jejich podmnožině. Příkladem spojitě náhodné veličiny může být váha nebo délka jedince, věk, a podobně.

Popis rozdělení spojitých náhodných veličin je podobný jako u diskrétních veličin. Jelikož však možných hodnot x je nekonečně mnoho, nemá smysl se snažit vyjmenovat všechny $P(X = x)$. Naopak u spojitých rozdělení je $P(X = x) = 0$ a místo toho se udává tzv. hustota rozdělení $f(x)$ s následujícím významem: pravděpodobnost, že náhodná veličina nabyde hodnoty x z intervalu (a, b) je rovna ploše pod křivkou $f(x)$ v intervalu (a, b) . Matematicky zapsáno:

$$P(X \in (a, b)) = \int_a^b f(x) dx.$$

Z tohoto výrazu je také patrné, že $P(X = a) = \int_a^a f(x) dx = 0$. Hustota $f(x)$ je spojitá funkce definovaná na reálných číslech a celková plocha pod ní je rovna 1.

Distribuční funkce $F(x)$ spojitě náhodné veličiny je opět neklesající spojitá funkce, která se v $-\infty$ přimyká k 0 a v $+\infty$ k 1 (to je asi nejlépe patrné na obrázcích níže). Je definovaná

vzorcem

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(y)dy.$$

Navíc platí, že derivací distribuční funkce je hustota, tedy $F'(x) = f(x) \forall x \in \mathbb{R}$.

Asi jste zpozorovali, že tam, kde byla u diskrétních rozdělení suma přes x , tj. \sum_x , je nyní integrál. To platí i u definice střední hodnoty, která má u spojitých rozdělení tvar

$$E X = \int_{-\infty}^{\infty} x f(x) dx,$$

rozptyl je opět

$$\text{var } X = E(X - E X)^2 = \int_{-\infty}^{\infty} (x - E X)^2 f(x) dx.$$

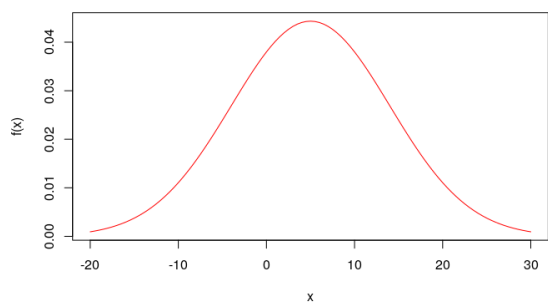
Normální rozdělení

Asi nejvýznamnějším zástupcem spojitých rozdělení je normální (neboli Gaussovo³) rozdělení. Fakt, že veličina X je normálně rozdělená označujeme symbolem $X \sim N(\mu, \sigma^2)$, kde parametr $\mu \in \mathbb{R}$ označuje střední hodnotu a parametr $\sigma^2 \geq 0$ rozptyl. Pro toto rozdělení platí:

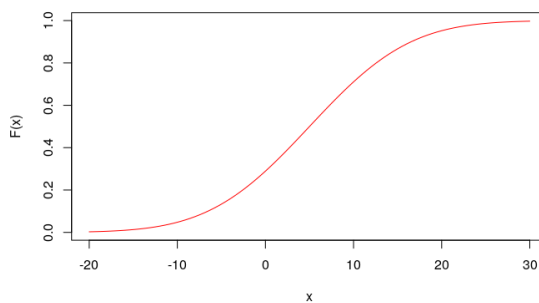
- $x \in \mathbb{R}$ (tedy X může nabývat libovolných reálných hodnot)
- hustota $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ má tvar dobře známé zvonovité křivky (viz Obrázek 7a) a často se jí říká "Gaussova křivka", popř. "gaussovka"
- distribuční funkci $F(x)$ si lze prohlédnout na Obrázku 7b
- $E X = \mu$
- $\text{var } X = \sigma^2$
- šikmost $\gamma_3 = 0$ (hustota $f(x)$ je symetrická kolem μ)
- špičatost $\gamma_4 = 3$, $\gamma'_4 = 0$

Je-li $\mu = 0$ a $\sigma^2 = 1$, pak jde o tzv. **normované normální rozdělení**.

³Carl Friedrich Gauss (německý matematik)



(a) Hustota



(b) Distribuční funkce

Obrázek 7: Normální rozdělení $N(5, 9)$

Normální rozdělení je základním předpokladem mnoha statistických metod a má ve statistice zvláštní postavení také díky tzv. centrální limitní větě. Můžeme na něj ale narazit i v běžném životě. Například se má za to, že výška lidí má přibližně normální rozdělení. Někdy ho lze zase zpozorovat na sešlapaných schodech :-)

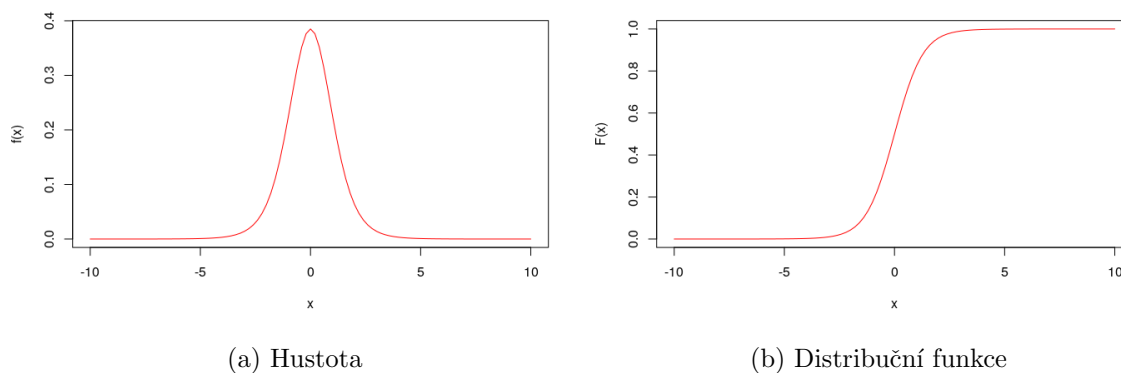


t-rozdělení

Někdy též nazýváno Studentovo⁴ t-rozdělení. Symbolicky ho zapisujeme jako $X \sim t_n$, kde výraz n je parametrem tohoto rozdělení a nazývá se "stupeň volnosti". Tento parametr je vždy větší nebo roven jedné, tj. $n \geq 1$. Pro toto rozdělení platí:

- $x \in \mathbb{R}$ (veličina s t-rozdělením tedy může nabývat libovolné reálné hodnoty)
- hustota $f(x)$ má podobný tvar jako hustota normálního rozdělení (viz Obrázek 8a), její analytický tvar je ale velmi složitý, proto ho zde nebudeme uvádět. Pro libovolné n je hustota symetrická kolem 0.
- distribuční funkce $F(x)$ - viz Obrázek 8b
- $EX = 0$
- $\text{var } X = \frac{n}{n-2}$ pro $n > 2$ (pro $1 \leq n \leq 2$ není rozptyl definován)

Také t-rozdělení hraje významnou roli ve statistických testech - zejména tam, kde není znám populační rozptyl zkoumaného znaku.



Obrázek 8: Studentovo t-rozdělení se 7 stupni volnosti, tj. t_7 .

Další spojitá rozdělení

Mezi další spojitá rozdělení, se kterými se lze ve statistice běžně setkat, patří χ^2 rozdělení a F-rozdělení (zvané též Fisherovo-Snedecorovo). Popis těchto rozdělení lze nalézt v učebnicích a na internetu.

⁴tento pseudonym si zvolil autor t-rozdělení - anglický matematik William Sealy Gosset