Poznámky k předmětu Aplikovaná statistika, 5.téma

5. Některá významná rozdělení

A. Diskrétní rozdělení

(*i*) Diskrétní rovnoměrné rozdělení na množině $\{1, \ldots, n\}$

Náhodná veličina X, která má diskrétní rovnoměrné rozdělení (značíme $X \sim R(\{1, \ldots, n\})$) nabývá pouze hodnot z množiny $\{1, \ldots, n\}$ a to s pravděpodobností $\frac{1}{n}$, tj.

$$\mathbb{P}(X = i) = \frac{1}{n}, \qquad i \in \{1, \ldots, n\}.$$

Střední hodnota: $\mathbb{E}X = \frac{n+1}{2}$.
Rozptyl: $\mathrm{var}X = \frac{n^2-1}{12}$.

(*ii*) Alternativní rozdělení $X \sim Alt(p)$, $p \in (0, 1)$

Náhodná veličina X mající alternativní rozdělení nabývá pouze hodnot 0 a 1 a to s následujícími pravděpodobnostmi:

$$\begin{aligned} \mathbb{P}(X = 1) &= p, \\ \mathbb{P}(X = 0) &= 1 - p. \end{aligned}$$

Střední hodnota: $\mathbb{E}X = p$.
Rozptyl: $\mathrm{var}X = p(1 - p)$.

Příklad 1. *Házíme na basketbalový koš ze šestky. V* 80% *případů se trefíme. Náhodná veličina* X*, která nabývá hodnoty* 1*, trefíme-li koš a hodnoty* 0*, jestliže koš netrefíme, má rozdělení* $Alt(0, 8)$.

(*iii*) Binomické rozdělení $X \sim Bi(n, p)$, $n \in \mathbb{N}$, $p \in (0, 1)$

Náhodná veličina X nabývá hodnot $k = 0, \ldots, n$ s pravděpodobnostmi

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \qquad k = 0, 1, \ldots, n.$$

Náhodnou veličinou $X \sim Bi(n, p)$ se modeluje počet úspěchů v n **nezávislých** pokusech, je-li pravděpodobnost úspěchu v jednotlivých pokusech rovna p. Zevrubně řečeno, označme náhodné veličiny

$$Y_i \ldots \text{výsledek } i\text{-tého pokusu}, \; i = 1, \ldots, n.$$

Pak

$$\begin{aligned} \mathbb{P}(Y_i = 1) &= p, \\ \mathbb{P}(Y_i = 0) &= 1 - p. \end{aligned}$$

a veličiny Y_i jsou nezávislé a mají alternativní rozdělení s parametrem p. Veličina X se potom dá zapsat jako

$$X = \sum_{i=1}^{n} Y_i.$$

Tedy veličina $X \sim Bi(n,p)$ se dá interpretovat jako součet n nezávislých veličin s rozdělením $Alt(p)$.

Bez této interpretace by pro nás počítání **střední hodnoty** a rozptylu bylo pro obecné n, p dosti obtížné (vyžaduje znalosti sčítání řad), ovšem takto je výpočet velmi přímočarý:

$$\mathbb{E}X = \mathbb{E}\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \mathbb{E}Y_i = \sum_{i=1}^{n} p = np.$$

Rozptyl se díky nezávislosti veličin Y_i vypočítá analogicky,

$$\mathrm{var}X = np(1-p).$$

Příklad 2. *Házíme desetkrát na koš. Pravděpodobnost, že se v jednotlivých hodech trefíme je opět* $0,8$. *Potom náhodná veličina* Y *popisující počet trefených košů při těchto deseti hodech má binomické rozdělení* $Bi(10; 0,8)$.

(*iv*) Geometrické rozdělení $X \sim Ge(p)$, $p \in (0,1)$

Náhodná veličina X nabývá hodnot $k = 0, 1, \ldots$ s pravděpodobnostmi

$$\mathbb{P}(X = k) = p(1-p)^k, \qquad k = 0, 1, \ldots$$

Geometrickým rozdělením se modeluje počet neúspěšných pokusů před prvním úspěchem, jsou-li jednotlivé pokusy nezávislé a pravděpodonost úspěchu je p.

Střední hodnota: $\mathbb{E}X = \frac{1-p}{p}$.
Rozptyl: $\mathrm{var}X = \frac{1-p}{p^2}$.

Příklad 3. *Házíme opět na koš. Počítáme si, jak dlouho, respektive kolik hodů, nám trvá, než se poprvé trefíme. Náhodná veličina* Z *udávající počet neúspěšných hodů před prvním úspěšným má geometrické rozdělení* $Ge(0,8)$.

(*v*) Poissonovo rozdělení $X \sim Po(\lambda)$, $\lambda > 0$

Rozdělení veličiny X je dáno pravděpodobnostní funkcí

$$\mathbb{P}(X = k) = e^{-\lambda}\,\frac{\lambda^k}{k!}, \text{ pro } k = 0, 1, \ldots$$

Poissonovým rozdělením modelujeme náhodnou veličinu, která vyjadřuje počet výskytů události v určitém intervalu (času, délky, objemu,...), jestliže události nastávají náhodně a nezávisle na sobě.

Střední hodnota: $\mathbb{E}X = \lambda$.
Rozptyl: $\mathrm{var}X = \lambda$.

Spojitost s binomickým rozdělením:

Uvažujme náhodnou veličinu $X \sim Bi(n,p)$, kde $n \to \infty$, $p \to 0$, a $np = \lambda$. Potom

$$P(X = k) = \frac{n(n-1)\ldots(n-k+1)}{k!} p^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \underset{n\to\infty, p\to 0}{\longrightarrow} \frac{\lambda^k}{k!} e^{-\lambda},$$

tedy veličina X má asymptoticky Poissonovo rozdělení.

Cvičení k tomuto tématu

(i) Ve velké várce vajec je 1/3 zkažených. Jaká je pravděpodobnost, že ze šesti náhodně vybraných vajec budou nejvýše dvě zkažená? Nadto nakreslete graf pravděpodobnostní funkce příslušné náhodné veličiny.

(ii) Na telefonní ústřednu přichází průměrně 100 hovorů za jednu hodinu, přičemž příchody hovorů jsou náhodné. Jaká je pravděpodobnost, že

(a) když telefonistka odejde na deset minut do kuchyňky, nezmešká žádný hovor?

(b) ve třech minutách přijdou více než dva hovory?

(iiia) Ve výrobní hale je 8 stejných žárovek. Pravděpodobnost, že během směny praskne jedna žárovka, je $0,4$. Žárovky praskají nezávisle a nelze je během směny vyměnit. Když jich praskne více než 5, musí pracovníci za tuto směnu dostat příplatek. Jaká je pravděpodobnost, že za příští směnu dostanou příplatek?

(iiib) Ve výrobní hale je velké množství žárovek. Ví se, že za směnu praskne průměrně $3,2$ žárovky. Jaká je pravděpodobnost, že během příští směny praskne víc než 5 žárovek (a pracovníci dostanou příplatek)? Nakreslete graf pravděpodobnostní funkce příslušné náhodné veličiny.

B. Spojitá rozdělení

(*i*) Rovnoměrné rozdělení na intervalu $[a,b]$ $X \sim R([a,b])$, $a < b \in \mathbb{R}$

Má hustotu pravděpodobnosti

$$f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b, \\ 0 & x < a, \quad x > b. \end{cases}$$

Tato hustota je konstantní na $[a,b]$, což skutečně odpovídá tomu, že pravděpodobnost, že X padne do podintervalu intervalu $[a,b]$, je úměrná velikosti tohoto podintervalu vůči $[a,b]$. (Srovnejte s geometrickou pravděpodobností.)
Distribuční funkce je pak

$$F(x) = \begin{cases} 0 & x < a, \\ \frac{x-a}{b-a} & a \le x \le b, \\ 1 & x \ge b. \end{cases}$$

Prostým integrováním obdržíme vzorce pro střední hodnotu a rozptyl. Povšimněme si, že střední hodnota je přesně uprostřed intervalu $[a,b]$.

Střední hodnota: $\mathbb{E}X = \frac{a+b}{2}$.
Rozptyl: $\mathrm{var}X = \frac{1}{12}(b-a)^2$.

(*ii*) Exponenciální rozdělení $X \sim Exp(\lambda)$, $\lambda > 0$

Veličina X nabývá hodnot z intervalu $(0,\infty)$.
Hustota závisí na parametru λ a je dána vztahem

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0 & \text{jinak.} \end{cases}$$

Distribuční funkce F:

$$F(x) = \begin{cases} 0 & \text{pro } x \leq 0, \\ 1 - e^{-\lambda x} & x > 0. \end{cases}$$

Střední hodnota: $\mathbb{E}X = \frac{1}{\lambda}$.
Rozptyl: $\mathrm{var}X = \frac{1}{\lambda^2}$.

Exponenciální rozdělení je rozdělení „bez paměti“:
Pro náhodnou veličinu X s exponenciálním rozdělením platí:

$$P(X > x + y | X > y) = P(X > x), \quad \forall x > 0, y > 0,$$

neboť užitím definice podmíněné pravděpodobnosti můžeme pravděpodobnost $P(X > x + y|X > y)$ přepsat ve tvaru

$$\frac{P(X > x + y)}{P(X > y)} = \frac{e^{-\lambda(x+y)}}{e^{-\lambda y}} = e^{-\lambda x}.$$

Exponenciálním rozdělením se modeluje doba čekání na událost a má velký význam v teorii přežití a v kombinaci s Poissonovým rozdělením i v systémech hromadné obsluhy (viz níže).

Souvislost s Poissonovým rozdělením:
Jestliže náhodá veličina X, která popisuje dobu čekání na výskyt události, má rozdělení $Exp(\lambda)$, potom n.v. Y popisující počet těchto událostí, jež nastaly v časovém intervalu délky T, má Poissonovo rozdělení $Po(\lambda T)$.

(*iii*) Normální rozdělení $X \sim \mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$

Říká se mu také **Gaussovo rozdělení**. Veličina X nabývá hodnot z celého $\mathbb{R}$.
Hustota normálního rozdělení je

$$f(x) = \frac{1}{\sqrt{2\pi\ \sigma^2}}\ e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty,$$

a říká se jí **Gaussovka křivka**.
Distribuční funkce je

$$F(x) = \frac{1}{\sqrt{2\pi\ \sigma^2}}\ \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \mathrm{d}t, \quad -\infty < x < \infty,$$

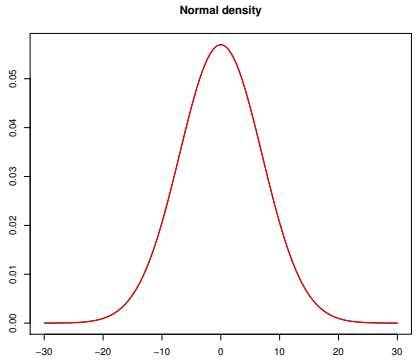
a **nelze** ji explicitně vyjádřit vzorcem.

Normální rozdělení má v teorii pravděpodobnosti a matematické statistice zásadní roli, jak se dozvíme v další kapitole. Velmi hrubě řečeno se dá normálním rozdělením modelovat součet velmi mnoha drobných náhodných faktorů.

Střední hodnota: $\mathbb{E}X = \mu$.
Rozptyl: $\mathrm{var}X = \sigma^2$.

(*iiib*) Normované normální rozdělení $X \sim \mathcal{N}(0, 1)$

Je znormovaná náhodná veličina s normálním rozdělením, tj. $\mu = \mathbb{E}X = 0$ a $\sigma^2 = \mathrm{var}X = 1$.



Obrázek 1: Hustota $\mathcal{N}(0, 49)$.

Hustota normovaného normálního rozdělení je

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}, \quad -\infty < x < \infty.$$

Distribuční funkce Φ je

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, \mathrm{d}t, \ -\infty < x < \infty,$$

a její hodnoty jsou tabelizovány (lze je nalézti ve statistických tabulkách).
Díky symetrii hustoty ϕ platí:

$$\Phi(x) = 1 - \Phi(-x), \qquad x \in \mathbb{R},$$

a tudíž je dostatečné tabelizovat hodnoty $\Phi(x)$ pro nezáporné hodnoty x.
Transformace normálně rozdělených veličin

1. Jestliže Y má normální rozdělení s parametry μ a σ^2, potom

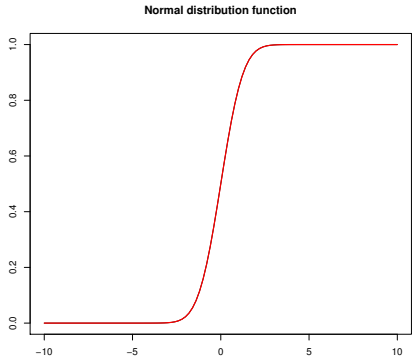
   $$X = \frac{Y - \mu}{\sigma}$$

   má normované normální rozdělení $\mathcal{N}(0, 1)$.
2. Jestliže X má normální rozdělení s parametry μ a σ^2 a $Y = a + bX$, potom Y má opět normální rozdělení s parametry $a + b\mu$ a $b^2\sigma^2$.
3. Uvažujme náhodné veličiny X a Y, $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, jsou **nezávislé**. Potom n.v. $Z = aX + bY + c$ má rozdělení $\mathcal{N}(a\mu_1 + b\mu_2 + c, a^2\sigma_1^2 + b^2\sigma_2^2)$.

Díky vztahu 1. není třeba tabelizovat hodnoty distribuční funkce F pro všechna rozdělení $\mathcal{N}(\mu, \sigma^2)$ (to by ani nebylo možné!), ale stačí nám tabulka pro Φ, neboť pro všechny distribuční funkce F platí vztah

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

U následujících rozdělení si uvedeme pouze jejich definice bez dalších vlastností. Tato rozdělení budou hrát důležitou roli při pozdějším testování hypotéz.



Obrázek 2: Distribuční funkce $\mathcal{N}(0,1)$.

(*iv*) χ^2- rozdělení o n stupních volnosti $X \sim \chi^2(n)$

Nechť $X_1, \ldots, X_n$ jsou nezávislé náhodné veličiny s normovaným normálním rozdělením $\mathcal{N}(0,1)$. Potom náhodná veličina

$$Y_n = \sum_{i=1}^{n} X_i^2$$

má tzv. χ^2- rozdělení o n stupních volnosti. Symbolem $\chi^2_\alpha(n)$ budeme značit $\alpha-$kvantily tohoto rozdělení.

(*v*) Studentovo rozdělení o n stupních volnosti $X \sim t(n)$

Nechť U a V jsou nezávislé náhodné veličiny, $U \sim \mathcal{N}(0,1)$ a $V \sim \chi^2(n)$. Potom náhodná veličina

$$W = \frac{U}{\sqrt{\frac{V}{n}}}$$

má tzv. Studentovo rozdělení (nebo též **t-rozdělení**) o n stupních volnosti. Symbolem $t_\alpha(n)$ budeme značit $\alpha-$kvantily tohoto rozdělení.

(*vi*) Fisherovo–Snedecorovo rozdělení $X \sim F(m,n)$

Uvažujme nezávislé náhodné veličiny $U_{n_1} \sim \chi^2(n_1)$ a $U_{n_2} \sim \chi^2(n_2)$. Potom náhodná veličina

$$Z = \frac{\frac{U_{n_1}}{n_1}}{\frac{U_{n_2}}{n_2}}$$

má tzv. Fisherovo–Snedecorovo rozdělení o n_1, n_2 stupních volnosti. Symbolem $F_\alpha(n_1,n_2)$ budeme značit $\alpha-$kvantily tohoto rozdělení. Platí:

$$F_\alpha(n_1,n_2) = \frac{1}{F_{1-\alpha}(n_2,n_1)}.$$

Cvičení k tomuto tématu

(iv) Babička chodí k vnoučkovi na návštěvu náhodně mezi první a šestou hodinou odpolední. Dnes však odchází vnouček ve čtyři hodiny na fotbal. Jaká je pravděpodobnost, že babička vnoučka doma nezastihne? Dále spočtěte střední hodnotu, rozptyl a distribuční funkci příslušné náhodné veličiny.

(v) Nechť $X \sim \mathcal{N}(2, 16)$. Pomocí tabulky distribuční funkce $\Phi(x)$ určete

(a) $\mathbb{P}(-4 \leq X < 5)$,

(b) $\mathbb{P}(X > 8, 5)$,

(c) $\mathbb{P}(X = 3)$,

(d) $\mathbb{P}(X < 18)$.

(vi) Náhodná veličina X udává odchylku obsahu cukru v jednotlivých bonbónech od normy (udaná v gramech). Ví se, že $X \sim \mathcal{N}(0, 1)$. Výrobce potřebuje znát takovou hodnotu h (v gramech), že bonbónů, v nichž obsah cukru převýší normu o více než h, bude právě

(a) 1%,

(b) 12%.

Dále výrobce potřebuje znát takové h, že bonbónů, v nichž se obsah cukru bude lišit od normy o více než h, bude právě 1%.

(vii) Nechť $X \sim \mathcal{N}(3, 10)$ a $Y \sim \mathcal{N}(1, 9)$ jsou nezávislé. Definujme náhodné veličiny

$$U := 3X - 1, \; V := 2X - 3Y + 2.$$

Hledejme

(a) $\mathbb{P}(U < 15)$,

(b) $\mathbb{P}(V > 2)$,

(c) $\mathbb{P}(|V| \leq 6)$,

(d) $\rho(X, Y)$ a $\rho(U, X)$,

(e) $\rho(V, X)$ a $\rho(U, V)$.

(viii) Na toalety ve čtvrtém patře budovy C přichází „uživatel“ v průměru jednou za tři minuty. Na těchto toaletách se nachází také sprchy a právě teď tu nikdo není. Jaká je pravděpodobnost, že během následujících deseti minut nikdo nepřijde a Vy se tak stihnete v klidu vysprchovat? Jaká je pravěpodobnost, že budete vyrušeni už za pět minut?

(dcv) Na trávníku o rozloze 50×100 m^2 se během sezóny udělá přibližně 120 krtinců na zcela náhodných místech. Majitel na něm chce udělat green o ploše 100 m^2. Jaká je pravděpodobnost, že na greenu

(a) nebude za sezónu žádný krtinec?

(b) budou za sezónu více než tři krtince?

Nakrestele pravděpodobnostní funkci příslušné náhodné veličiny.