

# MATEMATIKA PRO EKONOMIKU

Kateřina STAŇKOVÁ HELISOVÁ

# Obsah

<b>1</b>	<b>Základy pravděpodobnosti</b>	<b>5</b>
1.1	Základní pojmy . . . . .	5
1.1.1	Speciální typy pravděpodobnostních prostorů . . . . .	7
1.1.2	Podmíněná pravděpodobnost . . . . .	8
1.1.3	Nezávislost . . . . .	11
1.2	Náhodná veličina . . . . .	11
1.2.1	Charakteristiky náhodných veličin . . . . .	14
1.2.2	Příklady diskrétních náhodných veličin . . . . .	17
1.2.3	Příklady spojitých náhodných veličin . . . . .	19
1.3	Funkce náhodných veličin . . . . .	22
1.3.1	Jedna veličina funkcí druhé . . . . .	22
1.3.2	Součet náhodných veličin . . . . .	23
1.3.3	Náhodný vektor . . . . .	25
1.3.4	Nezávislost náhodných veličin . . . . .	27
1.4	Zákony velkých čísel, centrální limitní věta . . . . .	28
1.4.1	Zákony velkých čísel . . . . .	29
1.4.2	Centrální limitní věta . . . . .	29

<b>2</b>	<b>Základy statistiky</b>	<b>31</b>
2.1	Bodový odhad . . . . .	35
2.1.1	Metoda momentů . . . . .	36
2.1.2	Metoda maximální věrohodnosti . . . . .	37
2.2	Intervalový odhad . . . . .	38
2.2.1	Intervalové odhady parametrů normálního rozdělení . . . . .	38
2.2.2	Intervalové odhady založené na CLV . . . . .	39
2.3	Testování hypotéz . . . . .	40
2.3.1	Princip testování hypotéz . . . . .	40
2.3.2	Testování pomocí intervalových odhadů . . . . .	41
2.3.3	Testování střední hodnoty normálního rozdělení - $t$ -testy . . . . .	41
2.3.4	$\chi^2$ test dobré shody . . . . .	43
2.3.5	Test nezávislosti v kontingenční tabulce . . . . .	44
<b>3</b>	<b>Regresní analýza</b>	<b>46</b>
3.1	Metoda nejmenších čtverců . . . . .	46
3.2	Lineární regrese . . . . .	47
3.2.1	Bodový odhad parametrů modelu . . . . .	47
3.2.2	Intervalové odhady parametrů modelu . . . . .	48
3.2.3	Testování podmodelů . . . . .	49
3.2.4	Analýza reziduí a odlehlá pozorování . . . . .	50
3.2.5	Výběr vhodného modelu . . . . .	51
<b>4</b>	<b>Náhodné procesy</b>	<b>52</b>

4.1	Základní pojmy . . . . .	52
4.2	Markovské řetězce s diskrétním časem . . . . .	54
4.2.1	Základní definice a vlastnosti . . . . .	54
4.2.2	Klasifikace stavů Markovského řetězce . . . . .	56
4.2.3	Stacionární rozdělení . . . . .	58
4.3	Markovské řetězce se spojitým časem . . . . .	59
4.3.1	Základní definice a vlastnosti . . . . .	59
4.3.2	Klasifikace stavů Markovského řetězce . . . . .	60
4.3.3	Stacionární rozdělení . . . . .	61
4.4	Nejčastěji používané Markovské procesy . . . . .	62
4.5	Náhodné procesy ve finanční matematice . . . . .	63
4.5.1	Základní pojmy . . . . .	64
4.5.2	Martingal . . . . .	65
4.5.3	Stochastický integrál . . . . .	65
4.5.4	Itôův proces a jeho diferenciál, Itôova formule . . . . .	66
4.5.5	Stochastická diferenciální rovnice . . . . .	67
<b>5</b>	<b>Pojistná matematika</b>	<b>69</b>
5.1	Neživotní pojištění . . . . .	69
5.1.1	Výpočet pojistného . . . . .	69
5.1.2	Modelování celkové výše škod . . . . .	70
5.1.3	Technické rezervy . . . . .	72
5.2	Životní pojištění . . . . .	76

5.2.1	Modelování úmrtnosti . . . . .	76
5.2.2	Komutační čísla . . . . .	78
5.2.3	Druhy pojištění . . . . .	79
<b>6</b>	<b>Shluková analýza dat</b>	<b>82</b>
6.1	Vstupní data . . . . .	82
6.1.1	Typy proměnných . . . . .	82
6.1.2	Normování hodnot proměnných . . . . .	83
6.2	Měření podobnosti . . . . .	84
6.3	Metody shlukové analýzy . . . . .	85
6.3.1	Metody rozkladu . . . . .	85
6.3.2	Hierarchické metody . . . . .	86

# Kapitola 1

## Základy pravděpodobnosti

### 1.1 Základní pojmy

**Definice 1.1** *Nechť  $\mathcal{A}$  je neprázdný systém podmnožin množiny  $\Omega \neq \emptyset$  takový, že*

- a)  $\emptyset \in \mathcal{A}$ ,
- b) *je-li  $A \in \mathcal{A}$ , pak  $A^c \in \mathcal{A}$ , kde  $A^c$  značí doplněk množiny  $A$  do  $\Omega$ .*
- c) *jsou-li  $A_i \in \mathcal{A}$ ,  $i = 1, 2, \dots$ , pak  $\cup_{i=1}^{\infty} A_i \in \mathcal{A}$ .*

*Pak  $\mathcal{A}$  nazýváme  $\sigma$ -algebrou.*

**Definice 1.2** *Nechť  $\Omega \neq \emptyset$  a  $\mathcal{A}$  je  $\sigma$ -algebra definovaná na  $\Omega$ . Pak pravděpodobností nazveme reálnou funkci  $P$  definovanou na  $\mathcal{A}$ , která splňuje*

- a)  $P(\Omega) = 1$ ,  $P(\emptyset) = 0$ ,
- b)  $P(A) > 0$  pro všechna  $A \in \mathcal{A}$ ,
- c) *pro každou posloupnost disjunktních jevů  $\{A_n\}_{n=1}^{\infty}$  platí*

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i).$$

*Trojice  $(\Omega, \mathcal{A}, P)$  se nazývá pravděpodobnostní prostor.*

### Terminologie:

- 1)  $\emptyset$  ... jev nemožný
- 2)  $\Omega$  ... jev jistý
- 3)  $A \cup B$  ... sjednocení jevů  $A, B$  (jev, který nastane právě tehdy, nastane-li aspoň jeden z jevů  $A, B$ )
- 4)  $A \cap B$  ... průnik jevů  $A, B$  (jev, který nastane právě tehdy, nastanou-li oba dva jevy současně)
- 5)  $B - A$  ... rozdíl jevu  $B$  a  $A$  (jev, který nastane právě tehdy, když nastane jev  $B$  a zároveň nenastane jev  $A$ )
- 6)  $A \subset B$  ...  $A$  je podjev jevu  $B$  (jev  $A$  nastane, kdykoliv nastane jev  $B$ )
- 7)  $A^c = \Omega - A$  ... doplněk jevu  $A$  (jev, který nastane právě tehdy, když nenastane jev  $A$ )
- 8)  $A \cap B = \emptyset$  ... jevy  $A, B$  jsou disjunktní (nemohou nastat současně)

### Vlastnosti pravděpodobnosti:

- 1)  $0 \leq P(A) \leq 1, \quad \forall A \in \mathcal{A},$
- 2)  $P$  je monotónní:  $A, B \in \mathcal{A}, A \subset B \Rightarrow P(A) \leq P(B),$
- 3)  $P(A^c) = 1 - P(A), \quad \forall A \in \mathcal{A},$
- 4)  $P(A \cup B) = P(A) + P(B) - P(A \cap B), \quad \forall A, B \in \mathcal{A},$
- 5)  $A, B \in \mathcal{A}, A \subset B \Rightarrow P(B - A) = P(B) - P(A),$
- 6) pro každou posloupnost disjunktních jevů  $\{A_i\}_{i=1}^{\infty}$  takových, že  $\cup_{i=1}^{\infty} A_i = \Omega$ , platí  $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i) = 1.$

### 1.1.1 Speciální typy pravděpodobnostních prostorů

#### I. Klasický pravděpodobnostní prostor

Pravděpodobnostní prostor  $(\Omega, \mathcal{A}, P)$  nazveme klasickým pravděpodobnostním prostorem, jestliže

- a) množina  $\Omega$  je konečná a všechny možné výsledky jsou stejně pravděpodobné, tzn. označíme-li postupně  $p_1, \dots, p_m$  pravděpodobnosti jednotlivých výsledků elementárních jevů, pak  $p_1 = p_2 = \dots = p_m = \frac{1}{m}$  (je-li možných výsledků  $m$ ),
- b) za  $\sigma$ -algebru  $\mathcal{A}$  vezmeme systém všech podmnožin množiny  $\Omega$ ,
- c) pravděpodobnost  $P$  náhodného jevu  $A$  je rovna

$$P(A) = \frac{m_A}{m},$$

kde  $m_A$  je počet výsledků příznivých jevů  $A$  a  $m$  je počet všech možných výsledků náhodného pokusu.

#### II. Geometrický pravděpodobnostní prostor

Geometrickým pravděpodobnostním prostorem nazveme pravděpodobnostní prostor  $(\Omega, \mathcal{A}, P)$  takový, že

- a)  $\Omega \subset \mathbb{R}^d$  (obvykle  $d = 1, 2, 3$ ), neboli všechny elementární jevy lze vyjádřit jako body nějaké množiny,
- b)  $\mathcal{A} = \mathcal{B}(\Omega)$  je Borelovská  $\sigma$ -algebra na  $\Omega$  (tj. nejmenší  $\sigma$ -algebra obsahující všechny otevřené podmnožiny  $\Omega$ , tudíž i všechny uzavřené podmnožiny a kombinace obou těchto typů),
- c)  $P(A) = \frac{\mu^d(A)}{\mu^d(\Omega)}$ , kde  $\mu^d$  je  $d$ -rozměrná Lebesgueova míra. Pro naše účely postačí, pokud si pod  $\mu^1(A)$  představíme délku množiny  $A$ , pod  $\mu^2(A)$  obsah  $A$  a pod  $\mu^3(A)$  objem  $A$ .



### III. Obecný diskretní pravděpodobnostní prostor

Obecný diskretní pravděpodobnostní prostor je pravděpodobnostní prostor  $(\Omega, \mathcal{A}, P)$  takový, že

- a)  $\Omega = \{\omega_1, \omega_2, \dots\}$ ,
- b)  $\mathcal{A}$  je množina všech podmnožin  $\Omega$ ,
- c) jsou dány pravděpodobnosti elementárních jevů  $P(\omega_i)$ , které splňují:  $\sum_{i=1}^{\infty} P(\omega_i) = 1$ . Pak pravděpodobnost libovolného jevu je dána jednoznačně vztahem  $P(A) = \sum_{\omega_i \in A} P(\omega_i)$ .

### IV. Obecný spojitý pravděpodobnostní prostor

O spojitém pravděpodobnostním prostoru mluvíme tehdy, když

- a)  $\Omega = \mathbb{R}$ , neboli všechny elementární jevy lze vyjádřit jako body na reálné ose,
- b)  $\mathcal{A} = \mathcal{B}(\mathbb{R})$  je Borelovská  $\sigma$ -algebra nad  $\mathbb{R}$ ,
- c) Je dána funkce  $f: \mathbb{R} \rightarrow [0, \infty]$  taková, že  $\int_{\mathbb{R}} f(x)dx = 1$ . Pak pravděpodobnost libovolného jevu  $A \in \mathcal{A}$  je dána jednoznačně vztahem

$$P(A) = \int_A f(x)dx.$$

#### 1.1.2 Podmíněná pravděpodobnost

**Definice 1.3** *Nechť je dán pravděpodobnostní prostor  $(\Omega, \mathcal{A}, P)$  a náhodné jevy  $A, B$ , kde  $P(B) > 0$ . Podmíněnou pravděpodobnost jevu  $A$  za podmínky, že nastal jev  $B$ , definujeme vztahem*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \quad (1.1)$$

**Věta 1.1** *Nechť je dán pravděpodobnostní prostor  $(\Omega, \mathcal{A}, P)$  a náhodný jev  $B$ , kde  $P(B) > 0$ . Potom pro libovolný jev  $A \in \mathcal{A}$  platí:*

- a)  $P(A|B) \geq 0$ ,

b)  $P(\Omega|B) = 1,$

c)  $P(\cup_{n=1}^{\infty} A_n|B) = \sum_{n=1}^{\infty} P(A_n|B)$  pro každou posloupnost  $\{A_n\}$  disjunktních jevů.

*Důkaz.*

a) zřejmé,

b) z definice 1.3 plyne, že

$$P(\Omega|B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1,$$

c) protože  $A_1, A_2, \dots$  jsou disjunktní, tak i  $A_1 \cap B, A_2 \cap B, \dots$  jsou disjunktní. Z axiomu c) definice 1.2 a z definice 1.3 plyne

$$\begin{aligned} P(\cup_{n=1}^{\infty} A_n|B) &= \frac{P(\cup_{n=1}^{\infty} A_n \cap B)}{P(B)} = \frac{\sum_{n=1}^{\infty} P(A_n \cap B)}{P(B)} = \\ &= \sum_{n=1}^{\infty} P(A_n|B) \end{aligned}$$

□

**Poznámka 1.1** *Věta 1.1 říká, že podmíněná pravděpodobnost má všechny vlastnosti pravděpodobnosti nepodmíněné.*

**Věta 1.2 (o násobení pravděpodobnosti):**

*Pro libovolnou posloupnost náhodných jevů  $A_1, A_2, \dots, A_n$ ,  $P(A_1 \cap A_2 \cap \dots \cap A_{n-1}) > 0$  platí*

$$\begin{aligned} P(\cap_{i=1}^n A_i) &= P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \dots \\ &\dots P(A_n|A_1 \cap A_2 \cap \dots \cap A_{n-1}). \end{aligned} \tag{1.2}$$

*Důkaz.* Opakovaným použitím definice 1.3 podmíněné pravděpodobnosti dostáváme:

$$\begin{aligned} P(\cap_{i=1}^{n-1} A_i \cap A_n) &= P(\cap_{i=1}^{n-1} A_i)P(A_n|\cap_{i=1}^{n-1} A_i) = \\ &= P(\cap_{i=1}^{n-2} A_i)P(A_{n-1}|\cap_{i=1}^{n-2} A_i)P(A_n|\cap_{i=1}^{n-1} A_i) \dots \\ &= P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \dots P(A_n|\cap_{i=1}^{n-1} A_i). \end{aligned}$$

Vzhledem k monotonnii pravděpodobnosti a předpokladu věty máme

$$P(A_1) \geq P(A_1 \cap A_2) \geq \dots \geq P(A_1 \cap \dots \cap A_{n-1}) > 0$$

a tedy všechny podmíněné pravděpodobnosti v tvrzení věty jsou dobře definovány.

**Věta 1.3 (O celkové pravděpodobnosti)** *Nechť  $A_1, A_2, \dots$  jsou náhodné jevy tvořící rozklad jevu jistého, tzn.*

$$A_i \cap A_j = \emptyset, \quad \forall i \neq j \quad \text{a} \quad \bigcup_{i=1}^{\infty} A_i = \Omega.$$

*Nechť tyto náhodné jevy mají postupně pravděpodobnosti  $P(A_1), P(A_2), \dots$ , přičemž  $P(A_i) > 0, \forall i = 1, 2, \dots$ . Uvažujme libovolný náhodný jev  $B$ , u něhož známe podmíněné pravděpodobnosti*

$$P(B|A_i), \quad \forall i = 1, 2, \dots$$

*Potom*

$$P(B) = \sum_{i=1}^{\infty} P(A_i) \cdot P(B|A_i). \quad (1.3)$$

*Důkaz.* Jevy  $A_1, \dots, A_n$  tvoří disjunktní rozklad  $\Rightarrow (A_i \cap B) \cap (A_j \cap B) = \emptyset \quad \forall i \neq j$ ,  $\bigcup_{i=1}^{\infty} (A_i \cap B) = B$ . Potom

$$P(B) = P(\bigcup_{i=1}^{\infty} (A_i \cap B)) = \sum_{i=1}^{\infty} P(A_i \cap B) = \sum_{i=1}^{\infty} P(A_i) \cdot P(B|A_i).$$

□

**Věta 1.4 (Bayesova věta)** *Nechť jsou splněny předpoklady věty 1.3. Pak*

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{j=1}^{\infty} P(A_j) \cdot P(B|A_j)}, \quad i = 1, 2, \dots \quad (1.4)$$

*Důkaz.* Podle vzorce pro podmíněnou pravděpodobnost je

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)}. \quad (1.5)$$

Po dosazení (1.3) do (1.5) dostáváme

$$P(A_i|B) = \frac{P(A_i \cap B)}{\sum_{j=1}^{\infty} P(A_j) \cdot P(B|A_j)} = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{j=1}^{\infty} P(A_j) \cdot P(B|A_j)}. \quad (1.6)$$

□

### 1.1.3 Nezávislost

**Definice 1.4** Náhodné jevy  $A$  a  $B$  jsou nezávislé, jestliže platí

$$P(A \cap B) = P(A) \cdot P(B). \quad (1.7)$$

Pojem nezávislosti můžeme rozšířit i na skupinu náhodných jevů.

**Definice 1.5** Necht'  $A_1, A_2, \dots, A_n$  jsou náhodné jevy. Řekneme, že jsou skupinově (totálně) nezávislé, jestliže pro libovolnou posloupnost indexů  $\{k_1, k_2, \dots, k_r\} \subset \{1, \dots, n\}$ ,  $r = 2, \dots, n$  platí

$$P(A_{k_1} \cap A_{k_2} \cap \dots \cap A_{k_r}) = P(A_{k_1}) \cdot P(A_{k_2}) \cdot \dots \cdot P(A_{k_r}). \quad (1.8)$$

**Definice 1.6** Necht'  $A_1, \dots, A_n$  jsou náhodné jevy. Řekneme, že jsou po dvou nezávislé, jestliže jevy  $A_i, A_j$  jsou nezávislé pro všechna  $i, j = 1, \dots, n$ ,  $i \neq j$ .

**Věta 1.5** Necht'  $A, B$  jsou nezávislé náhodné jevy. Pak dvojice jevů  $(A, B^c)$ ,  $(A^c, B)$ ,  $(A^c, B^c)$  jsou nezávislé.

*Důkaz.*

$$\begin{aligned} P(A^c \cap B) &= P(B - A) = P(B - [A \cap B]) = P(B) - P(A \cap B) = \\ &= P(B) - P(B) \cdot P(A) = P(B) \cdot (1 - P(A)) = P(B) \cdot P(A^c). \end{aligned}$$

Nezávislost jevů  $A, B^c$  se dokáže analogicky.

Jsou-li nezávislé jevy  $A, B$ , pak podle předchozího jsou nezávislé jevy  $A, B^c$ , ale odtud opět podle předchozího jsou nezávislé i jevy  $A^c, B^c$ .

□

## 1.2 Náhodná veličina

**Definice 1.7** Necht'  $(\Omega, \mathcal{A}, P)$  je pravděpodobnostní prostor. Reálnou funkci  $X$  definovanou na  $\Omega$  nazýváme náhodnou veličinou, jestliže  $X$  je měřitelné zobrazení  $X : (\Omega, \mathcal{A}) \rightarrow (\mathbb{R}, \mathcal{B})$ , tj.

$$\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{A} \quad (1.9)$$

pro libovolnou borelovskou množinu  $B \in \mathcal{B}$  ( $\mathcal{B}$  je  $\sigma$ -algebra borelovských podmnožin, tj. nejmenší  $\sigma$ -algebra obsahující systém všech otevřených podmnožin  $\mathbb{R}$ ).

**Poznámka 1.2** Náhodné veličiny budeme značit velkými písmeny  $X, Y, Z \dots$ . Hodnoty, kterých mohou náhodné veličiny nabývat, budeme značit malými písmeny  $x, y, z \dots$ .

**Poznámka 1.3** Místo  $\{\omega \in \Omega : X(\omega) \in B\}$  budeme zjednodušeně psát  $\{X \in B\}$  a místo  $\{\omega \in \Omega : X(\omega) < x\}$  budeme zjednodušeně psát  $\{X < x\}$ .

**Poznámka 1.4** Součty, součiny, podíly, minima a maxima náhodných veličin jsou opět náhodné veličiny; umocnění náhodné veličiny přirozeným číslem, násobení náhodné veličiny skalárem jsou také náhodné veličiny.

**Definice 1.8** Necht'  $X$  je náhodná veličina. Její distribuční funkci nazýváme reálnou funkcí  $F_X$  reálné proměnné  $x$  definovanou

$$F_X(x) = P(X \leq x) = P(\{\omega : X(\omega) \leq x\}), \quad x \in \mathbb{R}. \quad (1.10)$$

**Vlastnosti distribuční funkce:**

Distribuční funkce  $F_X(x)$  náhodné veličiny  $X$  je

- a) neklesající, tj. pro libovolné  $a, b \in \mathbb{R}, a \leq b$ , platí  $F_X(a) \leq F_X(b)$ ,
- b) zprava spojitá v libovolném bodě  $x \in \mathbb{R}$ ,
- c)  $\lim_{x \rightarrow -\infty} F_X(x) = 0, \lim_{x \rightarrow \infty} F_X(x) = 1$ .

**Diskrétní náhodná veličina**

**Definice 1.9** Náhodná veličina  $X$  se nazývá diskrétní (nebo také říkáme, že  $X$  má diskrétní rozdělení), jestliže existuje (konečná nebo spočetná) posloupnost reálných čísel  $\{x_n\}$  a odpovídající posloupnost nezáporných čísel  $\{p_n\} = P(X = x_n)$  (obvykle  $p_n > 0$  pro všechna  $n$ ) taková, že

$$\sum_{n=1}^{\infty} p_n = 1. \quad (1.11)$$

Distribuční funkce diskrétní náhodné veličiny  $X$  má tvar

$$F_X(x) = P(X \leq x) = \sum_{\{n: x_n \leq x\}} P(X = x_n) = \sum_{\{n: x_n \leq x\}} p_n \quad (1.12)$$

a

$$P(a < X \leq b) = F_X(b) - F_X(a) = \sum_{\{n:a < x_n \leq b\}} P(X = x_n) = \sum_{\{n:a < x_n \leq b\}} p_n$$

pro libovolná reálná čísla  $a, b$ , kde  $a \leq b$ .

Distribuční funkce je tedy schodovitá funkce se skoky v bodech  $x_1, x_2, \dots$  a je konstantní na intervalech  $[x_n, x_{n+1})$ . Velikost skoku v bodě  $x_n$  je  $p_n = P(X = x_n)$ .

### Absolutně spojitá náhodná veličina

**Definice 1.10** Náhodná veličina  $X$  se nazývá absolutně spojitá (nebo také říkáme, že  $X$  má absolutně spojitě rozdělení), jestliže existuje nezáporná integrovatelná funkce  $f_X$  taková, že platí

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(t) dt, \quad x \in (-\infty, \infty). \quad (1.13)$$

Funkce  $f_X$  se nazývá hustotou rozdělení pravděpodobnosti.

**Poznámka 1.5** Místo " $P(X \text{ má vlastnost } V) = 1$ " budeme říkat " $X$  má vlastnost  $V$  skoro jistě." Často budeme užívat zkratku s.j.

### Vlastnosti hustoty:

a)  $f_X(x) = \frac{d}{dx} F_X(x)$  s.j.,

b)  $\int_{-\infty}^{\infty} f_X(x) dx = 1$ ,

c)  $P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x) dx$  pro libovolná reálná čísla  $a, b$ , kde  $a \leq b$ .

### Zobecnění

**Definice 1.11** Míra je množinová fce na  $(\Omega, \mathcal{A})$ , tj.

(i)  $\mu : \mathcal{A} \rightarrow [0, \infty]$ ,

(ii)  $\mu(\emptyset) = 0$ ,

(iii) jsou-li  $A_n \in \mathcal{A}, n \geq 1$  disjunktní, pak  $\mu(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$ .

Je-li  $\mu(\Omega) = 1$ , říkáme  $\mu$  pravděpodobnostní míra.

Každé náhodné veličině  $X$  a borelovské množině  $B \in \mathcal{B}$  pak lze přiřadit pravděpodobnostní míru na  $(\mathbb{R}, \mathcal{B})$

$$\mu_X(B) = P(\{\omega \in \Omega : X(\omega) \in B\}),$$

kteřou nazýváme *rozdělení pravděpodobnosti* náhodné veličiny  $X$ . Položíme-li speciálně  $B = (-\infty, x]$ , dostáváme

$$\mu_X(B) = P(\{\omega \in \Omega : X(\omega) \leq x\}) = F_X(x),$$

což není nic jiného než distribuční funkce. Vidíme tedy, že mezi distribuční funkcí a rozdělením pravděpodobnosti existuje vzájemně jednoznačný vztah. Speciálně pak, položíme-li  $B = (a, b]$ ;  $-\infty < a \leq b < \infty$ , dostáváme

$$P(X \in (a, b]) = F_X(b) - F_X(a) = \mu_X((a, b]).$$

Uvedený vztah je definován pouze pro intervaly, ovšem to nám stačí pro jednoznačné definování míry pro všechny borelovské množiny, tudíž platí

$$P(X \in B) = \mu_X(B) = \int_B 1 d\mu_X(x) = \int_B 1 dF_X(x), \quad \forall B \in \mathcal{B}.$$

Tento integrál je tedy zobecněním klasického integrálu  $\int 1 dx$ , v němž ale  $x$  mají různou váhu, a sumy, kde jednotlivé sčítance mají také různou váhu.

Tyto úvahy nás proto vedly k rozdělení náhodných veličin na dva základní typy - na diskrétní a absolutní spojitě náhodné veličiny. V praxi se však mohou vyskytovat i jejich kombinace.

## 1.2.1 Charakteristiky náhodných veličin

### Střední hodnota

**Definice 1.12** *Nechť  $X$  je náhodná veličina definovaná na pravděpodobnostním prostoru  $(\Omega, \mathcal{A}, P)$ .*

a) Nechť  $X$  je diskrétní náhodná veličina nabývající reálných hodnot  $x_1, x_2, x_3, \dots$ , tzn. taková, že  $P(X = x_i) = p_i$ . Pak střední hodnota  $\mathbb{E}X$  náhodné veličiny  $X$  je tvaru

$$\mathbb{E}X = \sum_{i=1}^{\infty} x_i \cdot p_i, \quad (1.14)$$

pokud řada v (1.14) konverguje.

b) Nechť  $X$  je absolutně spojitá náhodná veličina s hustotou  $f_X$ . Pak střední hodnota náhodné veličiny  $X$  je

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f_X(x) dx, \quad (1.15)$$

pokud integrál existuje.

Obecně řečeno, střední hodnotou  $\mathbb{E}X$  náhodné veličiny  $X$  nazveme integrál

$$\mathbb{E}X = \int_{-\infty}^{\infty} x dF_X(x), \quad (1.16)$$

pokud tento integrál existuje.

### Základní vlastnosti střední hodnoty

1)  $\mathbb{E}a = a$

2)  $\mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y$

3)  $X_1 \leq X \leq X_2$  s.j.  $\Rightarrow \mathbb{E}X_1 \leq \mathbb{E}X \leq \mathbb{E}X_2$ , speciálně  $X \geq 0$  s.j.  $\Rightarrow \mathbb{E}X \geq 0$

**Věta 1.6** Nechť  $X$  je náhodná veličina definovaná na pravděpodobnostním prostoru  $(\Omega, \mathcal{A}, P)$  a nechť  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ .

a) Má-li náhodná veličina  $X$  diskrétní rozdělení  $\{x_n, p_n\}_{n \in \mathbb{N}_0}$ , pak

$$\mathbb{E}\varphi(X) = \sum_{n \in \mathbb{N}_0} \varphi(x_n) p_n, \quad (1.17)$$

pokud jedna ze stran rovnosti existuje.

b) Má-li náhodná veličina  $X$  absolutně spojitě rozdělení s hustotou  $f$ , potom

$$\mathbb{E}\varphi(X) = \int_{-\infty}^{\infty} \varphi(x) f(x) dx, \quad (1.18)$$

pokud jeden z integrálů existuje.



Obecně napsáno,

$$\mathbb{E}\varphi(X) = \int_{-\infty}^{\infty} \varphi(x) dF_X(x), \quad (1.19)$$

pokud jeden z integrálů existuje.

## Momenty, rozptyl a kovariance

**Definice 1.13** Nechť  $X$  je náhodná veličina. Hodnota

$\mathbb{E}X^n$  se nazývá  $n$ -tý moment náhodné veličiny  $X$ ,

$\mathbb{E}(X - \mathbb{E}X)^n$  se nazývá  $n$ -tý centrální moment náhodné veličiny  $X$ ,

$\mathbb{E}|X - \mathbb{E}X|$  se nazývá absolutní moment náhodné veličiny  $X$ .

**Definice 1.14** Druhý centrální moment se nazývá rozptyl, značí se  $\text{var } X = \mathbb{E}(X - \mathbb{E}X)^2$ .

**Definice 1.15** Nechť  $X, Y$  jsou náhodné veličiny takové, že  $\mathbb{E}X^2 < \infty$  a  $\mathbb{E}Y^2 < \infty$ . Pak jejich kovariance je definována jako

$$\text{cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y).$$

**Poznámka 1.6** Z předešlých dvou definic vyplývá souvislost mezi rozptylem a kovariancí, a to  $\text{cov}(X, X) = \text{var}(X)$ .

## Vlastnosti rozptylu a kovariance

- 1) Nechť  $X$  je náhodná veličina. Pak  $\text{var } X = \mathbb{E}(X^2) - (\mathbb{E}X)^2$
- 2) Nechť  $c$  je konstanta. Pak  $\text{var } c = 0$ .
- 3) Nechť  $X$  je náhodná veličina a  $a$  je reálné číslo. Pak  $\text{var } (aX) = a^2 \text{var } X$ .
- 4) Nechť  $X$  je náhodná veličina a  $c$  je konstanta. Pak  $\text{var } (X + c) = \text{var } X$ .
- 5) Nechť  $X$  je náhodná veličina, která má konečnou střední hodnotu a konečný nenulový rozptyl. Nechť

$$Z = \frac{X - \mathbb{E}X}{\sqrt{\text{var } X}}.$$

Pak  $\mathbb{E}Z = 0$  a  $\text{var } Z = 1$ .

- 6) Nechť  $X, Y$  jsou náhodné veličiny. Pak  $\text{var } (X + Y) = \text{var } X + \text{var } Y + 2\text{cov}(X, Y)$ .

7) Necht  $X, Y$  jsou náhodné veličiny. Pak  $\text{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y$ .

**Věta 1.7 (Čebyševova nerovnost)** Necht  $X$  je náhodná veličina s konečným rozptylem. Pak pro libovolné  $\varepsilon > 0$  platí

$$P[|X - \mathbb{E}X| \geq \varepsilon] \leq \frac{\text{var } X}{\varepsilon^2}.$$

*Důkaz.* Uvažujme náhodnou veličinu  $Y = X - \mathbb{E}X$ . Pak

$$\begin{aligned} \text{var } X &= \mathbb{E}Y^2 = \int_{-\infty}^{\infty} x^2 dF_Y(x) \geq \int_{|x| \geq \varepsilon} x^2 dF_Y(x) \geq \\ &\geq \varepsilon^2 \int_{|x| \geq \varepsilon} dF_Y(x) = \varepsilon^2 P[|Y| \geq \varepsilon]. \end{aligned}$$

□

## 1.2.2 Příklady diskretních náhodných veličin

### 1. Alternativní rozdělení ( $Alt(p)$ )

má náhodná veličina  $X$ , která nabývá jen hodnot 0 a 1 s pravděpodobnostmi  $1 - p$  a  $p$ . Číslo  $p$  se nazývá parametr alternativního rozdělení,  $0 < p < 1$ .

Distribuční funkce je

$$F(x) = \begin{cases} 0 & \text{pro } x < 0 \\ 1 - p & \text{pro } 0 \leq x < 1 \\ 1 & \text{pro } x \geq 1 \end{cases}$$

Snadno se z definic spočte, že střední hodnota  $\mathbb{E}X = p$  a rozptyl  $\text{var } X = p(1 - p)$ .

**Použití:** Příkladem alternativně rozdělené náhodné veličiny je např. počet jedniček, které padnou při jednom hození kostkou, počet zmetků při náhodném výběru jednoho výrobku atd.

### 2. Binomické rozdělení ( $Bi(n, p)$ )

je rozdělení náhodné veličiny  $X$ , která nabývá hodnot  $k = 0, 1, 2, \dots, n$ . Binomické rozdělení je jednoznačně určeno dvěma parametry: přirozeným číslem  $n$  a číslem  $p \in (0, 1)$ .

Pravděpodobnosti  $p_k$  jsou tvaru

$$p_k = \binom{n}{k} p^k (1 - p)^{n-k}, \text{ pro } k = 0, 1, \dots, n.$$

Distribuční funkce  $F(x)$  je

$$F(x) = \begin{cases} 0 & x < 0 \\ \sum_{0 \leq k \leq x} \binom{n}{k} p^k (1 - p)^{n-k} & 0 \leq x < n \\ 1 & x \geq n. \end{cases}$$

Rozdělení  $Bi(n, p)$  má střední hodnotu  $np$ , neboť

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

K výpočtu rozptylu použijeme vztah

$$\text{var } X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \mathbb{E}X(X-1) + \mathbb{E}X - (\mathbb{E}X)^2$$

a provedeme výpočet  $\mathbb{E}X(X-1) = \sum_{k=0}^n k(k-1) \binom{n}{k} p^k (1-p)^{n-k}$  analogický tomu pro střední hodnotu. Dosazením pak získáme hodnotu pro rozptyl  $np(1-p)$ .

**Použití:** Binomickým rozdělením se řídí např. náhodná veličina  $X$ , která je rovna počtu úspěchů v posloupnosti  $n$  nezávislých alternativních pokusů, kde pravděpodobnost úspěchu v každém pokuse je  $p$ ,  $0 < p < 1$ . Tedy

$$X = \sum_{i=1}^n X_i,$$

kde

$$X_i = \begin{cases} 1 & \text{pokud v } i\text{-tém pokuse nastal úspěch,} \\ 0 & \text{pokud úspěch nenastal.} \end{cases}$$

$X$  je součtem  $n$  alternativních náhodných veličin.

### 3. Poissonovo rozdělení ( $Po(\lambda)$ )

je rozdělení náhodné veličiny  $X$ , která nabývá hodnot  $k = 0, 1, 2, \dots$  s pravděpodobnostmi

$$p_k = e^{-\lambda} \frac{\lambda^k}{k!}.$$

Číslo  $\lambda > 0$  je parametr Poissonova rozdělení.

Distribuční funkce je

$$F(x) = \begin{cases} 0 & \text{pro } x \leq 0 \\ \sum_{0 \leq j \leq x} e^{-\lambda} \frac{\lambda^j}{j!} & \text{pro } 0 \leq x < \infty. \end{cases}$$

Střední hodnota je  $\lambda$ . Rozptyl je rovněž roven  $\lambda$ . Jejich výpočet je obdobou výpočtu charakteristik binomického rozdělení

**Poznámka 1.7** *Poissonovo rozdělení je limitním případem binomického rozdělení pro  $n \rightarrow \infty, p \rightarrow 0, np \rightarrow \lambda$  (=konstanta).*

**Použití:** Tímto rozdělením se řídí např. náhodná veličina, kterou je počet výskytů sledovaného jevu v určitém časovém intervalu délky  $t$  (předpokládejme, že jev může nastat v kterémkoliv okamžiku a počet výskytů během časového intervalu závisí jen na jeho délce a ne na jeho počátku ani na tom, kolikrát jev nastal před jeho počátkem), např. počet telefonních zavolaní během nějakého časového intervalu atd.

#### 4. Geometrické rozdělení ( $Geom(p)$ )

je rozdělení náhodné veličiny  $X$ , která nabývá hodnot  $k = 0, 1, 2, \dots$  s pravděpodobnostmi  $p_k = p(1-p)^k$ . Parametr  $p$  je z intervalu  $(0,1)$ . Distribuční funkce je

$$F(x) = \begin{cases} 0 & \text{pro } x < 0 \\ \sum_{0 \leq k \leq x} p(1-p)^k & \text{pro } x \geq 0. \end{cases}$$

Pomocí vzorců pro geometrické řady se spočte střední hodnota  $\frac{1-p}{p}$  a rozptyl  $\frac{1-p}{p^2}$ .

**Použití:** Předpokládejme, že provádíme nezávislé pokusy a že pravděpodobnost úspěchu v jednom pokuse je pro všechny pokusy stejná a je rovna  $p$ . Pak vidíme, že náhodná veličina (počet pokusů do prvního úspěchu) se řídí geometrickým rozdělením. Název tohoto rozdělení vyplývá ze skutečnosti, že s rostoucími hodnotami  $k$  pravděpodobnosti  $p_k$  geometricky klesají.

### 1.2.3 Příklady spojitých náhodných veličin

#### 1. Rovnoměrné rozdělení na intervalu $[a, b]$ ( $Ro(a, b)$ )

je dáno hustotou

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

Distribuční funkce je

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

Lehce se spočte, že střední hodnota a rozptyl jsou

$$\mathbb{E}X = \frac{a+b}{2}, \quad \text{var } X = \frac{1}{12}(b-a)^2.$$

**Použití:** S rovnoměrným rozdělením se setkáváme např. při vyšetřování chyb ze zaokrouhlování v numerických výpočtech. Jsou-li čísla vstupující do výpočtů nekonečné desetinné zlomky, jež se zaokrouhlují na  $k$  desetinných míst, pak lze chybu ze zaokrouhlení považovat za náhodnou veličinu s rovnoměrným rozdělením na intervalu  $[-5 \cdot 10^{-k-1}, 5 \cdot 10^{-k-1}]$ , ale také u spousty dalších jevů.

## 2. Exponenciální rozdělení ( $Exp(\lambda)$ )

je rozdělení s hustotou

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

kde  $\lambda$  je parametrem rozdělení.

Distribuční funkce je

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

Integrací per partes se spočte střední hodnota

$$\mathbb{E}X = \int_0^{\infty} x \lambda e^{-\lambda x} dx = \frac{1}{\lambda}.$$

Dvojitým použitím per partes získáme

$$\mathbb{E}X^2 = \int_0^{\infty} x^2 \lambda e^{-\lambda x} dx = \frac{2}{\lambda^2}$$

a následně rozptyl

$$\text{var } X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{1}{\lambda^2}.$$

**Použití:** Exponenciální rozdělení  $Exp(\lambda)$  je vhodným modelem "doby čekání" do nastoupení určitého jevu, např. doby životnosti určitého zařízení a to tehdy, jestliže rozdělení zbývajících doby čekání nezávisí na tom, jak dlouho již čekáme. Říká se tomu, že exponenciální rozdělení nemá paměť. Přesně je tato vlastnost popsána tvrzením: Má-li náhodná veličina  $X$  exponenciální rozdělení, pak

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

neboť  $P(X > x + y | X > y)$  můžeme podle definice podmíněné pravděpodobnosti přepsat na

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

**Poznámka 1.8** *Souvislost s Poissonovým rozdělením: Má-li náhodná veličina  $X$  popisující dobu čekání na událost rozdělení  $Exp(\lambda)$ , pak náhodná veličina  $Y$  popisující počet těchto událostí v časovém intervalu délky  $T$  má rozdělení  $Po(\lambda T)$ .*

## 3. Obecné normální rozdělení ( $N(\mu, \sigma^2)$ )

je definováno hustotou

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

kde  $\mu$  reálné a  $\sigma^2$  kladné jsou parametry.

Distribuční funkce je

$$F(x) = \frac{1}{\sqrt{2\pi} \sigma} \int_{-\infty}^x e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt, \quad -\infty < x < \infty. \quad (1.20)$$

Střední hodnota  $\mathbb{E}X = \mu$  a rozptyl  $\text{var } X = \sigma^2$ .

#### 4. Normované normální rozdělení ( $N(0, 1)$ )

je speciálním případem obecného normálního rozdělení pro  $\mu = 0$  a  $\sigma^2 = 1$ , tj. je definováno hustotou

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

Jeho distribuční funkce se tradičně značí písmenem  $\Phi$  a platí

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

Je zřejmé, že střední hodnota je rovna 0 a rozptyl je roven 1.

**Použití:** Normální rozdělení má mimořádný význam v teorii pravděpodobnosti a matematické statistice, přestože se tímto rozdělením řídí přesně jen málo náhodných veličin. V následujících odstavcích bude řečeno, že součet velkého počtu nezávislých náhodných veličin (o jejichž rozdělení se činí jen velmi obecné předpoklady) má přibližně normální rozdělení, tím lze vysvětlit klíčovou roli tohoto rozdělení v teorii pravděpodobnosti a matematické statistice. Náhodné veličiny, s nimiž se v reálném světě setkáváme, lze velmi často považovat za výslednice působení velkého počtu drobných náhodných vlivů. Pak lze očekávat, že normální rozdělení bude vhodným modelem pro takové náhodné veličiny. Nejběžnějším typem takových veličin jsou náhodné chyby (chyby měření, způsobené velkým počtem neznámých a vzájemně nezávislých příčin). Normální rozdělení je vhodným modelem pro řadu fyzikálních, technických a biologických veličin jako například tělesná výška jedinců homogenní populace, roční částka, kterou pojišťovna vyplatí za pojistné případy atd.

**Poznámka 1.9** *Jelikož se s normálním rozdělením velmi často pracuje a výpočet distribuční funkce je zdlouhavý, jsou hodnoty distribuční funkce  $N(0, 1)$  tabelovány. Vzhledem k symetrii funkce ( $\Phi(x) = 1 - \Phi(-x)$ ) se tabelují hodnoty  $\Phi$  pouze pro nezáporné  $x$ .*

#### Věta 1.8 (Transformace veličin s normálním rozdělením)

- a) Má-li  $X$  normované normální rozdělení a  $Y = \mu + \sigma X$ , pak  $Y$  má normální rozdělení s parametry  $\mu$  a  $\sigma^2$ .
- b) Má-li  $X$  normální rozdělení s parametry  $\mu, \sigma^2$  a je-li  $Y = a + bX$ , pak  $Y$  má opět normální rozdělení s parametry  $a + b\mu$  a  $b^2\sigma^2$ .
- c) Necht  $X, Y$  jsou náhodné veličiny,  $X$  má rozdělení  $N(\mu_1, \sigma_1^2)$ ,  $Y$  má rozdělení  $N(\mu_2, \sigma_2^2)$  a  $\text{cov}(X, Y) = 0$ . Potom  $Z = X + Y$  má rozdělení  $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ .

## 1.3 Funkce náhodných veličin

### 1.3.1 Jedna veličina funkcí druhé

Při řešení některých pravděpodobnostních úloh se setkáváme se situací, kdy známe rozdělení náhodné veličiny  $X$  a hledáme rozdělení náhodné veličiny  $Y$ , která je funkcí veličiny  $X$

$$Y = \varphi(X).$$

**Věta 1.9** Necht  $X$  je náhodná veličina s distribuční funkcí  $F$  a necht  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ . Označme  $Y = \varphi(X)$  a  $G$  její distribuční funkci. Potom

$$G(y) = \int_{\{x; \varphi(x) \leq y\}} dF(x), \quad \forall y \in \mathbb{R}. \quad (1.21)$$

Speciálně, je-li  $F$  diskrétní  $\{x_n, p_n\}$ , je

$$G(y) = \sum_{\{x_n; \varphi(x_n) \leq y\}} p_n, \quad \forall y \in \mathbb{R} \quad (1.22)$$

a je-li absolutně spojitá s hustotou  $f$ , je

$$G(y) = \int_{\{x; \varphi(x) \leq y\}} f(x) dx, \quad \forall y \in \mathbb{R}. \quad (1.23)$$

*Důkaz.* Označme  $B_y = \{x; \varphi(x) \leq y\}$ . Pak

$$G(y) = P(Y \leq y) = P(\varphi(X) \leq y) = P(X \in B_y) = \int_{\{x; \varphi(x) \leq y\}} dF(x).$$

□

### 1.3.2 Součet náhodných veličin

Mějme dvě nezávislé (viz kapitola 1.3.4) náhodné veličiny  $X$  a  $Y$  s distribučními funkcemi  $F(x)$  a  $G(y)$ . Zajímá nás rozdělení součtu  $Z = X + Y$ . Distribuční funkci náhodné veličiny  $Z$  označme  $H(z)$ . Pak obecně platí

$$\begin{aligned} H(z) &= \int \int_{x+y \leq z} dF(x) dG(y) = \int_{-\infty}^{\infty} F(z-y) dG(y) = \\ &= \int_{-\infty}^{\infty} G(z-x) dF(x). \end{aligned} \tag{1.24}$$

**Definice 1.16** *Rozdělení s distribuční funkcí  $H(z)$  se nazývá konvoluce rozdělení s distribučními funkcemi  $F(x)$  a  $G(y)$ .  $H$  se nazývá konvoluce distribučních funkcí  $F$  a  $G$ .*

Operaci konvoluce budeme značit  $H = F * G$ .

Mají-li náhodné veličiny  $X$  a  $Y$  diskrétní, resp. absolutně spojitě, rozdělení, pak s rozdělením jejich součtů pracujeme následovně:

**Věta 1.10** *Nechť  $F, G$  jsou diskrétní distribuční funkce se skoky v přirozených číslech o velikosti  $\{p_n\}, \{q_n\}$ , tj.*

$$F(x) = \sum_{0 \leq n \leq x} p_n, \quad G(y) = \sum_{0 \leq n \leq y} q_n.$$

*Nechť  $H = F * G$ . Potom  $H$  je diskrétní distribuční funkce se skoky v přirozených číslech a platí*

$$H(z) = \sum_{0 \leq n \leq z} h_n, \quad \text{kde } h_n = \sum_{k=0}^n p_k q_{n-k}.$$

**Věta 1.11** *Nechť náhodné veličiny  $X$  a  $Y$  jsou nezávislé a mají absolutně spojitě distribuční funkce  $F(x)$  a  $G(y)$  s hustotami  $f(x)$  a  $g(y)$ . Potom také  $H = F * G$  je absolutně spojitá a pro její hustotu  $h(z)$  (tj. pro hustotu náhodné veličiny  $Z = X + Y$ ) platí*

$$h(z) = \int_{-\infty}^{\infty} f(x)g(z-x)dx = \int_{-\infty}^{\infty} f(z-y)g(y)dy. \tag{1.25}$$

**Poznámka 1.10** *Funkce  $h(z)$  definovaná vztahem (1.25) se nazývá konvoluce hustot  $f(x)$  a  $g(y)$  a budeme ji značit  $h = f * g$ . Je to skutečně hustota, neboť z (1.25) plyne, že*



$h(z) \geq 0$  a

$$\begin{aligned}\int_{-\infty}^{\infty} h(z) dz &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x-y)g(y)dydx = \\ &= \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} f(x-y)dx \right) g(y)dy = \\ &= \int_{-\infty}^{\infty} 1 \cdot g(y)dy = 1.\end{aligned}$$

### Konvoluce některých základních pravděpodobnostních rozdělání:

#### 1. Konvoluce binomických rozdělání

Nechť  $X, Y$  jsou nezávislé náhodné veličiny  $X \sim Binom(n_1, p)$  a  $Y \sim Binom(n_2, p)$ .  
Potom náhodná veličina  $Z = X + Y$  má rozdělání  $Binom(n_1 + n_2, p)$ .

#### 2. Konvoluce Poissonových rozdělání

Nechť  $X \sim Po(\lambda_1)$  a  $Y \sim Po(\lambda_2)$  jsou nezávislé. Potom náhodná veličina  $Z = X + Y$  má rozdělání  $Po(\lambda_1 + \lambda_2)$ .

#### 3. Konvoluce rovnoměrných rozdělání

Nechť

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{pro } a \leq x \leq b \\ 0 & \text{jinak} \end{cases}$$

a

$$g(y) = \begin{cases} \frac{1}{d-c} & \text{pro } c \leq y \leq d \\ 0 & \text{jinak.} \end{cases}$$

Předpokládejme, že  $d - c \geq b - a$ . Pak pro konvoluci hustot  $h(z)$  náhodných veličin s hustotami  $f(x)$  a  $g(y)$  platí

$$h(z) = \begin{cases} 0 & \text{pro } z \leq a + c \text{ nebo } b + d \leq z \\ \frac{z-(a+c)}{(b-a)(d-c)} & \text{pro } a + c \leq z \leq b + c \\ \frac{1}{d-c} & \text{pro } b + c \leq z \leq a + d \\ \frac{(b+d)-z}{(b-a)(d-c)} & \text{pro } a + d \leq z \leq b + d \end{cases}$$

Grafem je lichoběžník se základnou ve vodorovné ose. Vidíme, že hustota  $h(z)$  je všude spojitá, ačkoliv  $f(x)$  a  $g(y)$  mají body nespojitosti (konvoluce "vyhlazuje" nespojitosti). Ve speciálním případě, kdy obě náhodné veličiny  $X$  a  $Y$  mají stejné rozdělání (tj.  $a = c, b = d$ ), má hustota  $h(z)$  tvar trojúhelníku. Toto rozdělání se nazývá *Simpsonovo rozdělání*.

#### 4. Konvoluce normálních rozdělení

Nechť  $X, Y$  jsou nezávislé náhodné veličiny,  $X \sim N(\mu_1, \sigma_1^2)$  a  $Y \sim N(\mu_2, \sigma_2^2)$ . Potom  $Z = X + Y$  má rozdělení  $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ .

#### 5. Konvoluce exponenciálních rozdělení

Jsou-li  $X, Y$  nezávislé náhodné veličiny s tímž exponenciálním rozdělením s parametrem  $\lambda > 0$ , pak náhodná veličina  $Z = X + Y$  má rozdělení s hustotou

$$h(z) = \begin{cases} \lambda^2 z \exp\{-z\lambda\} & z > 0, \\ 0 & z \leq 0. \end{cases}$$

### 1.3.3 Náhodný vektor

**Definice 1.17** *Nechť  $(\Omega, \mathcal{A}, P)$  je pravděpodobnostní prostor a nechť na tomto prostoru jsou definovány náhodné veličiny  $X_1, X_2, \dots, X_n$ . Pak vektor  $\mathbb{X} = (X_1, \dots, X_n)^T$  nazýváme náhodný vektor.*

**Poznámka 1.11** *Náhodný vektor je tedy zobrazení z  $\Omega$  do  $\mathbb{R}^n$ . Hodnoty náhodného vektoru je možno geometricky interpretovat jako bod v  $n$ -rozměrném prostoru.*

**Definice 1.18** *Nechť  $\mathbb{X} = (X_1, \dots, X_n)^T$  je náhodný vektor definovaný na pravděpodobnostním prostoru  $(\Omega, \mathcal{A}, P)$ . (Sdruženou) distribuční funkcí  $F_{\mathbb{X}}$  náhodného vektoru  $\mathbb{X}$  nazveme reálnou funkci  $n$  proměnných definovanou na  $\mathbb{R}^n$  vztahem*

$$\begin{aligned} F_{\mathbb{X}}(x_1, \dots, x_n) &= P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n) = \\ &= P(\cap_{i=1}^n \{\omega : X_i(\omega) \leq x_i\}), \quad -\infty < x_i < \infty, \quad i = 1, \dots, n. \end{aligned} \tag{1.26}$$

#### Vlastnosti distribuční funkce náhodného vektoru

1.  $F_{\mathbb{X}}(x_1, \dots, x_n)$  je neklesající funkce v každé ze svých proměnných při pevných hodnotách ostatních proměnných.
2.  $F_{\mathbb{X}}(x_1, \dots, x_n)$  je zprava spojitá v každé proměnné.
3.  $\lim_{x_i \rightarrow -\infty} F_{\mathbb{X}}(x_1, \dots, x_n) = 0$ ,  $i = 1, \dots, n$ , hodnoty  $x_j$  jsou pevné,  $j \neq i$ ,  $j = 1, \dots, n$ .

$$4. \lim_{x_1 \rightarrow \infty} F_{\mathbb{X}}(x_1, \dots, x_n) = 1.$$

$$\begin{array}{c} x_2 \rightarrow \infty \\ \vdots \\ x_n \rightarrow \infty \end{array}$$

**Definice 1.19** Náhodný vektor  $\mathbb{X}$  má diskrétní rozdělení, jestliže existuje posloupnost  $\{\mathbf{x}_k\}_{k=1}^{\infty}$ ,  $\mathbf{x}_k \in \mathbb{R}^n$ , a odpovídající posloupnost kladných čísel  $\{p_k\}_{k=1}^{\infty}$  taková, že

$$\sum_{k=1}^{\infty} p_k = 1, \quad \text{kde } p_k = P(\mathbb{X} = \mathbf{x}_k) = P(\{\omega \in \Omega : \mathbb{X}(\omega) = \mathbf{x}_k\}). \quad (1.27)$$

Distribuční funkce náhodného vektoru  $\mathbb{X}$  diskrétního typu má tvar

$$F_{\mathbb{X}}(\mathbf{x}) = \sum_{\{k: \mathbf{x}_k \leq \mathbf{x}\}} p_k, \quad \forall \mathbf{x} \in \mathbb{R}^n, \quad (1.28)$$

kde nerovnost  $\mathbf{x}_k < \mathbf{x}$  je uvažována po složkách, tj.  $x_k^i < x^i$  pro všechna  $i = 1, \dots, n$ .

**Definice 1.20** Náhodný vektor  $\mathbb{X} = (X_1, \dots, X_n)^T$  má absolutně spojitě rozdělení, jestliže existuje nezáporná funkce  $f_{\mathbb{X}}$   $n$  reálných proměnných taková, že

$$F_{\mathbb{X}}(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} f_{\mathbb{X}}(t_1, \dots, t_n) dt_1, \dots, dt_n, \quad (1.29)$$

kde funkci  $f_{\mathbb{X}}$  nazýváme hustotou rozdělení pravděpodobnosti náhodného vektoru  $\mathbb{X}$ , nebo též sdruženou hustotou náhodných veličin  $X_1, \dots, X_n$ .

**Poznámka 1.12** Stejně jako v případě náhodných veličin, i pro náhodný vektor lze jeho rozdělení zobecnit pomocí pravděpodobnostních měr. Pro naše účely ale úplně postačí základní dělení na náhodné vektory diskrétní a spojitě.

**Definice 1.21** Rozdělení (distribuční funkce, hustota, pravděpodobnosti) náhodného vektoru  $(X_1, \dots, X_k)^T$ , který je podvektorem náhodného vektoru  $\mathbb{X} = (X_1, \dots, X_n)^T$ , se nazývá marginální rozdělení (distribuční funkce, hustota, pravděpodobnosti).

**Způsob výpočtu marginálního rozdělení:**

1. Má-li náhodný vektor  $\mathbb{X} = (X_1, \dots, X_n)^T$  diskrétní rozdělení se sdruženými pravděpodobnostmi  $P(X_1 = \cdot, \dots, X_{i-1} = \cdot, X_i = \cdot, X_{i+1} = \cdot, \dots, X_n = \cdot)$ , kde náhodné veličiny  $X_i$  nabývají hodnot  $x_{i,1}, \dots, x_{i,k_i}$ , pak marginální pravděpodobnosti jsou

$$P(X_i = x) = \sum_{j_1=1}^{k_1} \dots \sum_{j_{i-1}=1}^{k_{i-1}} \sum_{j_{i+1}=1}^{k_{i+1}} \dots \sum_{j_n=1}^{k_n} P(X_1 = x_{1,j_1}, \dots, X_{i-1} = x_{i-1,j_{i-1}}, X_i = x, X_{i+1} = x_{i+1,j_{i+1}}, \dots, X_n = x_{n,j_n}).$$

2. Má-li náhodný vektor  $\mathbb{X} = (X_1, \dots, X_n)^T$  spojitě rozdělení se sdruženou hustotou  $f_{\mathbb{X}}$ , pak marginální hustota náhodné veličiny  $X_i$  je  $(n-1)$ -násobný integrál

$$f_{X_i}(x) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{\mathbb{X}}(x_1, \dots, x_{i-1}, x, x_{i+1}, \dots, x_n) dx_1, \dots, dx_{i-1} dx_{i+1}, \dots, dx_n.$$

### Nejčastěji používané číselné charakteristiky náhodného vektoru

$\mathbb{X} = (X_1, \dots, X_n)^T$ :

- a) Vektor středních hodnot

$$\mathbb{E}\mathbb{X} = (\mathbb{E}X_1, \dots, \mathbb{E}X_n)^T.$$

- b) Varianční matice  $\text{var } \mathbb{X}$  definována jako matice typu  $n \times n$  s prvky

$$\text{cov}(X_i, X_j) = \mathbb{E}(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j), \quad 1 \leq i, j \leq n.$$

- c) Korelační matice  $\text{corr}\mathbb{X}$  s prvky

$$\text{corr}(X_i, X_j) = \frac{\text{cov}(X_i, X_j)}{\sqrt{\text{var}X_i}\sqrt{\text{var}X_j}}, \quad 1 \leq i, j \leq n.$$

**Poznámka 1.13** *Korelace je vynormovaná kovariance v tom smyslu, že zatímco kovariance může nabývat libovolně velkých hodnot podle toho, jaké náhodné veličiny popisuje, pro korelaci platí*

$$-1 \leq \text{corr}(X, Y) \leq 1.$$

### 1.3.4 Nezávislost náhodných veličin

**Definice 1.22** *Náhodné veličiny  $X_1, X_2, \dots, X_n$  jsou vzájemně nezávislé, jestliže*

$$P(\cap_{j=1}^r \{\omega : X_{i_j}(\omega) < x_{i_j}\}) = \prod_{j=1}^r P(\{\omega : X_{i_j}(\omega) < x_{i_j}\}) \quad (1.30)$$

$$\forall \{i_1, i_2, \dots, i_r\} \subset \{1, 2, \dots, n\}, 1 \leq r \leq n, \forall x_{i_j} \in \mathbb{R}.$$

**Poznámka 1.14** Podobně jako u náhodných jevů můžeme zde definovat nezávislost náhodných veličin  $X_1, X_2, \dots, X_n$  po dvou. Definici nezávislosti po dvou bychom dostali z definice 1.22 pro  $r = 2$ .

**Věta 1.12 (Ověřování nezávislosti náhodných veličin v praxi)**

a) Nechť  $\mathbb{X} = (X_1, X_2, \dots, X_n)^T$  je náhodný vektor diskrétního typu. Náhodné veličiny  $X_1, X_2, \dots, X_n$  jsou vzájemně nezávislé právě tehdy, když platí

$$P(X_1 = x_1^{(i)}, \dots, X_n = x_n^{(i)}) = \prod_{j=1}^n P(X_j = x_j^{(i)})$$

pro všechna  $\mathbf{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)})$ ,  $i = 1, 2, \dots$ , kterých může  $\mathbb{X}$  nabývat.

b) Nechť  $\mathbb{X} = (X_1, X_2, \dots, X_n)^T$  je náhodný vektor absolutně spojitého typu. Náhodné veličiny  $X_1, X_2, \dots, X_n$  jsou vzájemně nezávislé právě tehdy, platí-li

$$f_{\mathbb{X}}(x_1, x_2, \dots, x_n) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdot \dots \cdot f_{X_n}(x_n), \quad \forall (x_1, x_2, \dots, x_n) \in \mathbb{R}^n.$$

**Věta 1.13** Jsou-li  $X, Y$  nezávislé náhodné veličiny s konečnými středními hodnotami, pak

a)  $\mathbb{E}XY = (\mathbb{E}X)(\mathbb{E}Y)$ .

b) Jsou-li navíc  $\mathbb{E}X^2 < \infty$  a  $\mathbb{E}Y^2 < \infty$ , pak  $\text{cov}(X, Y) = 0$ .

Platí-li  $\text{cov}(X, Y) = 0$ , pak říkáme, že náhodné veličiny jsou nekorelované. Z nekorelovanosti však obecně ještě neplyne nezávislost!

## 1.4 Zákony velkých čísel, centrální limitní věta

**Definice 1.23** Mějme posloupnost náhodných veličin  $X_1, X_2, X_3, \dots$  a náhodnou veličinu  $X$ . Nechť jsou všechny tyto veličiny definovány na témže pravděpodobnostním prostoru  $(\Omega, \mathcal{A}, P)$ .

Říkáme, že  $X_n$  konverguje k  $X$  skoro jistě, jestliže

$$P\{\omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\} = 1.$$

Jestliže pro každé  $\varepsilon > 0$  platí

$$\lim_{n \rightarrow \infty} P\{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\} = 0,$$

pak říkáme, že  $X_n$  konverguje k  $X$  v pravděpodobnosti.

**Věta 1.14** *Z konvergence skoro jistě plyne konvergence v pravděpodobnosti.*

Implikaci nelze bez dodatečných předpokladů obrátit!

### 1.4.1 Zákony velkých čísel

**Věta 1.15 (Slabý zákon velkých čísel)**

*Nechť  $\{X_n\}_{n=1}^{\infty}$  je posloupnost nezávislých náhodných veličin se stejnými středními hodnotami  $\mu$  a stejnými rozptyly  $\sigma^2 < \infty$ . Pak pro  $n \rightarrow \infty$  platí*

$$\frac{1}{n}(X_1 + X_2 + \dots + X_n) \rightarrow \mu$$

*v pravděpodobnosti.*

**Věta 1.16 (Silný zákon velkých čísel)**

*Nechť  $\{X_n\}_{n=1}^{\infty}$  je posloupnost nezávislých stejně rozdělených náhodných veličin s konečnou střední hodnotou  $\mathbb{E}X_1 = \mu$ . Pak pro  $n \rightarrow \infty$  platí*

$$\frac{1}{n}(X_1 + X_2 + \dots + X_n) \rightarrow \mu$$

*v pravděpodobnosti i skoro jistě.*

### 1.4.2 Centrální limitní věta

Podstatou centrální limitní věty (CLV) je tvrzení, že náhodná veličina  $X$ , která vznikla jako součet velkého počtu vzájemně nezávislých náhodných veličin  $X_1, X_2, \dots, X_n$ , má za velmi obecných podmínek přibližně normální rozdělení. Budeme říkat, že náhodná veličina  $X$  má tzv. asymptoticky normální rozdělení.

Centrální limitní věta má několik verzí s různými předpoklady. Asi nejčastěji používanou verzí je tzv. Lévy-Lindebergova CLV.

**Věta 1.17 (Lévy-Lindebergova CLV)**

*Nechť  $X_1, X_2, \dots$  jsou nezávislé stejně rozdělené náhodné veličiny se střední hodnotou  $\mu$  a konečným rozptylem  $\sigma^2$ . Označme*

$$Z_n = \frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{n\sigma^2}} \quad n = 1, 2, \dots$$

*a označme  $F_n(x)$  distribuční funkci  $Z_n$ . Potom  $\lim_{n \rightarrow \infty} F_n(x) = \Phi(x)$  pro všechna  $-\infty < x < \infty$ , kde  $\Phi(x)$  je distribuční funkce  $N(0, 1)$ .*

**Poznámka 1.15** *Předpoklad, že veličiny jsou nezávislé a stejně rozdělené, se v literatuře často označuje zkratkou i.i.d. (z anglického "independent, identically distributed").*

# Kapitola 2

## Základy statistiky

Úlohu matematické statistiky lze zjednodušeně popsat takto: Pro danou reálnou situaci (nasbíraná data) vytvoříme model (např. nějaké pravděpodobnostní rozdělení), z dat odhadneme jeho parametry a na závěr testujeme hypotézy o těchto parametrech. Zatímco pro první krok stačí srovnat graficky znázorněná data (např. histogram) s grafickým znázorněním teoretického rozdělení (např. tvar grafu hustoty), k dalším dvěma krokům je již zapotřebí statistických výpočetních metod, které lze zjednodušeně rozdělit do tří skupin, a to:

1. Bodové odhady parametrů
2. Intervalové odhady parametrů
3. Testování hypotéz

A právě těmto metodám bude věnována tato kapitola.

Předtím, než popíšeme výše zmíněné statistické metody, uvedeme si další pravděpodobnostní rozdělení, kterých se využívá - na základě věty 2.1 - zejména při tvorbě intervalových odhadů a při testování hypotéz (viz níže). Hodnoty jejich distribučních funkcí jsou stejně jako hodnoty distribuční funkce normovaného normálního rozdělení uvedeny ve statistických tabulkách.

$\chi_n^2$ -rozdělení



Nechť  $X_1, X_2, \dots, X_n$  jsou nezávislé stejně rozdělené náhodné veličiny s rozděleními  $N(0, 1)$ .

Pak náhodná veličina

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

má tzv.  $\chi_n^2$ -rozdělení (čteno "chí-kvadrát rozdělení o  $n$  stupních volnosti).

## Studentovo $t$ -rozdělení

Nechť  $X$  je náhodná veličina s rozdělením  $N(0, 1)$  a  $Y$  je náhodná veličina s rozdělením  $\chi_n^2$ . Pak náhodná veličina

$$Z = \frac{X}{\sqrt{Y}} \sqrt{n}$$

má tzv.  $t_n$ -rozdělení nazváno také Studentovo  $t$ -rozdělení o  $n$  stupních volnosti.

**Definice 2.1** Náhodnému vektoru  $\mathbb{X} = (X_1, X_2, \dots, X_n)^T$  nezávislých stejně rozdělených náhodných veličin s distribuční funkcí  $F_\theta$  závisející na parametru  $\theta$  ve statistice říkáme náhodný výběr. Číslo  $n$  říkáme rozsah výběru.

**Definice 2.2** Funkce náhodného výběru

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

se nazývá výběrový průměr a funkce

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

se nazývá výběrový rozptyl.  $S_n = \sqrt{S_n^2}$  je pak výběrová směrodatná odchylka.

**Věta 2.1** Necht'  $\mathbb{X} = (X_1, X_2, \dots, X_n)^T$  je náhodný výběr z rozdělení  $N(\mu, \sigma^2)$ ,  $\mu \in \mathbb{R}$ ,  $\sigma^2 > 0$ . Pak

1. výběrový průměr  $\bar{X}_n$  a výběrový rozptyl  $S_n^2$  jsou nezávislé náhodné veličiny,
2. rozdělení výběrového průměru  $\bar{X}_n$  je  $N(\mu, \sigma^2/n)$ ,
3. náhodná veličina  $(n-1)S_n^2/\sigma^2$  má  $\chi^2$ -rozdělení o  $(n-1)$  stupních volnosti,
4. náhodná veličina  $T = \frac{\bar{X}_n - \mu}{S_n} \sqrt{n}$  má  $t$ -rozdělení o  $(n-1)$  stupních volnosti.

**Definice 2.3** Necht' distribuční funkce  $F$  je spojitá, ryze monotónní a necht'  $0 < \beta < 1$ . Pak číslo  $z_\beta$  takové, že  $F(z_\beta) = \beta$  se nazývá  $\beta$ -kvantil tohoto rozdělení.

## Označení kvantilů používaných v této přednášce

Kvantily v této přednášce budeme často označovat v souladu s označením rozdělení příslušné náhodné veličiny. Používat budeme

- $u_\beta \dots$   $\beta$ -kvantil normovaného normálního rozdělení,
- $t_{\beta,n} \dots$   $\beta$ -kvantil Studentova  $t$ -rozdělení o  $n$  stupních volnosti,
- $\chi_{\beta,n}^2 \dots$   $\beta$ -kvantil  $\chi^2$  rozdělení o  $n$  stupních volnosti.

**Poznámka 2.1** Je-li  $X$  náhodná veličina s distribuční funkcí  $F$  a kvantily  $z_\beta$ , pak

$$P(z_{\alpha/2} < X < z_{1-\alpha/2}) = F(z_{1-\alpha/2}) - F(z_{\alpha/2}) = 1 - \alpha.$$

**Definice 2.4** Nechť  $(x_1, x_2, \dots, x_n)^T$  je realizace náhodného výběru  $\mathbb{X} = (X_1, X_2, \dots, X_n)^T$   
Pak

$$F_{emp}(x) = \frac{\#\{x_i : x_i \leq x\}}{n},$$

kde  $\#$  značí počet prvků, se nazývá empirická distribuční funkce.

**Definice 2.5** Nechť  $(x_1, x_2, \dots, x_n)^T$  je realizace náhodného výběru  $\mathbb{X} = (X_1, X_2, \dots, X_n)^T$   
Pak

- $z = \min(x_i : F_{emp}(x_i) \geq 1/4)$  se nazývá 1.kvartil,
- $z = \min(x_i : F_{emp}(x_i) \geq 3/4)$  se nazývá 3.kvartil,
- $z = \min(x_i : F_{emp}(x_i) \geq 1/2)$  se nazývá medián (2.kvartil),
- nejčastěji zastoupený prvek se nazývá modus.

**Poznámka 2.2** Grafickým znázorněním kvartilů je tzv. krabicový graf (boxplot) - viz obrázek 2.2 vpravo.

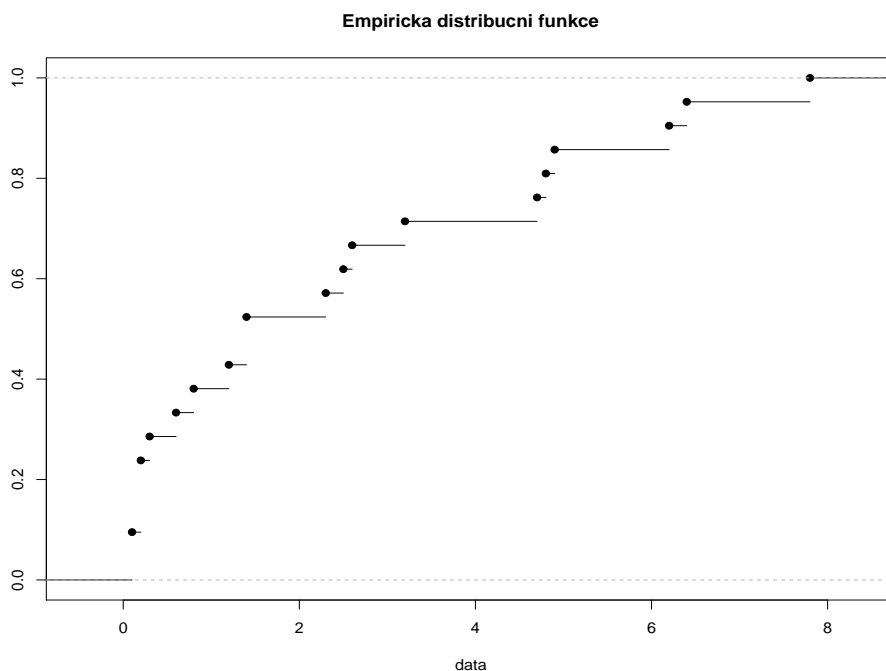
Dalším často používaným grafickým znázorněním dat pak je tzv. histogram znázorňující počet hodnot spadající do (vhodně zvolených) ekvidistančních intervalů - viz obrázek 2.2 vlevo.

**Příklad 2.1** 21× po sobě byly sledovány doby (v měsících) do poruchy daného přístroje. Naměřeny byly hodnoty:

4.9, 6.2, 2.6, 0.6, 0.3, 2.3, 3.2, 1.4, 6.4, 4.8, 1.2

2.5, 0.2, 0.2, 0.8, 0.1, 0.1, 1.4, 7.8, 0.2, 4.7.

Uspořádáme-li pro lepší přehled tato data od nejmenší hodnoty k největší, dostaneme řadu hodnot:



Obrázek 2.1: Distribuční funkce pro data z příkladu 2.1.

0.1, 0.1, 0.2, 0.2, 0.2, 0.3, 0.6, 0.8, 1.2, 1.4, 1.4,  
2.3, 2.5, 2.6, 3.2, 4.7, 4.8, 4.9, 6.2, 6.4, 7.8.

Graficky znázorněná empirická distribuční funkce je na obrázku 2.1.

Z dat pak máme:

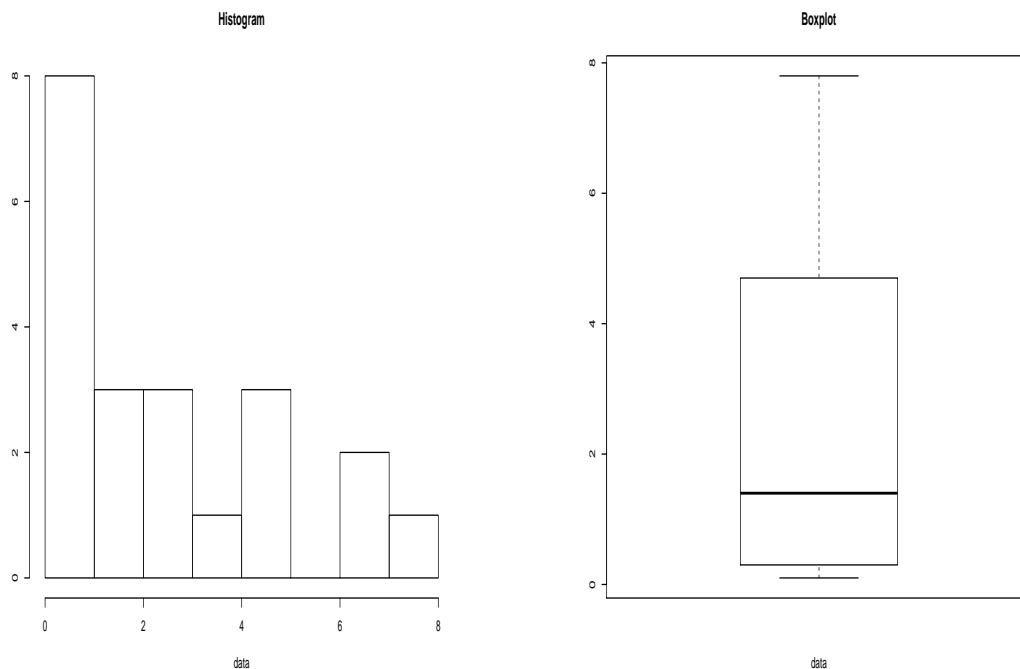
výběrový průměr  $\bar{X}_{21} = 2.471$ ,

výběrový rozptyl  $S_{21}^2 = 5.81$  a výběrovou směrodatnou odchylku  $S_{21} = \sqrt{5.81} = 2.21$ ,

1.kvartil = 0.3, medián (tj. 2.kvartil) = 1.4, 3.kvartil = 4.7 a modus = 0.2.

## 2.1 Bodový odhad

**Definice 2.6** Mějme náhodný výběr  $\mathbb{X} = (X_1, X_2, \dots, X_n)^T$ , kde rozdělení náhodné veličiny  $X_1$  závisí na parametru  $\theta$ . Bodový odhad parametru  $\theta$  je jakákoliv funkce náhodného výběru  $\theta^*(\mathbb{X})$ , jejíž funkční předpis nezávisí na  $\theta$ . Tento odhad se nazývá nestranný, jestliže  $\mathbb{E}\theta^*(\mathbb{X}) = \theta$ .



Obrázek 2.2: Histogram a boxplot pro data z příkladu 2.1.

Výše zmíněná definice je sice matematicky přesná, avšak pro naše potřeby příliš obecná a nepřehledná. Budeme si proto pod bodovým odhadem představovat nikoliv funkci náhodného výběru  $\theta^*(\mathbb{X})$ , nýbrž samotné číslo  $\hat{\theta}$ , které získáme z realizace  $(x_1, x_2, \dots, x_n)^T$  náhodného výběru  $\mathbb{X} = (X_1, X_2, \dots, X_n)^T$  a které co nejpřesněji popisuje hledaný parametr  $\theta$ .

### 2.1.1 Metoda momentů

Mějme  $(x_1, x_2, \dots, x_n)^T$  realizaci náhodného výběru  $\mathbb{X} = (X_1, X_2, \dots, X_n)^T$ , kde s distribuční funkce  $F$  náhodné veličiny  $X_1$  závisí na parametrech  $\theta_1, \dots, \theta_k \in \Theta$ , kde  $\Theta$  je množina, z níž může parametr pocházet (např. nezáporná reálná čísla). Předpokládejme, že tzv.  $i$ -té momenty  $\mathbb{E}X_1^i$  jsou konečné pro všechna  $i = 1, \dots, k$ . Tyto momenty rovněž závisí na  $\theta_1, \dots, \theta_k$ . Pak položením

$$\mathbb{E}X_1^i = m_i,$$

kde  $m_i$  je  $i$ -tý výběrový moment získaný jako

$$m_i = \frac{1}{n} \sum_{j=1}^n x_j^i$$

pro všechna  $i = 1, \dots, k$ , získáme soustavu  $k$  rovnic o  $k$  neznámých  $\theta_1, \dots, \theta_k$ , jejímž řešením jsou odhady  $\hat{\theta}_1, \dots, \hat{\theta}_k$ .

Alternativa: Pokud  $k = 2$ , pak místo vztahů pro  $i$ -té momenty,  $i = 1, 2$ , můžeme uvažovat rovnice  $\mathbb{E}X_1 = \bar{X}_n$  a  $\text{var } X_1 = S_n^2$ .

Nevýhoda: tento odhad má velký rozptyl.

## 2.1.2 Metoda maximální věrohodnosti

Předpokládejme, že máme  $(x_1, x_2, \dots, x_n)^T$  realizaci náhodného výběru  $\mathbb{X} = (X_1, X_2, \dots, X_n)^T$  z rozdělení s pravděpodobnostmi  $P_\theta(X_1 = \cdot)$  nebo s hustotou  $f_\theta$  a nechť tyto pravděpodobnosti, resp. hustota, závisí na nějakém parametru  $\theta \in \Theta$ . Odhad  $\hat{\theta}$  je maximálně věrohodným odhadem, jestliže

$$\prod_{i=1}^n P_{\hat{\theta}}(X_1 = x_i) = \max_{\theta \in \Theta} \prod_{i=1}^n P_\theta(X_1 = x_i),$$

resp.

$$\prod_{i=1}^n f_{\hat{\theta}}(x_i) = \max_{\theta \in \Theta} \prod_{i=1}^n f_\theta(x_i).$$

Většinou je však výhodnější pracovat s logaritmem součinu, neboť tak získáme součet logaritmů. Postup je tudíž následující:

a) Diskrétní případ:

1. Vyjádřit věrohodnostní funkci  $L(\theta) = \prod_{i=1}^n P_\theta(X_1 = x_i)$ .
2. Zlogaritmovat a tím získat logaritmicko-věrohodnostní funkci  $l(\theta) = \sum_{i=1}^n \log P_\theta(X_1 = x_i)$ .
3. Položit  $\frac{\partial l(\theta)}{\partial \theta} = 0$ .
4. Řešení rovnice  $\frac{\partial l(\theta)}{\partial \theta} = 0$  je hledaný maximálně věrohodný odhad  $\hat{\theta}$ .

b) Spojitý případ:

1. Vyjádřit věrohodnostní funkci  $L(\theta) = \prod_{i=1}^n f_\theta(x_i)$ .

2. Zlogaritmovat a tím získat logaritmicko-věrohodnostní funkci

$$l(\theta) = \sum_{i=1}^n \log f_{\theta}(x_i).$$

3. Položit  $\frac{\partial l(\theta)}{\partial \theta} = 0$ .

4. Řešení rovnice  $\frac{\partial l(\theta)}{\partial \theta} = 0$  je hledaný maximálně věrohodný odhad  $\hat{\theta}$ .

## 2.2 Intervalový odhad

**Definice 2.7** Mějme náhodný výběr  $\mathbb{X} = (X_1, X_2, \dots, X_n)^T$  a číslo  $\alpha \in (0, 1)$ .

1. Dvojice  $(\theta_L^*(X_1, \dots, X_n), \theta_U^*(X_1, \dots, X_n))$  se nazývá intervalový odhad parametru  $\theta$  o spolehlivosti  $1 - \alpha$ , jestliže

$$P(\theta_L^*(X_1, \dots, X_n) < \theta < \theta_U^*(X_1, \dots, X_n)) = 1 - \alpha.$$

2.  $(\theta_D^*(X_1, \dots, X_n))$  se nazývá dolní odhad parametru  $\theta$  o spolehlivosti  $1 - \alpha$ , jestliže

$$P(\theta_D^*(X_1, \dots, X_n) < \theta) = 1 - \alpha.$$

3.  $(\theta_H^*(X_1, \dots, X_n))$  se nazývá horní odhad parametru  $\theta$  o spolehlivosti  $1 - \alpha$ , jestliže

$$P(\theta_H^*(X_1, \dots, X_n) > \theta) = 1 - \alpha.$$

Konstrukce intervalového odhadu závisí na typu odhadovaného parametru a samozřejmě také na rozdělení, z něhož náhodný výběr pochází. Existuje sice obecná metoda pro konstrukci intervalového odhadu. Ta je však pro naše účely zbytečně komplikovaná, a proto se raději zaměříme jen na odhady nejčastěji používané v praxi.

### 2.2.1 Intervalové odhady parametrů normálního rozdělení

**Věta 2.2** Necht'  $\mathbb{X} = (X_1, X_2, \dots, X_n)^T$  je náhodný výběr z rozdělení  $N(\mu, \sigma^2)$ ,  $\mu \in \mathbb{R}$  je neznámý parametr,  $\sigma^2 > 0$  známá konstanta. Pak

1.  $(\bar{X}_n - u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X}_n + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}})$  je intervalový odhad  $\mu$  o spolehlivosti  $1 - \alpha$ ,

2.  $\bar{X}_n - u_{1-\alpha} \frac{\sigma}{\sqrt{n}}$  je dolní intervalový odhad  $\mu$  o spolehlivosti  $1 - \alpha$ ,

3.  $\bar{X}_n + u_{1-\alpha} \frac{\sigma}{\sqrt{n}}$  je horní intervalový odhad  $\mu$  o spolehlivosti  $1 - \alpha$ .

**Věta 2.3** Necht'  $\mathbb{X} = (X_1, X_2, \dots, X_n)^T$  je náhodný výběr z rozdělení  $N(\mu, \sigma^2)$ ,  $\mu \in \mathbb{R}$ ,  $\sigma^2 > 0$ , oba parametry neznámé. Pak

1.  $(\bar{X}_n - t_{1-\alpha/2, n-1} \frac{S_n}{\sqrt{n}}, \bar{X}_n + t_{1-\alpha/2, n-1} \frac{S_n}{\sqrt{n}})$  je intervalový odhad  $\mu$  o spolehlivosti  $1 - \alpha$ ,
2.  $\bar{X}_n - t_{1-\alpha, n-1} \frac{S_n}{\sqrt{n}}$  je dolní intervalový odhad  $\mu$  o spolehlivosti  $1 - \alpha$ ,
3.  $\bar{X}_n + t_{1-\alpha, n-1} \frac{S_n}{\sqrt{n}}$  je horní intervalový odhad  $\mu$  o spolehlivosti  $1 - \alpha$ ,
4.  $(\frac{(n-1)S_n^2}{\chi_{1-\alpha/2, n-1}^2}, \frac{(n-1)S_n^2}{\chi_{\alpha/2, n-1}^2})$  je intervalový odhad  $\sigma^2$  o spolehlivosti  $1 - \alpha$ ,
5.  $\frac{(n-1)S_n^2}{\chi_{1-\alpha, n-1}^2}$  je dolní intervalový odhad  $\sigma^2$  o spolehlivosti  $1 - \alpha$ ,
6.  $\frac{(n-1)S_n^2}{\chi_{\alpha, n-1}^2}$  je horní intervalový odhad  $\sigma^2$  o spolehlivosti  $1 - \alpha$ .

## 2.2.2 Intervalové odhady založené na CLV

**Věta 2.4** Necht'  $\mathbb{X} = (X_1, X_2, \dots, X_n)^T$  je náhodný výběr z libovolného rozdělení, pro které  $0 < \sigma^2 < \infty$ . Pak intervalovým odhadem  $\mu = \mathbb{E}X$  o asymptotické spolehlivosti  $1 - \alpha$  je

$$(\bar{X}_n - u_{1-\alpha/2} \frac{S_n}{\sqrt{n}}, \bar{X}_n + u_{1-\alpha/2} \frac{S_n}{\sqrt{n}}).$$

*Důkaz.* Víme, že pro velká  $n$  platí  $\frac{S_n}{\sigma} \rightarrow 1$ , tj.  $S_n$  se asymptoticky blíží hodnotě  $\sigma$ . Z CLV víme, že  $\frac{\sum X_i - n\mu}{\sqrt{n\sigma^2}}$  má přibližně normální rozdělení. To znamená, že

$$\begin{aligned} P(u_{\frac{\alpha}{2}} \leq \frac{\sum X_i - n\mu}{\sqrt{n\sigma^2}} \leq u_{1-\frac{\alpha}{2}}) &= 1 - \alpha \\ P(u_{\frac{\alpha}{2}} \leq \frac{\sum X_i - n\mu}{\sqrt{n}S_n} \leq u_{1-\frac{\alpha}{2}}) &= 1 - \alpha \\ P(\frac{\sum X_i}{n} + u_{1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}} \geq \mu \geq \frac{\sum X_i}{n} + u_{\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}}) &= 1 - \alpha \\ P(\bar{X}_n + u_{1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}} \geq \mu \geq \bar{X}_n - u_{1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}}) &= 1 - \alpha, \end{aligned}$$

což je definice intervalového odhadu o spolehlivosti  $1 - \alpha$ .

Speciálními případy Věty 2.4 jsou následující dvě tvrzení:



**Věta 2.5** 1. Necht  $\mathbb{X} = (X_1, X_2, \dots, X_n)^T$  je náhodný výběr z alternativního rozdělení s parametrem  $0 < p < 1$ . Pak intervalovým odhadem  $p$  o asymptotické spolehlivosti  $1 - \alpha$  je

$$\left( \bar{X}_n - u_{1-\alpha/2} \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + u_{1-\alpha/2} \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \right).$$

2. Necht  $\mathbb{X} = (X_1, X_2, \dots, X_n)^T$  je náhodný výběr z Poissonova rozdělení s parametrem  $0 < \lambda < \infty$ . Pak intervalovým odhadem  $\lambda$  o asymptotické spolehlivosti  $1 - \alpha$  je

$$\left( \bar{X}_n - u_{1-\alpha/2} \sqrt{\frac{\bar{X}_n}{n}}, \bar{X}_n + u_{1-\alpha/2} \sqrt{\frac{\bar{X}_n}{n}} \right).$$

*Důkaz.* Myšlenka důkazu plyne z charakteristik daných rozdělení: pro alternativní rozdělení máme  $\mathbb{E}X = p$  a  $\text{var } X = p(1-p)$  a v Poissonově rozdělení platí  $\mathbb{E}X = \text{var } X = \lambda$ .

□

## 2.3 Testování hypotéz

Matematická statistika popisuje obrovské množství testů o hodnotách získaných odhadů, shodnosti parametrů různých rozdělení, nezávislosti apod. V této přednášce si pouze popíšeme obecný princip testování hypotéz a jen pro ukázkou jejich aplikace zde uvedeme ty nejčastěji používané testy.

### 2.3.1 Princip testování hypotéz

Necht  $\mathbb{X} = (X_1, X_2, \dots, X_n)^T$  je náhodný výběr z rozdělení, které závisí na parametru  $\theta \in \Theta$ . Chceme-li zjistit, jestli  $\theta$  patří do nějaké užší podmnožiny parametrů  $\Theta_0$ , nazveme tuto domněnku nulovou hypotézou, označme  $H_0 : \theta \in \Theta_0$ , a na základě výběru  $\mathbb{X} = (X_1, X_2, \dots, X_n)^T$  testujeme tuto hypotézu proti alternativě  $H_A : \theta \in \Theta \setminus \Theta_0$ . To se zpravidla provede tak, že se určí množina  $W$  taková, že pokud  $\mathbb{X} \in W$ , pak  $H_0$  zamítáme.

**Poznámka 2.3** Většinou testujeme  $H_0 : \theta = \theta_0$ , kde  $\theta_0$  je nějaká konkrétní hodnota. Přírozenou alternativou je pak  $H_A : \theta \neq \theta_0$ . Někdy je však smysluplnější testovat např. proti alternativě  $H_A : \theta > \theta_0$ . I když totiž teoreticky může být  $\theta < \theta_0$ , nemá smysl se tou situací zabývat, např. pokud při testování střední hodnoty  $H_0 : \mu = 3$  z dat víme, že  $\bar{X}_n = 5$ , pak nemá smysl uvažovat situaci  $\mu < 3$ .

Přítom mohou nastat následující situace:

- $H_0$  ve skutečnosti platí a test ji nezamítá. ✓
- $H_0$  ve skutečnosti neplatí a test ji zamítá. ✓
- $H_0$  ve skutečnosti platí, ale test ji zamítá  $\rightarrow$  nastala chyba 1.druhu.
- $H_0$  ve skutečnosti neplatí, ale test ji nezamítá  $\rightarrow$  nastala chyba 2.druhu.

Zvolíme si tzv. hladinu testu  $\alpha$  (obvykle 0.05, někdy ale také 0.01 i méně) a  $W$  volíme tak, aby pravděpodobnost chyby 1.druhu nebyla větší než  $\alpha$  (obvykle tak, aby byla rovna  $\alpha$ ).

**Definice 2.8** *Nejmenší hladina  $\alpha$ , pro kterou test hypotézu  $H_0$  ve prospěch  $H_A$  zamítá, se nazývá  $p$ -hodnota.*

### 2.3.2 Testování pomocí intervalových odhadů

Intervalové odhady uvedené v kapitole 2.2 lze využít pro testování hypotéz o střední hodnotě rozdělení, z něhož náhodný výběr  $\mathbb{X} = (X_1, X_2, \dots, X_n)^T$  pochází. Uvažujme hladinu testu  $\alpha$  a předpokládejme, že intervalový odhad pro střední hodnotu  $\mu$  o spolehlivosti  $1 - \alpha$  vypočítaný z výběru  $\mathbb{X}$  je  $(\mu_L, \mu_U)$ . Pak hypotézu  $H_0 : \mu = \mu_0$ , kde  $\mu_0$  je libovolná konstanta, zamítáme ve prospěch hypotézy  $H_A : \mu \neq \mu_0$ , jestliže  $\mu_0 \notin (\mu_L, \mu_U)$ .

Podobně pak můžeme testovat také  $H_0 : \mu = \mu_0$  oproti  $H_A : \mu > \mu_0$ . V tomto případě  $H_0$  zamítáme ve prospěch hypotézy  $H_A$ , jestliže  $\mu_0 \notin (-\infty, \mu_H)$ , kde  $\mu_H$  je horní odhad střední hodnoty  $\mu$ . Analogicky se pak postupuje i při testování  $H_0 : \mu = \mu_0$  oproti  $H_A : \mu < \mu_0$ .

### 2.3.3 Testování střední hodnoty normálního rozdělení - $t$ -testy

Obvykle se však ve statistice používá postup "opačný", tj. místo hledání intervalů pro testovaný parametr se z dat a testované hodnoty vypočítá hodnota, která je realizací nějaké náhodné veličiny, a ta se poté porovná s příslušným kvantilem.

Tímto postupem lze samozřejmě testovat také hypotézy o střední hodnotě normálního rozdělení. Jelikož k výpočtu využijeme větu 2.1 o  $t$ -rozdělení transformace výběrového průměru, nazýváme tyto testy tzv.  $t$ -testy.

## Jednovýběrový $t$ -test

Nechť  $(X_1, X_2, \dots, X_n)^T$  je náhodný výběr z  $N(\mu, \sigma^2)$ , kde  $\sigma^2 > 0$ , a předpokládejme, že ani jeden parametr není znám. Testujeme-li  $H_0 : \mu = \mu_0$  oproti  $H_A : \mu \neq \mu_0$ , je postup následující:

1. Vypočteme hodnotu  $T_0 = \frac{\bar{X}_n - \mu_0}{S_n} \sqrt{n}$ .
2. Pokud  $|T_0| \geq t_{1-\alpha/2, n-1}$ , zamítáme  $H_0$ . V opačném případě pak  $H_0$  nezamítáme.

Jednostranný test  $H_0 : \mu = \mu_0$  oproti  $H_A : \mu > \mu_0$  pak probíhá podobně:

1. Vypočteme hodnotu  $T_0 = \frac{\bar{X}_n - \mu_0}{S_n} \sqrt{n}$ .
2. Pokud  $T_0 \geq t_{1-\alpha, n-1}$ , zamítáme  $H_0$ . V opačném případě pak  $H_0$  nezamítáme.

$H_0 : \mu = \mu_0$  oproti  $H_A : \mu < \mu_0$  pak zamítáme v případě, že  $T_0 \leq t_{\alpha, n-1}$ .

## Párový $t$ -test

Ve statistice se často vyskytuje situace, kdy u jednoho objektu sledujeme dva spolu související znaky najednou (např. výnosy dvou různých poboček jedné firmy apod.). Máme tedy náhodný výběr  $(Y_1, Z_1), (Y_2, Z_2), \dots, (Y_n, Z_n)^T$  a chceme testovat  $H_0 : \mathbb{E}Y_i - \mathbb{E}Z_i = \mu_0$  (nejčastěji  $\mu_0 = 0$ , tj. střední hodnoty se rovnají) oproti některé z výše uvedených alternativ. Pak postupujeme tak, že utvoříme rozdíly

$$X_1 = Y_1 - Z_1, \dots, X_n = Y_n - Z_n$$

a pokud veličiny  $X_1, \dots, X_n$  pocházejí z normálního rozdělení, můžeme na ně použít výše popsáný jednovýběrový  $t$ -test.

## Dvouvýběrový $t$ -test

Mějme dva náhodné výběry  $(X_1, X_2, \dots, X_m)^T$  z  $N(\mu_1, \sigma^2)$  a  $(Y_1, Y_2, \dots, Y_n)^T$  z  $N(\mu_2, \sigma^2)$ , kde  $\sigma^2 > 0$ , a předpokládejme, že tyto výběry na sobě nezávisí.

Označme  $\bar{X}$  výběrový průměr výběru  $(X_1, X_2, \dots, X_m)^T$ ,  $\bar{Y}$  výběrový průměr výběru  $(Y_1, Y_2, \dots, Y_n)^T$ ,  $S_X^2$  výběrový rozptyl výběru  $(X_1, X_2, \dots, X_m)^T$  a  $S_Y^2$  výběrový rozptyl výběru  $(Y_1, Y_2, \dots, Y_n)^T$ .

Platí, že náhodná veličina

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{(m-1)S_X^2 + (n-1)S_Y^2}} \sqrt{\frac{mn(m+n-2)}{m+n}}$$

má  $t_{m+n-2}$  rozdělení.

Chceme-li tudíž testovat  $H_0 : \mu_1 - \mu_2 = \mu_0$  oproti  $H_A : \mu_1 - \mu_2 \neq \mu_0$ , je postup následující:

1. Vypočteme hodnotu  $T_0 = \frac{\bar{X} - \bar{Y} - \mu_0}{\sqrt{(m-1)S_X^2 + (n-1)S_Y^2}} \sqrt{\frac{mn(m+n-2)}{m+n}}$ .
2. Pokud  $|T_0| \geq t_{1-\alpha/2, m+n-2}$ , zamítáme  $H_0$ . V opačném případě pak  $H_0$  nezamítáme.

## Testování normality

Základním předpokladem pro  $t$ -testy je normalita dat. Tu lze testovat

- analytickými testy (např. Shapiro-Wilkův test, D'Agostinův test apod. - v této přednášce vynecháme)
- graficky (např. histogram, Q-Q plot apod.)

Q-Q plot je graf, v němž jsou porovnávány teoretické kvantily testovaného - tj. v našem případě normálního - rozdělení (vodorovná osa) s empirickými kvantily získanými z dat (svislá osa). Data pak považujeme za výběr z normálního rozdělení, je-li graf přibližně lineární.

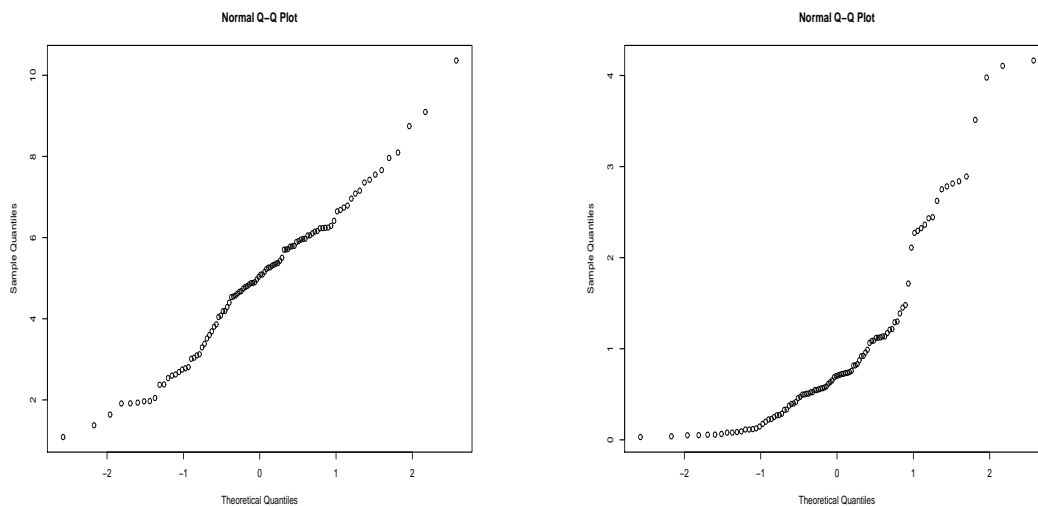
**Poznámka 2.4** *Hodnoty parametrů testovaného rozdělení se přitom mohou lišit od parametrů dat, důležitý je typ rozdělení.*

Ukázka Q-Q plotu je na obrázku 2.3.

### 2.3.4 $\chi^2$ test dobré shody

Nejprve si uveďme zobecnění binomického rozdělení pro více možných výsledků jednoho pokusu než jen úspěch nebo neúspěch.

## Multinomické rozdělení



Obrázek 2.3: Q-Q plot pro data pocházející z rozdělení  $N(5, 4)$  (vlevo) a  $Exp(1)$  (vpravo).

Předpokládejme, že v určitém pokusu může nastat právě jeden z jevů  $A_1, A_2, \dots, A_k$  a označme pravděpodobnosti  $p_i = P(A_i)$ . Opakujme tento pokus  $n$ -krát a označme  $X_i$  počet výskytu jevu  $A_i$  v těchto  $n$  pokusech. Pak

$$P(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k}, \quad \sum_{i=1}^k p_i = 1, \quad \sum_{i=1}^k x_i = n$$

a rozdělení vektoru  $(X_1, X_2, \dots, X_k)^T$  se nazývá multinomické.

Chceme-li testovat  $H_0$ , že skutečné hodnoty dílčích pravděpodobností jsou rovny předem zadaným číslům  $p_1, \dots, p_k$ , oproti alternativě, že aspoň jedna hodnota  $p_i$  je jiná, postupujeme následovně:

1. Vypočteme hodnotu  $\chi^2 = \sum_{i=1}^k \frac{(X_i - np_i)^2}{np_i}$ .
2. Pokud  $\chi^2 > \chi_{1-\alpha, k-1}^2$ , zamítáme  $H_0$ . V opačném případě pak  $H_0$  nezamítáme.

### 2.3.5 Test nezávislosti v kontingenční tabulce

Tento test na rozdíl od předešlých netestuje hodnotu parametru, nýbrž vyhodnocuje, zda je možné dva výběry z veličin s diskrétním rozdělením považovat za nezávislé.

Předpokládejme, že máme výběr  $(Y_1, Z_1), (Y_2, Z_2), \dots, (Y_n, Z_n)$ , kde  $Y_k$  může nabývat hodnot  $1, \dots, r$  a  $Z_k$  může nabývat hodnot  $1, \dots, c$ . Označíme-li  $n_{ij}$  počet výskytů dvojice

$(Y_k = i, Z_k = j)$ , pak matici o rozměru  $r \times c$  s prvky  $n_{ij}$  říkáme kontingenční tabulka a prvkům  $n_{ij}$  říkáme sdružené četnosti. Marginální četnosti jsou pak dány vztahy

$$n_{i.} = \sum_j n_{ij}, \quad n_{.j} = \sum_i n_{ij}.$$

K testování  $H_0$ : "Y a Z jsou nezávislé" oproti  $H_A$ : "Y a Z nejsou nezávislé" vypočteme hodnotu

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}.$$

Pokud  $\chi^2 \geq \chi_{1-\alpha, (r-1)(c-1)}^2$ , pak zamítáme  $H_0$  ve prospěch  $H_A$ .

# Kapitola 3

## Regresní analýza

Uvažujme náhodný vektor  $\mathbb{X} = (X_1, \dots, X_r)$ , náhodnou veličinu  $Y$  a předpokládejme, že  $Y$  na  $\mathbb{X}$  nějakým způsobem závisí. Úkolem regresní analýzy je nalézt funkční závislost  $Y$  na  $\mathbb{X}$ , tj. najít takovou funkci  $f$ , že

$$Y = f_{\theta_1, \dots, \theta_p}(X_1, \dots, X_r) + \epsilon, \quad (3.1)$$

kde tzv. chybový člen  $\epsilon$  je náhodná veličina s nulovou střední hodnotou a rozptylem  $\sigma^2$ , přičemž aby tato funkce dobře popisovala situaci, je třeba, aby tento rozptyl nebyl příliš velký. Často navíc předpokládáme normalitu  $\epsilon$ .

**Definice 3.1** *Vztah (3.1) se nazývá regresní model. Funkce  $f$  se nazývá regresní funkce. Číslům  $\theta_1, \dots, \theta_p$  se říká parametry regrese. Náhodný vektor  $\mathbb{X} = (X_1, \dots, X_r)$  se nazývá vysvětlující proměnná. Náhodná veličina  $Y$  je pak vysvětlovaná proměnná.*

### 3.1 Metoda nejmenších čtverců

Metoda nejmenších čtverců slouží k odhadu parametrů  $\theta_1, \dots, \theta_p$  z dat.

Nechť  $(y_i, x_{1_i}, \dots, x_{r_i}), i = 1, \dots, n$  je  $n$  pozorování vektoru  $(Y, X_1, \dots, X_r)$ . Metoda spočívá v minimalizaci

$$\begin{aligned} (\theta_1, \dots, \theta_p) &= \arg \min_{\theta_1, \dots, \theta_p} S(\theta_1, \dots, \theta_p) \\ &= \arg \min_{\theta_1, \dots, \theta_p} \sum_{i=1}^n (y_i - f(x_{1_i}, \dots, x_{r_i}, \theta_1, \dots, \theta_p))^2, \end{aligned}$$

příčemž obdobně jako v případě použití maximální věrohodnosti se tato minimalizace provádí pomocí parciálních derivací podle jednotlivých parametrů, kde tyto derivace položíme rovné 0 a vyřešíme soustavu  $p$  rovnic o  $p$  neznámých (viz následující sekce).

**Definice 3.2** Výraz  $S(\theta_1, \dots, \theta_p)$  se nazývá reziduální součet čtverců.

## 3.2 Lineární regrese

Nejjednodušší formou regrese je případ, kdy funkce  $f$  je lineární. Obecný tvar (nazývaný vícenásobná regrese) je

$$Y = a + b_1X_1 + \dots + b_rX_r + \epsilon.$$

Často však hledáme závislost veličiny  $Y$  pouze na jediné veličině  $X$  (tj.  $\mathbb{X}$  je jednorozměrný vektor). Tento případ, tj. vztah

$$Y = a + bX + \epsilon,$$

se nazývá jednoduchá regrese.

Máme-li data  $(y_i, x_{1i}, \dots, x_{ri}), i = 1, \dots, n$  a uvažujeme-li jednoduchou regresi, pak hodnoty

$$e_i = y_i - a - bx_i, \quad i = 1, \dots, n$$

se nazývají rezidua a lze je považovat za realizace chybového členu  $\epsilon$ .

Reziduální součet čtverců pro jednoduchou regresi je

$$S(a, b) = \sum_{i=1}^n (y_i - a - bx_i)^2.$$

### 3.2.1 Bodový odhad parametrů modelu

K nalezení odhadů parametrů  $a$  a  $b$  v modelu jednoduché regrese použijeme výše popsanou metodu nejmenších čtverců. Hledáme-li minimum  $S(a, b)$ , dostáváme rovnice

$$\begin{aligned} \frac{\partial}{\partial a} S(a, b) &= -2 \sum_{i=1}^n (y_i - a - bx_i) = 0 \\ \frac{\partial}{\partial b} S(a, b) &= -2 \sum_{i=1}^n (y_i - a - bx_i)x_i = 0 \end{aligned}$$



↓

$$\hat{b} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{a} = \bar{y} - \hat{b} \bar{x},$$

kde  $\hat{a}, \hat{b}$  jsou hledané bodové odhady parametrů  $a$  a  $b$ .

### 3.2.2 Intervalové odhady parametrů modelu

Označme  $S_R = S(a, b)$  reziduální součet čtverců, pak bodový odhad rozptylu  $s^2$  chybového členu  $\epsilon$  je

$$s^2 = \frac{S_R}{n - 2}.$$

Pomocí  $s^2$  lze vyjádřit odhady rozptylu regresních parametrů

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

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

Statistiky  $T_a = \frac{\hat{a}-a}{S_a}$  a  $T_b = \frac{\hat{b}-b}{S_b}$  pak navíc mají studentovo  $t$ -rozdělení o  $(n - 2)$  stupních volnosti, tudíž intervalové odhady pro  $a$  a  $b$  jsou

$$\hat{a} - S_a t_{n-2, 1-\frac{\alpha}{2}} \leq a \leq \hat{a} + S_a t_{n-2, 1-\frac{\alpha}{2}},$$

$$\hat{b} - S_b t_{n-2, 1-\frac{\alpha}{2}} \leq b \leq \hat{b} + S_b t_{n-2, 1-\frac{\alpha}{2}},$$

kde  $t_{n-2, 1-\frac{\alpha}{2}}$  je  $(1 - \frac{\alpha}{2})$ -kvantil  $t$ -rozdělení o  $n - 2$  stupních volnosti.

**Příklad 3.1** Mějme následující pozorování:

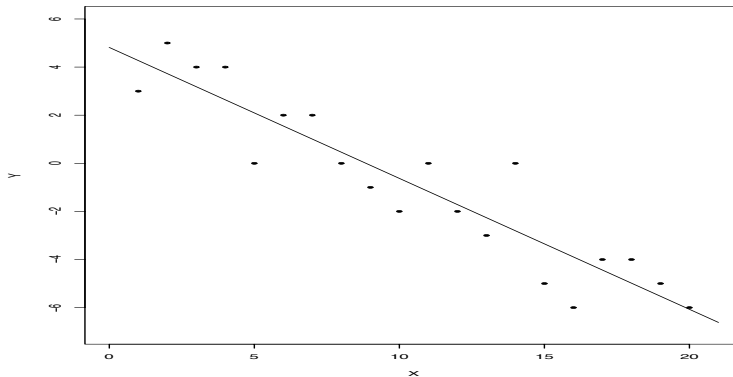
$x_i$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$y_i$	3	5	4	4	0	2	2	0	-1	-2	0	-2	-3	0	-5	-6	-4	-4	-5	-6

Bodové odhady parametrů  $a$  a  $b$  jsou  $\hat{a} = 4.816, \hat{b} = -0,544$ .

Daty proložená regresní přímka je znázorněná na Obrázku 3.1.

Intervalové odhady pro parametry  $a$  a  $b$  jsou

$$(3.499543 < a < 6.132036) \quad a \quad (-0.654239 < b < -0.4344828).$$



Obrázek 3.1: Regrení přímka pro data z Příkladu 3.1.

### 3.2.3 Testování podmodelů

Pro testování, zda se koeficienty modelu výrazně liší od nuly, používáme statistiky

$$T_a = \frac{\hat{a}}{S_a}, \quad \text{resp.} \quad T_b = \frac{\hat{b}}{S_b},$$

které mají za platnosti hypotézy  $H_0 : a = 0$ , resp.  $H_0 : b = 0$ ,  $t$ -rozdělení o  $(n - 2)$  stupních volnosti. Tedy hypotézu  $H_0 : a = 0$ , resp.  $H_0 : b = 0$ , zamítáme ve prospěch  $H_A : a \neq 0$ , resp.  $H_A : b \neq 0$ , na hladině významnosti  $\alpha$ , pokud  $|T_a| \geq t_{n-2, 1-\frac{\alpha}{2}}$ , resp.  $|T_b| \geq t_{n-2, 1-\frac{\alpha}{2}}$ .

**Příklad 3.2** Pro data z příkladu 3.1 dostáváme

$$|T_a| = |7.6867| = 7.6867 > 2.1009$$

*a*

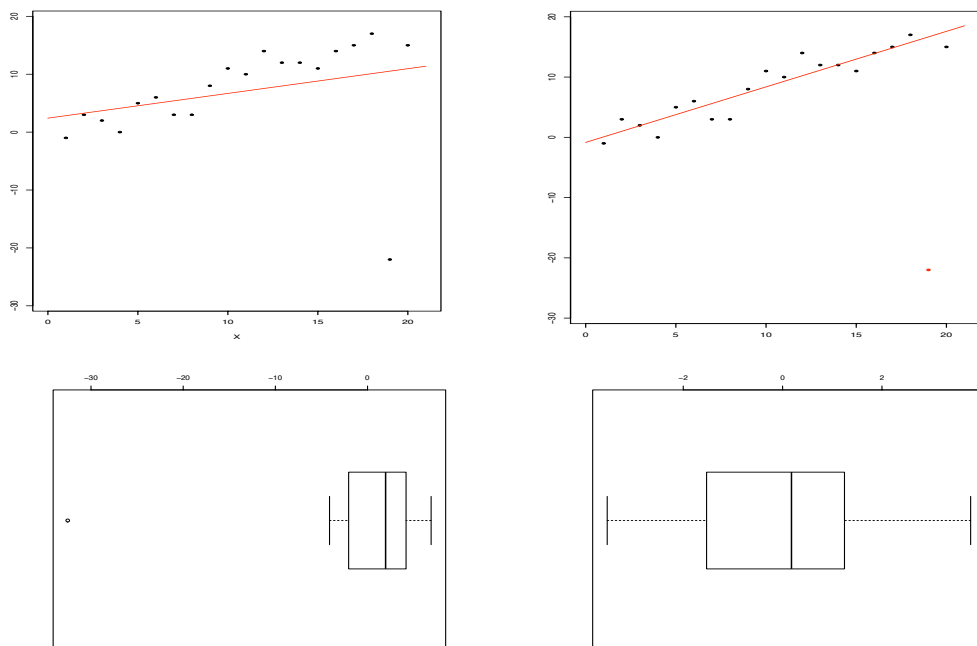
$$|T_b| = |-10.40844| = 10.4084 > 2.1009,$$

*tedy zamítáme jak hypotézu  $H_0 : a = 0$ , tak  $H_0 : b = 0$ , ve prospěch zmíněných alternativ.*

Pokud některý z koeficientů lze považovat za nulový, můžeme jej z modelu vynechat. Takto vytvořený nový model nazýváme podmodelem původního modelu. Tj. pokud např. vynecháme koeficient  $a$ , dostáváme speciální případ lineární regrese

$$Y = bX + \epsilon,$$

jehož regresní přímka prochází počátkem soustavy souřadnic. V takovém modelu navíc stačí odhadnout pouze jeden parametr, takže celková práce s tímto modelem se výrazně zjednoduší.



Obrázek 3.2: Regresní přímka proložená všemi daty, regresní přímka proložená daty bez odlehlého pozorování a boxploty příslušných reziduí.

**Poznámka 3.1** *Alternativou pro testování, zda lze daný koeficient považovat za nulový, je sestavení intervalů spolehlivosti (viz předešlá podkapitola) a zjištění, zda některý z těchto intervalů nulu obsahuje.*

### 3.2.4 Analýza reziduí a odlehlá pozorování

Na začátku kapitoly jsme uvedli předpoklad, že chybový člen  $\epsilon$  je náhodná veličina s nulovou střední hodnotou a většinou také s normálním rozdělením. Je tudíž třeba ověřit tyto předpoklady na odhadech chybových členů  $e_i = y_i - a - bx_i$ . Toto ověření se nazývá analýza reziduí a používají se pro ni již dříve popsané metody (histogram, boxplot, Q-Q plot apod.).

Při analýze reziduí lze navíc detekovat i odlehlá pozorování, což jsou pozorování nacházející se očividně stranou od ostatních dat. Jejich výskyt lze předpokládat právě ve chvíli, kdy rozdělení chybových členů vykazuje známky nesymetrie (viz Obrázek 3.2.4).

### 3.2.5 Výběr vhodného modelu

Pro kontrolu, jak vhodný je zvolený model, lze využít koeficientu determinace  $R^2$ , který určuje podíl části variability  $Y$  vysvětlené modelem ku celkové variabilitě  $Y$ . Ten je dán vztahem

$$R^2 = \frac{\sum_{i=1}^n (a + bx_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Jeho hodnoty se pohybují v intervalu  $(0,1)$ , přičemž větší hodnoty koeficientu determinace znamenají větší úspěšnost regrese.

**Příklad 3.3** Pro data z obrázku dostáváme  $R^2 = 0.084$  (špatná data), resp.  $R^2 = 0.876$  (data bez odlehlého pozorování).

# Kapitola 4

## Náhodné procesy

### 4.1 Základní pojmy

**Definice 4.1** *Nechť  $(\Omega, \mathcal{A}, P)$  je pravděpodobnostní prostor a  $T \subset \mathbb{R}$ . Rodina reálných náhodných veličin  $\{X_t, t \in T\}$  definovaných na  $(\Omega, \mathcal{A}, P)$  se nazývá náhodný (nebo také stochastický) proces.*

**Definice 4.2** *Je-li  $T = \mathbb{Z}$  nebo  $T = \mathbb{N}$ , mluvíme o náhodném procesu s diskrétním časem. Je-li  $T = [a, b]$ , kde  $-\infty \leq a < b \leq \infty$ , mluvíme o náhodném procesu se spojitým časem.*

**Definice 4.3** *Dvojice  $(S, \mathcal{E})$ , kde  $S$  je množina hodnot náhodných veličin  $X_t$  a  $\mathcal{E}$  je  $\sigma$ -algebra na množině  $S$ , se nazývá stavový prostor.*

*Pokud náhodné veličiny  $X_t$  nabývají pouze diskrétních hodnot, mluvíme o náhodném procesu s diskrétními stavy. Pokud náhodné veličiny  $X_t$  nabývají spojitých hodnot, mluvíme o náhodném procesu se spojitými stavy.*

Náhodný proces  $\{X_t, t \in T\}$  můžeme chápat jako funkci dvou proměnných  $\omega$  a  $t$ . Pro pevné  $t$  je tato funkce náhodnou veličinou, pro pevné  $\omega$  se jedná o funkci jedné reálné proměnné  $t$ .

**Definice 4.4** *Mějme pevné  $\omega \in \Omega$ . Pak funkce  $t \rightarrow X_t$  se nazývá trajektorie procesu  $\{X_t, t \in T\}$ .*

**Definice 4.5** *Proces se nazývá spojitý, jsou-li všechny jeho trajektorie spojité.*

**Definice 4.6** *Nechť  $\{X_t, t \in T\}$  je náhodný proces takový, že pro každé  $t \in T$  existuje střední hodnota  $\mathbb{E}X_t$ . Potom funkce  $\mu_t = \mathbb{E}X_t$  definovaná na  $T$  se nazývá střední hodnota procesu  $\{X_t\}$ . Jestliže platí  $\mathbb{E}|X_t|^2 < \infty$  pro všechna  $t \in T$ , potom funkce dvou proměnných definovaná na  $T \times T$  předpisem  $R(s, t) = \mathbb{E}(X_s - \mu_s)(X_t - \mu_t)$  se nazývá autokovarianční funkce procesu  $\{X_t\}$ . Hodnota  $R(t, t)$  se nazývá rozptyl procesu  $\{X_t\}$  v čase  $t$ .*

**Definice 4.7** *Řekneme, že náhodný proces  $\{X_t, t \in T\}$  je slabě stacionární, jestliže  $R(s, t)$  je funkcí pouze rozdílu  $s - t$ , tj.*

$$R(s, t) = \tilde{R}(s - t)$$

Důsledkem této definice je rovnost  $R(s, t) = R(s + h, t + h)$  pro každé  $h \in \mathbb{R}$  takové, že  $s + h \in T$  a  $t + h \in T$ .

Každé konečné podmnožině  $\{t_1, \dots, t_n\} \in T$  lze přiřadit systém náhodných veličin, které mají sdruženou distribuční funkci

$$F_{t_1, \dots, t_n}(x_1, \dots, x_n) = P(X_{t_1} \leq x_1, \dots, X_{t_n} \leq x_n).$$

**Definice 4.8** *Řekneme, že náhodný proces  $\{X_t, t \in T\}$  je striktně stacionární, jestliže pro libovolné  $n \in \mathbb{N}$ , pro libovolná reálná  $x_1, \dots, x_n$  a pro libovolná reálná  $t_1, \dots, t_n$  a  $h$  taková, že  $t_k \in T$ ,  $t_k + h \in T$ ,  $1 \leq k \leq n$ , platí*

$$F_{t_1, \dots, t_n}(x_1, \dots, x_n) = F_{t_1+h, \dots, t_n+h}(x_1, \dots, x_n). \quad (4.1)$$

**Poznámka 4.1** *Vztah (4.1) je ekvivalentní vztahu*

$$P(X_{t_1} = x_1, \dots, X_{t_n} = x_n) = P(X_{t_1+h} = x_1, \dots, X_{t_n+h} = x_n).$$

**Definice 4.9** *Nechť náhodné procesy  $\{X_t, t \in T\}$  a  $\{Y_t, t \in T\}$  definované na stejném pravděpodobnostním prostoru s hodnotami ve stejném stavovém prostoru. Pak*

a)  $\{X_t\}$  a  $\{Y_t\}$  jsou stochasticky ekvivalentní, jestliže

$$P(X_t = Y_t) = P(\omega : X_t(\omega) = Y_t(\omega)) = 1, \quad \forall t \in T.$$

Říkáme také, že proces  $\{X_t\}$  je stochastickou verzí, popř. modifikací, procesu  $\{Y_t\}$ .

b)  $\{X_t\}$  a  $\{Y_t\}$  jsou nerozlišitelné, jestliže

$$P(X_t = Y_t, \forall t \in T) = P(\omega : X_t(\omega) = Y_t(\omega), \forall t \in T) = 1.$$

## 4.2 Markovské řetězce s diskretním časem

V této a následující kapitole budeme často místo pojmu proces používat pojmu řetězec. Důvodem je, že množina hodnot, kterých může proces nabývat je spočetná a ke změnám nabývaných hodnot dochází skokovitě, tzn. nabývané hodnoty tvoří řetězec hodnot. Všechny definice pro procesy uvedené v kapitole 4.1 samozřejmě platí i zde.

### 4.2.1 Základní definice a vlastnosti

Mějme pravděpodobnostní prostor  $(\Omega, \mathcal{A}, P)$ , na něm posloupnost náhodných veličin  $\{X_n, n \in \mathbb{N}\}$  a stavový prostor  $(S, \mathcal{E})$ , kde množina  $S$  může být buď konečná nebo spočetná, bez újmy na obecnosti předpokládejme, že  $S = \{0, 1, \dots, N\}$ , resp.  $S = \{0, 1, \dots\}$

**Definice 4.10** *Posloupnost náhodných veličin  $\{X_n, n \in \mathbb{N}\}$  nazveme Markovský řetězec s diskretním časem, jestliže*

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P(X_{n+1} = j | X_n = i)$$

pro všechna  $n = 0, 1, \dots$  a všechna  $i, j, i_{n-1}, \dots, i_0 \in S$  taková, že  $P(X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) > 0$ .

**Příklad 4.1** *Typickým příkladem Markovského řetězce je tzv. náhodná procházka. Nechť  $Y_1, Y_2, \dots$  jsou nezávislé stejně rozdělené náhodné veličiny nabývající hodnot  $\pm 1$  s pravděpodobnostmi  $1/2$ . Definujme  $X_0 = 0$  a  $X_n = \sum_{i=1}^n Y_i$ . Pak posloupnost (proces, řetězec)  $\{X_n, n \in \mathbb{N}\}$  se nazývá náhodná procházka.*

**Definice 4.11** *Podmíněné pravděpodobnosti*

- a)  $P(X_{n+1} = j | X_n = i) = p_{ij}(n, n+1)$  nazveme pravděpodobnostmi přechodu ze stavu  $i$  v čase  $n$  do stavu  $j$  v čase  $n+1$  nebo také pravděpodobnostmi přechodu 1.řádu;
- b)  $P(X_{n+m} = j | X_n = i) = p_{ij}(n, n+m)$  nazveme pravděpodobnostmi přechodu ze stavu  $i$  v čase  $n$  do stavu  $j$  v čase  $n+m$  nebo také pravděpodobnostmi přechodu  $m$ -tého řádu.

**Definice 4.12** Jestliže pravděpodobnosti přechodu  $p_{ij}(n, n + m)$  nezávisí na časových okamžicích  $n$  a  $n + m$ , ale pouze na rozdílu  $m$ , nazývá se příslušný Markovský řetězec homogenní.

Uvažujme homogenní řetězec a označme zkráceně  $p_{ij}(n, n + 1)$  jako  $p_{ij}$ . Tyto prvky lze seřadit do čtvercové matice  $\mathbf{P} = \{p_{ij}, i, j \in S\}$ , pro niž zřejmě platí

$$p_{ij} \geq 0, \forall i, j \in S; \quad \text{a} \quad \sum_{j \in S} p_{ij} = 1, \forall i \in S.$$

**Definice 4.13** Matice  $\mathbf{P} = \{p_{ij}, i, j \in S\}$  se nazývá matice pravděpodobností přechodu.

Označme dále

$$p_i = P(X_0 = i), \quad \forall i \in S,$$

pro které zřejmě platí

$$p_i \geq 0, \forall i \in S \quad \text{a} \quad \sum_{i \in S} p_i = 1.$$

**Definice 4.14** Vektor  $\mathbf{p} = \{p_i, i \in S\}$  se nazývá počáteční rozdělení Markovského řetězce.

Dá se pak ukázat, že konečněrozměrná rozdělení procesu  $\{X_n, n \in \mathbb{N}\}$  s počátečním rozdělením  $\mathbf{p}$  a maticí pravděpodobností přechodu  $\mathbf{P}$  je dáno vztahem

$$P(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) = p_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i_n}.$$

Označme dále  $p_{ij}^{(1)} = p_{ij}$  a definujme pro přirozené  $n \geq 1$  postupně

$$p_{ij}^{(n+1)} = \sum_{k \in S} p_{ik}^{(n)} p_{kj}. \quad (4.2)$$

Lze ukázat, že  $p_{ij}^{(n)} \leq 1$  a navíc pro matice pravděpodobností přechodů platí

$$\mathbf{P}^{(2)} = \mathbf{P} \cdot \mathbf{P} = \mathbf{P}^2 \quad \text{a} \quad \text{obecně} \quad \mathbf{P}^{(n+1)} = \mathbf{P}^{(n)} \cdot \mathbf{P} = \mathbf{P} \cdot \mathbf{P}^{(n)} = \mathbf{P}^{n+1}.$$

**Věta 4.1** Necht'  $\{X_n, n \in \mathbb{N}\}$  je homogenní Markovský řetězec s maticí přechodu  $\mathbf{P}$ . Potom pro pravděpodobnosti přechodu  $n$ -tého řádu platí

$$P(X_{m+n} = j | X_m = i) = p_{ij}^{(n)}, \quad \forall i, j \in S$$

pro všechna přirozená  $m$  a  $n$  a pro  $P(X_m = i) > 0$ .



Vztah (4.2) lze zobecnit. Toto zobecnění se nazývá **Chapman-Kolmogorova rovnost** a je definována jako

$$p_{ij}^{(m+n)} = \sum_{k \in S} p_{ik}^{(m)} p_{kj}^{(n)},$$

zapsáno maticově

$$\mathbf{P}^{(m+n)} = \mathbf{P}^{(m)} \cdot \mathbf{P}^{(n)}.$$

## 4.2.2 Klasifikace stavů Markovského řetězce

V této kapitole se budeme zabývat pouze homogenními Markovskými řetězci. Používat pak budeme následující značení:

- Vychází-li řetězec  $\{X_n, n \in \mathbb{N}\}$  ze stavu  $j$ , tj.  $P(X_0 = j) = 1$ , pak označíme

$$P(\cdot | X_0 = j) = P_j(\cdot).$$

- Položme  $\tau_j(0) = 0$  a definujme náhodné veličiny

$$\tau_j(1) = \inf\{n > 0 : X_n = j\}$$

čas prvního návratu řetězce do stavu  $j$ .

- Střední hodnotu doby prvního návratu označíme  $\mu_j = \mathbb{E}[\tau_j(1) | X_0 = j]$ .
- Největší společný dělitel čísel  $n \geq 1$ , pro které  $p_{jj}^{(n)} > 0$ , označíme  $d_j$ .

**Definice 4.15** Stav  $j$  Markovského řetězce se nazývá *trvalý*, jestliže

$$P_j(\tau_j(1) < \infty) = 1.$$

Stav  $j$  Markovského řetězce se nazývá *přechodný*, jestliže

$$P_j(\tau_j(1) = \infty) > 0.$$

**Definice 4.16** Trvalý stav  $j$  Markovského řetězce se nazývá *nenulový*, jestliže  $\mu_j < \infty$  a *nulový*, jestliže  $\mu_j = \infty$ .

**Definice 4.17** Je-li  $d_j > 1$ , stav  $j$  Markovského řetězce se nazývá *periodický* s periodou  $d_j$ , je-li  $d_j = 1$ , stav  $j$  Markovského řetězce se nazývá *neperiodický*.

#### Věta 4.2 (Konvergence pravděpodobností přechodu)

- a) Nechť  $j$  je přechodný stav. Potom  $\lim_{n \rightarrow \infty} p_{ij}^{(n)} = 0, \forall i \in S$ .
- b) Nechť  $j$  je trvalý nulový stav. Potom  $\lim_{n \rightarrow \infty} p_{ij}^{(n)} = 0, \forall i \in S$ .
- c) Nechť  $j$  je trvalý nenulový a neperiodický stav. Potom  $\lim_{n \rightarrow \infty} p_{jj}^{(n)} = \frac{1}{\mu_j}$ .
- d) Nechť  $j$  je trvalý nenulový stav s periodou  $d_j$ . Potom  $\lim_{n \rightarrow \infty} p_{jj}^{(nd_j)} = \frac{d_j}{\mu_j}$ .

**Věta 4.3** Trvalý stav  $j$  je nulový právě tehdy, když  $\lim_{n \rightarrow \infty} p_{jj}^{(n)} = 0$ .

Uvažujme nyní řetězec s množinou přechodných stavů  $R$  a definujme náhodnou veličinu

$$\tau = \inf\{n \geq 0 : X_n \notin R\}$$

značící čas výstupu z množiny přechodných stavů.

**Věta 4.4** V řetězci s konečně mnoha stavy je

$$P_i(\tau = \infty) = 0, \quad i \in R.$$

**Definice 4.18** Řekneme, že stav  $j$  je dosažitelný ze stavu  $i$ , jestliže existuje  $n \in \mathbb{N}$  takové, že  $p_{ij}^{(n)} > 0$ . Jestliže  $p_{ij}^{(n)} = 0$ , pro všechna  $n \in \mathbb{N}$ , pak říkáme, že stav  $j$  je nedosažitelný ze stavu  $i$ .

**Definice 4.19** Množina stavů  $C$  se nazývá uzavřená, jestliže žádný stav vně  $C$  není dosažitelný z žádného stavu uvnitř  $C$ .

**Věta 4.5** Množina stavů je uzavřená právě tehdy, je-li  $p_{ij} = 0$  pro všechna  $i \in C, j \notin C$ .

**Definice 4.20** Markovský řetězec se nazývá nerozložitelný, jestliže každý jeho stav je dosažitelný z každého jiného stavu. V opačném případě je řetězec rozložitelný.

**Definice 4.21** Je-li jednobodová množina stavů  $\{j\}$  uzavřená, tj. je-li  $p_{jj} = 1$ , pak se stav  $j$  nazývá absorpční.

**Definice 4.22** Řetězec s konečně mnoha stavy, jehož všechny trvalé stavy jsou absorpční, se nazývá absorpční řetězec.

### 4.2.3 Stacionární rozdělení

**Definice 4.23** *Nechť  $\{X_n, n \in \mathbb{N}\}$  je homogenní řetězec s množinou stavů  $S$  a maticí pravděpodobností přechodu  $\mathbf{P}$ . Nechť  $\pi = \{\pi_j, j \in S\}$  je nějaké pravděpodobnostní rozdělení na množině  $S$ , tj.  $\pi_j \geq 0, j \in S, \sum_{j \in S} \pi_j = 1$ . Potom  $\pi$  se nazývá stacionární rozdělení daného řetězce, jestliže platí*

$$\pi^T = \pi^T \mathbf{P},$$

neboli

$$\pi_j = \sum_{k \in S} \pi_k p_{kj}, j \in S.$$

**Věta 4.6** *Nechť počáteční rozdělení homogenního Markovského řetězce je stacionární. Pak je tento řetězec striktně stacionární a pro všechna  $n \in \mathbb{N}$  platí*

$$p_j(n) = P(X_n = j) = \pi_j, \quad j \in S,$$

kde  $\pi_j$  jsou počáteční stacionární pravděpodobnosti.

**Věta 4.7** *Pro nerozložitelný Markovský řetězec platí*

a) *Jsou-li všechny jeho stavy přechodné nebo všechny trvalé nulové, stacionární rozdělení neexistuje.*

b) *Jsou-li všechny jeho stavy trvalé nenulové, stacionární rozdělení existuje a je jednoznačné.*

(i) *Jsou-li všechny stavy neperiodické, potom pro stacionární pravděpodobnosti platí*

$$\pi_j = \lim_{n \rightarrow \infty} p_{ij}^{(n)} > 0, \quad i, j \in S,$$

$$\pi_j = \lim_{n \rightarrow \infty} p_j(n) > 0, \quad j \in S.$$

(ii) *Jsou-li všechny stavy periodické, potom pro stacionární pravděpodobnosti platí*

$$\pi_j = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n p_{ij}^{(k)} > 0, \quad i, j \in S,$$

$$\pi_j = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n p_j(k) > 0, \quad j \in S.$$

– *V nerozložitelném řetězci s konečně mnoha stavy stacionární rozdělení existuje.*

## 4.3 Markovské řetězce se spojitým časem

### 4.3.1 Základní definice a vlastnosti

**Definice 4.24** *Systém celočíselných náhodných veličin  $\{X_t, t \geq 0\}$  definovaných na pravděpodobnostním prostoru  $(\Omega, \mathcal{A}, P)$  nazveme Markovský řetězec se spojitým časem, jestliže*

$$P(X_t = j | X_s = i, X_{t_n} = i_n, \dots, X_{t_1} = i_1) = P(X_t = j | X_s = i) \quad (4.3)$$

pro všechna  $0 \leq t_1 < \dots < t_n < s < t$  a všechna  $i, j, i_n, \dots, i_1 \in S$  taková, že  $P(X_s = i, X_{t_n} = i_n, \dots, X_{t_1} = i_1) > 0$ .

**Poznámka 4.2** *Budeme používat podobné názvosloví a značení jako v předešlé kapitole:*

- $P(X_t = j | X_s = i) = p_{ij}(s, t)$  nazveme pravděpodobnostmi přechodu ze stavu  $i$  v čase  $s$  do stavu  $j$  v čase  $t$ ;
- pro homogenní řetězec, kde pravděpodobnosti přechodu závisí pouze na rozdílech časů, budeme značit  $p_{ij}(s, s + t)$  jako  $p_{ij}(t)$ ;
- absolutní pravděpodobnosti budeme značit  $p_j(t) = P(X_t = j), j \in S$  a  $p_j(0) = P(X_0 = j), j \in S$  pak budou počáteční pravděpodobnosti.

Obdobně jako v případě diskrétních řetězců v tomto případě pro každé  $t \geq 0$  platí

$$p_{ij}(t) \geq 0, \forall i, j \in S \quad \text{a} \quad \sum_{j \in S} p_{ij}(t) = 1, \forall i \in S,$$

$$p_i(t) \geq 0, \forall i \in S \quad \text{a} \quad \sum_{i \in S} p_i(t) = 1.$$

Pro každé  $t$  tak lze tyto prvky seřadit do čtvercové matice  $\mathbf{P}(t) = \{p_{ij}(t), i, j \in S\}$ , čímž dostáváme systém matic pravděpodobností přechodu  $\{\mathbf{P}(t), t \geq 0\}$ , přičemž  $\{\mathbf{P}(0) = I\}$ , kde  $I$  značí jednotkovou matici (tj. matici s jedničkami na diagonále a nulami jinde).

Pro absolutní pravděpodobnosti dostáváme analogicky s předešlou kapitolou pro  $i, j \in S$  vztah

$$p_j(t) = \sum_{i \in S} p_i(0) p_{ij}(t),$$

zapsáno maticově

$$\mathbf{p}(t)^T = \mathbf{p}(0)^T \cdot \mathbf{P}(t).$$

Ten lze opět zobecnit na

$$p_{ij}(s+t) = \sum_{k \in S} p_{ik}(s)p_{kj}(t),$$

zapsáno maticově

$$\mathbf{P}(s+t) = \mathbf{P}(s) \cdot \mathbf{P}(t),$$

což je opět **Chapman-Kolmogorova rovnost**.

V dalším textu budeme předpokládat, že

$$\lim_{t \rightarrow 0^+} p_{ij}(t) = \delta_{ij}, \quad i, j \in S, \quad (4.4)$$

kde  $\delta_{ij}$  značí Dirackovu funkci, tj.  $\delta_{ij} = 1$  pro  $i = j$  a  $\delta_{ij} = 0$  jinak. Tento předpoklad společně se skutečností, že  $p_{ij}(0) = \delta_{ij}$  znamená, že řetězec je zprava spojitý v 0.

Dále si označme

$$\lim_{h \rightarrow 0^+} \frac{1 - p_{ii}(h)}{h} := q_i \quad \text{a} \quad \lim_{h \rightarrow 0^+} \frac{p_{ij}(h)}{h} := q_{ij}. \quad (4.5)$$

**Definice 4.25** *Nezáporná čísla  $q_{ij}$  z (4.5) se nazývají intenzity přechodu, číslo  $q_i$  z (4.5) je pak celková intenzita. Matice  $\mathbf{Q} = \{q_{ij}, i, j \in S\}$ , kde  $q_{ii} = -q_i$  se nazývá matice intenzit přechodu.*

**Věta 4.8** *Pro homogenní Markovský řetězec se spočetnou množinou stavů platí pro všechna  $s \geq 0$  a všechna  $h \geq 0$*

$$P(X_t = i \text{ pro } t \in (s, s+h) | X_s = i) = e^{-q_i h}.$$

**Věta 4.9** *Je-li  $q_i = 0$ , pak  $p_{ii}(t) = 1$  pro všechna  $t \geq 0$ . Je-li  $0 < q_i < \infty$ , má doba, po kterou řetězec setrvává ve stavu  $i$ , exponenciální rozdělení s parametrem  $q_i$ .*

**Věta 4.10** *Nechť  $0 < q_i < \infty$ . Potom pravděpodobnost, že řetězec z počátečního stavu  $i$  přejde nejdříve do stavu  $j$ , je rovna  $\frac{q_{ij}}{q_i}$  pro všechna  $j \neq i$ .*

### 4.3.2 Klasifikace stavů Markovského řetězce

**Definice 4.26** *Stav  $j \in S$  se nazývá absorpční, jestliže  $q_j = 0$ . Jestliže  $q_i > 0$ , pak se stav  $j$  nazývá stabilní, pokud  $q_j < \infty$ , a nestabilní, pokud  $q_j = \infty$ .*

Dále uvažujme  $P(X_0 = j) = 1$ , tj. řetězec startující ze stavu  $j$ . Označme  $J$  čas, kdy tento řetězec poprvé opustí stav  $j$ , a

$$\tau_j(1) = \inf\{t \geq J : X_t = j\}$$

čas prvního návratu tohoto řetězce do  $j$ .

**Definice 4.27** Stav  $j$  Markovského řetězce se nazývá *trvalý*, jestliže buď  $q_j = 0$  nebo  $q_j > 0$  a zároveň

$$P_j(\tau_j(1) < \infty) = 1.$$

Stav  $j$  Markovského řetězce se nazývá *přechodný*, jestliže  $q_j > 0$  a zároveň

$$P_j(\tau_j(1) = \infty) > 0.$$

**Definice 4.28** Trvalý stav  $j$  Markovského řetězce se nazývá *nenulový*, jestliže buď  $q_j = 0$  nebo  $\mathbb{E}[\tau_j(1)] < \infty$ . V opačném případě se řetězec nazývá *nulový*.

**Definice 4.29** Řekneme, že stav  $j$  je *dosažitelný* ze stavu  $i$ , jestliže existuje  $t > 0$  takové, že  $p_{ij}(t) > 0$ .

**Poznámka 4.3** Analogicky jako pro řetězce s diskrétním časem můžeme definovat také nerozložitelnost řetězce se spojitým časem.

### 4.3.3 Stacionární rozdělení

**Definice 4.30** Nechť  $\{X_t, t \geq 0\}$  je homogenní řetězec se spojitým časem, množinou stavů  $S$  a maticemi pravděpodobností přechodu  $\mathbf{P}(t), t \geq 0$ . Potom  $\pi$  se nazývá *stacionární rozdělení* daného řetězce, jestliže platí

$$\pi^T = \pi^T \mathbf{P}(t), \quad \forall t \geq 0.$$

**Věta 4.11** Nechť počáteční rozdělení homogenního Markovského řetězce  $\{X_t, t \geq 0\}$  je stacionární. Pak je tento řetězec striktně stacionární a pro všechna  $t \geq 0$  platí

$$p_j(t) = P(X_t = j) = \pi_j, \quad j \in S,$$

kde  $\pi_j$  jsou počáteční stacionární pravděpodobnosti.

## 4.4 Nejčastěji používané Markovské procesy

V této části si uvedeme dva procesy, jichž se hojně využívá v matematice pojištění (zejména Poissonův) a finanční (zejména Wienerův). Jejich konkrétní aplikace budou společně s dalšími aplikacemi teorie náhodných procesů ukázány v dalších kapitolách.

### Poissonův proces

*Poissonův proces*, obvykle značený  $\{N_t, t \geq 0\}$ , je proces popisující počet nějakých událostí v čase (např. počet pojistných událostí, které nastaly do času  $t$ ). Pro tento proces se předpokládá, že

- počty událostí v disjunktních časových intervalech jsou nezávislé náhodné veličiny, tj. pro  $t_1 < t_2 \leq t_3 < t_4$  jsou veličiny  $N_{t_2} - N_{t_1}$  a  $N_{t_4} - N_{t_3}$  nezávislé - zkráceně se takovému procesu říká proces s nezávislými přírůstky,
- počty událostí v časovém intervalu  $(t, t + h)$  závisí pouze na délce intervalu  $h$ ,
- pro počty událostí v časovém intervalu  $(t, t + h)$  platí

$$P(N_{t+h} - N_t = 0) = 1 - \lambda h + o(h),$$

$$P(N_{t+h} - N_t = 1) = \lambda h + o(h),$$

$$P(N_{t+h} - N_t \geq 2) = o(h),$$

kde symbol  $o(h)$  značí, že  $o(h)/h \rightarrow 0$  při  $h \rightarrow 0+$ , a  $\lambda$  je konstanta, která se nazývá intenzita Poissonova procesu.

Z předpokladu nezávislosti přírůstků plyne Markovská vlastnost tohoto procesu a pro pravděpodobnosti přechodu platí

$$\begin{aligned} P(N_{t+h} = j | N_t = i) &= \lambda h + o(h) && j = i + 1 \\ &= 1 - \lambda h + o(h) && j = i \\ &= o(h) && j > i + 1 \\ &= 0 && j < i. \end{aligned}$$

Intenzity přechodu jsou

$$q_{i,i+1} = \lambda, \quad q_i = -q_{ii} = \lambda, \quad q_{ij} = 0 \text{ jinak.}$$

Navíc se dá ukázat, že pro tento proces platí

$$P(N_t = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \quad k = 0, 1, \dots,$$

což je Poissonovo rozdělení.

## Wienerův proces

Tento proces není sice Markovským řetězcem, neboť množina hodnot, kterých může nabývat, je nespočetná (všechna reálná čísla - je to tudíž náhodný proces se spojitými stavy), ale je potřeba ho zde uvést, neboť patří také mezi Markovské procesy. Splňuje totiž Markovskou vlastnost definovanou vztahem (4.3) a navíc je velice významným procesem, kterého se využívá zejména ve finančních modelech, kde popisuje odchylku hodnoty daného aktiva od trendové složky.

*Wienerův proces* (někdy také nazýván *Brownův pohyb*)  $\{W_t, t \geq 0\}$  je definován následujícími vlastnostmi:

- $\{W_t, t \geq 0\}$  má spojitě trajektorie,
- $W_0 = 0$ ,
- $\{W_t, t \geq 0\}$  má nezávislé přírůstky,
- přírůstek hodnoty v časovém intervalu  $(s, t)$  má normální rozdělení s nulovou střední hodnotou a rozptylem  $\sigma^2(t - s)$ , kde  $\sigma^2$  je kladná konstanta.

S Wienerovým procesem a jeho dalšími vlastnostmi plynoucími z této definice se setkáme v další kapitole.

## 4.5 Náhodné procesy ve finanční matematice

Tato kapitola popisuje konstrukce matematických modelů používaných hojně např. ve financích pro popis vývoje cen aktiv na finančních trzích.



### 4.5.1 Základní pojmy

Než se budeme věnovat této části náhodných procesů, uvedeme nejprve několik potřebných značení a pojmů.

V celé kapitole budeme uvažovat pravděpodobnostní prostor  $(\Omega, \mathcal{A}, P)$  a používat značení  $L_p(\Omega, \mathcal{A}, P)$  podle kontextu buď pro množinu náhodných veličin  $X$  definovaných na  $(\Omega, \mathcal{A}, P)$  splňujících  $\mathbb{E}|X|^p < \infty$  nebo pro množinu náhodných procesů  $\{X_t, t \geq 0\}$  definovaných na  $(\Omega, \mathcal{A}, P)$  splňujících  $\mathbb{E}|X_t|^p < \infty$  pro každé  $t$  (zpravidla budeme používat  $p = 1$  nebo  $p = 2$ ).

Dalším novým značením pak bude

$$\int_A X dP = \int_A X(\omega) dP(\omega), \quad \text{pro } A \in \mathcal{A}, \quad (4.6)$$

jehož interpretace je následující:

- v případě, že  $X$  je diskrétní náhodná veličina, je vztah (4.6) roven

$$\sum_{k: X(\omega)=k, \omega \in A} kP(X = k),$$

- v případě, že  $X$  je spojitá náhodná veličina s hustotou  $f$ , odpovídá vztah (4.6) výrazu

$$\int_{H: X(\omega) \in H, \omega \in A} x f(x) dx.$$

Připomeňme také pojem měřitelnosti náhodné veličiny:

Náhodná veličina je měřitelná vzhledem k  $\sigma$ -algebře  $\mathcal{A}$ , jestliže pro každé  $c \in \mathbb{R}$  platí

$$\{\omega : X(\omega) > c\} \in \mathcal{A}.$$

**Definice 4.31** *Nechť  $X \in L_1(\Omega, \mathcal{A}, P)$  a  $\mathcal{B} \subset \mathcal{A}$  je  $\sigma$ -algebra. Pak náhodnou veličinu  $Y$  nazveme podmíněnou střední hodnotou vzhledem k  $\mathcal{B}$ , jestliže*

a)  $Y \in L_1(\Omega, \mathcal{B}, P)$

b) pro každou množinu  $B \in \mathcal{B}$  platí

$$\int_B Y dP = \int_B X dP$$

a označovat ji budeme jako  $\mathbb{E}[X|\mathcal{B}]$ .

Vraťme se nyní k náhodným procesům.

**Definice 4.32** Filtrace  $\{\mathcal{F}_t, t \geq 0\}$  je systém  $\sigma$ -algeber  $\mathcal{F}_t \subseteq \mathcal{A}$  takových, že pro každé  $0 \leq s < t$  platí  $\mathcal{F}_s \subseteq \mathcal{F}_t$ .

**Definice 4.33** Řekneme, že stochastický proces  $\{X_t, t \geq 0\}$  je adaptovaný na filtraci  $\{\mathcal{F}_t\}$  (zkráceně  $\mathcal{F}_t$ -adaptovaný), jestliže pro každé  $t$  je  $X_t$  náhodná veličina měřitelná vzhledem k  $\sigma$ -algebře  $\mathcal{F}_t$ .

Množinu  $\mathcal{F}_t$ -adaptovaných procesů z  $L_p(\Omega, \mathcal{A}, P)$ ,  $p = 1, 2$ , budeme značit  $L_p(\mathcal{F}_t)$ .

## 4.5.2 Martingal

**Definice 4.34** Necht'  $\{X_t\} \in L_1(\mathcal{F}_t)$ . Pak se tento proces nazývá  $\mathcal{F}_t$ -martingal, jestliže

$$\mathbb{E}[X_t|\mathcal{F}_s] = X_s \quad \text{s.j.} \quad 0 \leq s \leq t < \infty.$$

**Příklad 4.2** Následující procesy jsou příklady martingalů:

1. Wienerův proces,
2.  $\{W_t^2 - t\}$ , kde  $\{W_t\}$  je Wienerův proces,
3. Poissonův proces s kompenzátorem  $\{M_t\} = \{N_t - \lambda t\}$ , kde  $\{N_t\}$  je Poissonův proces s intenzitou  $\lambda$  (připomeňme, že  $\mathbb{E}N_t = \lambda t$ ).

## 4.5.3 Stochastický integrál

**Definice 4.35** Necht'  $\{\mathcal{F}_t\}$  je filtrace,  $\Delta = \{0 = t_0 < t_1 < \dots < t_n < \infty\}$  je dělení časového intervalu  $[0, \infty)$  a  $\{G^{(j)}\}$  posloupnost náhodných veličin takových, že pro každé  $j$  je  $G^{(j)}$  měřitelná vzhledem k  $\mathcal{F}_{t_j}$ . Jednoduchým procesem nazveme proces  $G = \{G_t, t \geq 0\}$ , kde

$$G_t = \sum_{j=0}^{\infty} G^{(j)} \mathbf{1}_{[t_j, t_{j+1})}(t),$$

přičemž indikátor  $\mathbf{1}_{[t_j, t_{j+1})}(t) = 1$ , pokud  $t \in [t_j, t_{j+1})$ , a  $\mathbf{1}_{[t_j, t_{j+1})}(t) = 0$  jinak.

**Definice 4.36** *Nechť  $G$  je jednoduchý  $\mathcal{F}_t$ -adaptovaný proces a  $M$  je martingal. Označme pro každé  $t \in [t_k, t_{k+1})$*

$$\int_0^t G dM := \sum_{j=0}^{k-1} G_{t_j} (M_{t_{j+1}} - M_{t_j}) + G_t (M_t - M_{t_k}).$$

*Pak proces  $\{\int_0^t G dM, t \geq 0\}$  nazýváme (Itôův) stochastický integrál procesu  $G$  podle procesu  $M$ .*

Budeme-li pak dělení časového intervalu  $\Delta$  neustále zjemňovat, získáme tak limitním přechodem z jednoduchého procesu obecný proces, čímž máme stochastický integrál definovaný i pro procesy, které nejsou jednoduché.

#### 4.5.4 Itôův proces a jeho diferenciál, Itôova formule

**Definice 4.37** *Nechť  $W$  je  $\mathcal{F}_t$ -adaptovaný Wienerův proces. Mějme procesy  $A \in L_1(\mathcal{F}_t)$ ,  $B \in L_2(\mathcal{F}_t)$  a náhodnou veličinu  $X_0 \in L_1(\Omega, \mathcal{A}, P)$ . Pak proces  $X = \{X_t, t \geq 0\}$ , kde*

$$X_t = X_0 + \int_0^t A_s ds + \int_0^t B_s dW_s,$$

*se nazývá Itôův proces a jeho přírůstek v čase*

$$dX_t = A_t dt + B_t dW_t$$

*se nazývá stochastický diferenciál.*

Ve finanční terminologii se pro proces  $A$  používá název *drift* a procesu  $B$  se říká *volatilita*.

Právě Itôův je tím nejčastěji používaným modelem pro vývoj hodnoty aktiv na finančním trhu. Jelikož ale potřebujeme mnohdy pracovat nikoliv s daným aktivem, ale s nějakým derivátem, jehož hodnota je na hodnotě daného aktiva závislá, nepotřebujeme pracovat s hodnotou aktiva samotnou, nýbrž s nějakou její funkcí. K tomu tam slouží následující věta.

#### **Věta 4.12 (Itôova formule)**

*Nechť  $X$  je Itôův proces a  $f$  je spojitá, dvakrát diferencovatelná reálná funkce. Pak*

$$f(X_t) = f(X_0) + \int_0^t f'(X_s) A_s ds + \int_0^t f'(X_s) B_s dW_s + \frac{1}{2} \int_0^t f''(X_s) B_s^2 ds \quad (4.7)$$

**Poznámka 4.4** *Dvě analogie Itôovy formule:*

1. *Itôova formule je jakousi analogií Taylorova vzorce pro aproximaci funkce  $f$  v bodě  $a$  polynomem daný vztahem  $f(x) = f(a) + f'(a)(x - a) + f''(a)(x - a)^2/2 + o(x^2)$ .*
2. *Výraz " $ds$ " v posledním členu (4.7) lze heuristicky odvodnit vlastností Wienerova procesu uvedenou ve druhém bodě příkladu 4.2.*

### 4.5.5 Stochastická diferenciální rovnice

Stochastická diferenciální rovnice je rovnice popisující přírůstky procesu v čase. Ve finanční matematice se velice často používá pro popis přírůstku hodnoty nějakého aktiva.

Stochastických diferenciálních rovnic je mnoho typů. V této kapitole si však pro ukázkou práce s těmito rovnicemi popíšeme jen tu základní s driftem závislejícím pouze na aktuální hodnotě a konstantní volatilitou.

Uvažujme rovnici

$$dX_t = \mu X_t dt + \sigma dW_t \quad \text{s počáteční podmínkou } X_0 = x_0. \quad (4.8)$$

Budeme předpokládat, že řešení je ve tvaru

$$X_t = a(t)[X_0 + \int_0^t b(s)dW_s], \quad a(0) = 1,$$

kde  $a$  a  $b$  jsou reálné funkce. Odsud

$$dX_t = a(t)b(t)dW_t + a'(t)[X_0 + \int_0^t b(s)dW_s]dt$$

a porovnáním koeficientů u  $dW_t$ , resp.  $dt$ , se zadáním dostáváme

$$a(t)b(t) = \sigma \quad (4.9)$$

$$a'(t)[X_0 + \int_0^t b(s)dW_s] = \mu a(t)[X_0 + \int_0^t b(s)dW_s]. \quad (4.10)$$

Vykrácením (4.10) dostáváme rovnici

$$\frac{a'(t)}{a(t)} = \mu,$$

jejímž řešením je

$$a(t) = a(0) \exp\{\mu t\} = \exp\{\mu t\}.$$

Dosazením do (4.9) dostáváme

$$b(t) = \sigma \exp\{-\mu t\},$$

takže výsledným procesem je

$$X_t = \exp\{\mu t\} \left[ X_0 + \sigma \int_0^t \exp\{-\mu s\} dW_s \right].$$

# Kapitola 5

## Pojistná matematika

Pojistná matematika je důležitým nástrojem pro práci **pojišťovny**, jejíž úlohou je na základě **pojistné smlouvy** vyplatit v případě **pojistné události** tzv. **pojistné plnění**. Jelikož je pojistná událost náhodná, využívá se v pojistné matematice převážně pravděpodobnosti a matematické statistiky, s nimiž jsme se seznámili v předešlých kapitolách.

V praxi jsou od sebe odděleny dvě základní oblasti pojištění, a to:

- **životní pojištění**, do něhož spadá výplata předem sjednané částky v případě smrti nebo dožití se určitého věku;
- **neživotní pojištění**, do něhož spadají ostatní události, jejichž společným rysem je, že vyplácená částka - náhrada škody, která v souvislosti s touto událostí vznikla - není předem známa.

### 5.1 Neživotní pojištění

Dva nejdůležitější úkoly pojišťovny jsou stanovit výši ceny za pojištění, tzv. **pojistné**, a stavovit si tzv. **technickou rezervu**, tj. částku, kterou musí mít k dispozici na události, které jsou nahlášeny se zpožděním.

#### 5.1.1 Výpočet pojistného

Je zřejmé, že hlavním údajem pro výpočet pojistného je souhrná výše škod, za něž musí být vyplaceno pojistné plnění. Tato hodnota je však náhodná, proto ji budeme uvažovat

jako náhodnou veličinu  $S$ . Pravděpodobnostními a statistickými metodami odhadneme rozdělení této náhodné veličiny včetně jejích parametrů a základnem pro výpočet pojistného pak bude střední hodnota této náhodné veličiny  $\mathbb{E}S$ . Ta se nazývá **ryzí** (nebo také **netto**) **pojistné**.

Jelikož se ale musí pojišťovna jistit proti nepříznivému průběhu a také má správné náklady, je pojistné, které je nakonec předepsáno pojištěnci, navíc navýšeno o tzv. **bezpečnostní přírůžku**. Toto navýšené pojistné se nazývá **brutto pojistným**.

Nejběžnější způsoby stanovení brutto pojistného  $BP$  (pro všechny pojištěnce dohromady) jsou:

1. princip střední hodnoty:  $BP = (1 + a)\mathbb{E}S$ , kde  $a > 0$ ;
2. princip směrodatné odchylky:  $BP = \mathbb{E}S + a\sqrt{\text{var } S}$ , kde  $a > 0$ ;
3. princip rozptylu:  $BP = \mathbb{E}S + a\text{var } S$ , kde  $a > 0$ .

Hodnoty  $a\mathbb{E}S$ , resp.  $a\sqrt{\text{var } S}$ , resp.  $a\text{var } S$ , jsou zmiňovanými bezpečnostními přírůžkami. Druhá a třetí metoda mají nevýhodu, že je nutné počítat kromě střední hodnoty navíc rozptyl, avšak jsou zase přesnější, neboť berou v úvahu i velikost fluktuací rizika.

### 5.1.2 Modelování celkové výše škod

Souhrn pojistných smluv daného typu pojištění se nazývá **pojistný kmen** nebo také **pojistné portfolio**. Předpokládejme, že pojistný kmen je **homogenní**, tzn. že škody, které mohou nastat na jednotlivých smlouvách, jsou nezávislé stejně rozdělené náhodné veličiny  $X_i$ . Počet škodních událostí je pak také náhodná veličina  $N$ . Celkový úhrn škod je tudíž náhodná veličina

$$S = \sum_{i=1}^N X_i.$$

Jelikož je tato náhodná veličina daná součtem náhodného počtu náhodných veličin, říkáme, že má **složené rozdělení**.

Pro složená rozdělení

$$S = \sum_{i=1}^N X_i.$$

platí

$$\mathbb{E}S = \mathbb{E}N\mathbb{E}X_1$$

a

$$\text{var } S = \mathbb{E}N \text{var } X_1 + \text{var } N (\mathbb{E}X_1)^2.$$

## Rozdělení výší jednotlivých škod

Rozdělení náhodných veličin  $X_i$  musí být samozřejmě nezáporné a také se vyžaduje, aby bylo spojitě. Dalším přirozeným požadavkem pak je, aby pravděpodobnost extrémně velkých hodnot byla minimální (na automobilu zřídka vznikne škoda převyšující 1 mil. korun).

Příkladem takového rozdělení je rozdělení exponenciální, se kterým jsme se již setkali v dřívější kapitole a které se k modelování výší škod používá.

Dalšími používanými rozděleními jsou např.:

### 1.) Weibullovo rozdělení

s distribuční funkcí

$$F(x) = 1 - e^{-\alpha x^k}, \quad x \geq 0, k > 0, \alpha > 0$$

a hustotou

$$f(x) = \alpha k x^{k-1} e^{-\alpha x^k},$$

které je  $k$ -tou odmocninou exponenciálního rozdělení  $Exp(\alpha)$ .

### 2.) Paretovo (také logaritmicko-exponenciální) rozdělení

s distribuční funkcí

$$F(x) = 1 - \left(\frac{x}{a}\right)^{-\alpha}, \quad x \geq a, \alpha > 0, a > 0$$

a hustotou

$$f(x) = \alpha a^\alpha x^{-\alpha-1},$$

které vzniklo transformací

$$X = ae^Y,$$

kde  $Y$  má exponenciální rozdělení  $Exp(\alpha)$ .

### 3.) Logaritmicko-normální rozdělení

s hustotou

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-(\log x - \mu)^2 / 2\sigma^2}, \quad x > 0,$$

které vzniklo transformací

$$X = e^Y,$$

kde  $Y$  má normální rozdělení  $N(\mu, \sigma^2)$ .



## Rozdělení počtu škod

Předpokládejme, že portfolio obsahuje  $n$  smluv a že pravděpodobnost škodní události na jedné smlouvě je  $p$ . Pak má počet škod binomické rozdělení  $Bi(n, p)$ . Jelikož ale rozsah pojistného kmene bývá hodně velký a pravděpodobnost škodní události hodně malá, platí za předpokladu  $np = \lambda$  vztah

$$P(N = k) = \frac{n(n-1)\dots(n-k+1)}{k!} p^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \xrightarrow{n \rightarrow \infty, p \rightarrow 0} \frac{\lambda^k}{k!} e^{-\lambda},$$

což je vztah pro Poissonovo rozdělení, s nímž se lépe počítá, a tudíž je to rozdělení používané pro modelování počtů škod.

### 5.1.3 Technické rezervy

Technické rezervy slouží k zabezpečení prostředků potřebných k úhradě závazků pojišťovny v následujících obdobích. Těchto rezerv je několik druhů, např. **vyrovnávací rezerva** sloužící k vyrovnávání výkyvů v nákladech na pojistná plnění způsobená nepříznivými vlivy, **rezerva na nezasloužené pojistné** související s prováděním účetnictví na konci roku, tj. v době, kdy je ještě smlouva platná a tudíž na ní ještě může vzniknout pojistná událost, dále **rezerva na prémie a slevy** atd.

My se zde však budeme zabývat pouze tou nejdůležitější rezervou, a to **rezervou na pojistná plnění** (nebo též **škodní rezervou**), která udržuje prostředky k výplatě pojistného plnění pojistných událostí, které jsou nahlášeny v pozdějším období než se staly.

#### Rezervy na pojistná plnění - trojúhelníková schémata

Označme  $X_{j,s}$  celkovou výši škod, které vznikly v roce  $j$  a byly uhrazeny do konce roku  $j + s$  (tj.  $s$  hraje roli zpoždění). Předpokládejme, že jsme v roce  $t$ . Data, která máme k dispozici, můžeme seřadit následovně:

	0	1	...	$s$	...	$t-2$	$t-1$
1	$X_{1,0}$	$X_{1,1}$	...	$X_{1,s}$	...	$X_{1,t-2}$	$X_{1,t-1}$
2	$X_{2,0}$	$X_{2,1}$	...	$X_{2,s}$	...	$X_{2,t-2}$	
$\vdots$							
$t-1$	$X_{t-1,0}$	$X_{t-1,1}$					
$t$	$X_{t,0}$						

Tomuto řazení říkáme trojúhelníkové schéma nebo také kumulativní trojúhelník.

**Poznámka 5.1** *Někdy se místo škod, které vznikly v roce  $j$  a byly uraženy do konce roku  $j + s$ , pracuje s hodnotami  $Y_{j,s}$  škod, které vznikly v roce  $j$  a byly uraženy právě v roce  $j + s$ . Pak mluvíme o nekumulativním trojúhelníku.*

Cílem je nalézt hodnotu  $\hat{X}_{j,\infty}$ , která je odhadem celkové výše škod vzniklých v roce  $j$ . Rezervou na pojistná plnění je pak hodnota  $\hat{X}_{j,\infty} - X_{j,t-j}$ .

**Poznámka 5.2** *Samozřejmě se předpokládá, že po nějakém konečném počtu let jsou již všechna pojistná plnění pro daný rok vyplacena. Za tuto dobu je považován právě čas  $t$ , proto metody odhadu  $\hat{X}_{j,\infty}$  spočívají v doplnění kumulativního trojúhelníku na čtverec.*

### Metoda chain-ladder

Tato metoda předpokládá, že sloupce jsou si úměrné, tj. že

$$X_{j,s+1} \doteq c_s X_{j,s}, \quad s = 0, \dots, t-2, \quad j = 1, \dots, t-s-1.$$

Odhadem parametru  $c_s$  je hodnota

$$\hat{c}_s = \frac{\sum_{j=1}^{t-s-1} X_{j,s+1}}{\sum_{j=1}^{t-s-1} X_{j,s}}.$$

Trojúhelník na čtverec pak tedy doplníme pomocí vztahu

$$\hat{X}_{j,r} = X_{j,t-j} \hat{c}_{t-j} \cdots \hat{c}_{r-1}$$

a pro odhad konečné celkové výše plnění tak dostáváme

$$\hat{X}_{j,\infty} = \hat{X}_{j,t-1}$$

a výše rezervy je tudíž

$$\hat{X}_{j,t-1} - X_{j,t-j}.$$

### Zobecnění metody chain-ladder

Předpokládejme, že tzv. **vývojové faktory**

$$d_{j,s} = \frac{X_{j,s+1}}{X_{j,s}}, \quad s = 0, \dots, t-2, \quad j = 1, \dots, t-1,$$

závisí na řádkovém indexu  $j$ . Pro lepší představivost můžeme tyto faktory vypočtené ze známých dat seřadit zase do trojúhelníku

	0	1	...	$s$	...	$t-2$
1	$d_{1,0}$	$d_{1,1}$	...	$d_{1,s}$	...	$d_{1,t-2}$
2	$d_{2,0}$	$d_{2,1}$	...	$d_{2,s}$	...	
$\vdots$						
$t-1$	$d_{t-1,0}$					

a následně počítáme

$$\hat{d}_s = \frac{\sum_{j=1}^{t-s-1} \omega_{j,s} d_{j,s}}{\sum_{j=1}^{t-s-1} \omega_{j,s}}, \quad s = 0, \dots, t-2,$$

kde  $\omega_{j,s}$  jsou váhy jednotlivých faktorů  $d_{j,s}$ . Větší váhy se dávají novějším hodnotám.

Trojúhelník na čtverec pak opět doplníme pomocí vztahu

$$\hat{X}_{j,r} = X_{j,t-j} \hat{d}_{t-j} \cdots \hat{d}_{r-1}.$$

**Poznámka 5.3** *Pokud by byly splněny předpoklady klasické metody chain-ladder, byly by hodnoty faktorů ve sloupcích konstantní. Tato metoda tuto konstantnost nevyžaduje, proto je obecnější.*

*Klasickou metodu chain-ladder získáme, pokud volíme  $\omega_{j,s} = X_{j,s}$ .*

### Londýnský řetězec

Tato metoda stejně jako klasická metoda chain ladder předpokládá, že sloupce na sobě závisí bez ohledu na řádek, tentokrát vztahem

$$X_{j,s+1} \doteq a_s + c_s X_{j,s}, \quad s = 0, \dots, t-2, \quad j = 1, \dots, t-s-1.$$

Parametry  $a_s$  a  $c_s$  se určí tzv. metodou nejmenších čtverců, tj. minimalizací výrazu

$$\sum_{j=1}^{t-s-1} (X_{j,s+1} - a_s - c_s X_{j,s})^2, \quad s = 0, \dots, t-3, \quad (5.1)$$

pro  $s = t-2$  pak volíme  $a_{t-2} = 0$  a  $c_{t-2} = X_{1,t-1}/X_{1,t-2}$ .

Derivací výrazu (5.1) podle parametrů  $a_s$  a  $c_s$  a položením této derivace rovno nule dostáváme soustavu dvou rovnic o dvou neznámých, jejímž řešením je

$$\hat{a}_s = \frac{\sum_{j=1}^{t-s-1} X_{j,s+1} \sum_{j=1}^{t-s-1} X_{j,s}^2 - \sum_{j=1}^{t-s-1} X_{j,s} \sum_{j=1}^{t-s-1} X_{j,s+1} X_{j,s}}{(t-s-1) \sum_{j=1}^{t-s-1} X_{j,s}^2 - (\sum_{j=1}^{t-s-1} X_{j,s})^2} \quad (5.2)$$

$$\hat{c}_s = \frac{(t-s-1) \sum_{j=1}^{t-s-1} X_{j,s+1} X_{j,s} - \sum_{j=1}^{t-s-1} X_{j,s+1} \sum_{j=1}^{t-s-1} X_{j,s}}{(t-s-1) \sum_{j=1}^{t-s-1} X_{j,s}^2 - (\sum_{j=1}^{t-s-1} X_{j,s})^2}. \quad (5.3)$$

Na čtverec pak doplňujeme postupně počítáním

$$\hat{X}_{j,s+1} = \hat{a}_s + \hat{c}_s \hat{X}_{j,s}, \quad s = t-j, \dots, t-2, \quad j = 2, \dots, t,$$

kde  $\hat{X}_{j,t-j} = X_{j,t-j}$  je známá hodnota na diagonále.

## 5.2 Životní pojištění

Jak již bylo zmíněno výše, životní pojištění se zabývá výplatou předem sjednané částky nebo pravidelných splátek v případě úmrtí nebo dožití se nějakého věku, čímž se liší od pojištění neživotního. Některé prvky jsou však stejné.

### Společné prvky životního a neživotního pojištění

- Výše pojistného plnění je náhodná, modeluje se pomocí náhodné veličiny  $Z$ .
- Výše netto pojistného se tedy počítá jako  $NP = \mathbb{E}Z$ .
- Brutto pojistné je netto pojistné navýšené o bezpečnostní přírážku.
- V životním pojištění rovněž platí povinnost tvorby rezerv, způsob výpočtu je však odlišný (a nebude předmětem této přednášky).

### Odlišné prvky životního a neživotního pojištění

- Jelikož životní pojištění se uzavírá na delší dobu (řádově roky až desetiletí), je třeba vzít v úvahu ztrátu hodnoty peněz. Ta se provádí tak, že se do výpočtu zavede tzv. diskontní faktor

$$v = \frac{1}{1+i},$$

kde  $i$  je technická úroková míra, jehož interpretace je taková, že hodnota 1 koruny po  $k$  letech je pouze  $v^k$ .

- K výpočtu  $\mathbb{E}Z$  se nevyužívá známých rozdělení, nýbrž pravděpodobností úmrtí v daném věku, resp. dožití se daného věku, které lze vyčíst z úmrtnostních tabulek.
- Pojistné se většinou platí nikoliv jednorázově, nýbrž na splátky po dobu několika let. Tímto rozdělením splátek se však nebudeme zabývat a pojistné, které budeme počítat, tj.  $\mathbb{E}Z$ , budeme nazývat jednorázovým netto pojistným.

#### 5.2.1 Modelování úmrtnosti

Označme  $T_0$  náhodnou veličinu popisující délku života právě narozeného jedince a obecněji pak  $T_x$  náhodnou veličinu popisující zbývající délku života jedince ve věku  $x$ . Kromě již

známé distribuční funkce

$$F_x(t) = P(T_x \leq t)$$

se v životním pojištění pracuje s tzv. funkcí přežití

$$S_x(t) = P(T_x > t) = 1 - F_x(t).$$

Hodnoty těchto funkcí jsou pro celočíselné hodnoty  $x$  a  $t$  uváděny v úmrtnostních tabulkách, v nichž se používá následujících zjednodušených symbolů:

•

$$q_x = F_x(1) = P(T_x \leq 1)$$

pro pravděpodobnost, že jedinec, který je naživu ve věku  $x$ , zemře před dosažením věku  $x + 1$ ;

•

$$p_x = S_x(1) = P(T_x > 1)$$

pro pravděpodobnost, že jedinec, který je naživu ve věku  $x$ , se dožije věku  $x + 1$ ;

•

$${}_tq_x = F_x(t) = P(T_x \leq t)$$

pro pravděpodobnost, že jedinec, který je naživu ve věku  $x$ , zemře před dosažením věku  $x + t$ ;

•

$${}_tp_x = S_x(t) = P(T_x > t)$$

pro pravděpodobnost, že jedinec, který je naživu ve věku  $x$ , se dožije věku  $x + t$ .

V životním pojištění se pak využívá dvou významných vztahů mezi těmito pravděpodobnostmi, a to:

$$P(T_x > k) = {}_k p_x = p_x \cdot p_{x+1} \cdot p_{x+k-1} \quad (5.4)$$

a

$$P(k \leq T_x < k + 1) = {}_{k+1} q_x - {}_k q_x = {}_k p_x \cdot q_{x+k}. \quad (5.5)$$

Hodnoty  ${}_k p_x$  a  ${}_k q_x$  se získají jednoduchým způsobem. Označme v nějaké populaci  $l_0$  počet nově narozených jedinců a  $l_x$  počet jedinců, kteří se dožili věku  $x$ . Pak

$${}_k p_x = \frac{l_{x+k}}{l_x}$$

a

$${}_kq_x = \frac{l_x - l_{x+k}}{l_x}.$$

**Poznámka 5.4** Zatímco výpočet  ${}_kp_x$  a  ${}_kq_x$  je velice jednoduchý, při volbě populace, z níž hodnoty odhadujeme, je třeba brát v úvahu spoustu vlivů jako např. změnu způsobu života, války apod. Tímto problémem se zabývá sociologie a demografie.

Dalším užitečným značením je  $d_x = l_x - l_{x+1}$  počet lidí, kteří zemřeli ve věku  $x$ . Toho se využívá zejména pro výpočet pravděpodobnosti, že pojištěný ve věku  $x$  zemře v  $(k + 1)$ -ním roce pojištění, která se počítá jako

$${}_kp_x \cdot q_{x+k} = \frac{d_{x+k}}{l_x}.$$

## 5.2.2 Komutační čísla

**Příklad 5.1** Jaké je (jednorázové) netto pojistné pro pojištění, které sjedná 40-letý muž, kde pojišťovna vyplatí 1 mil. Kč, pokud pojištěný do 5 let zemře, a pokud nezemře, pojištění zanikne bez náhrady.

Řešení: Zemře-li pojištěný v  $k$ -tém roce pojištění, dotane 1 mil. Kč. Vezmeme-li v úvahu ztrátu hodnoty peněz, má částka, kterou dostane, současnou hodnotu  $10^6 \cdot v^k$ . Pravděpodobnost, že pojištěný v  $k$ -tém roce pojištění zemře, je  ${}_kp_{40} \cdot q_{40+k}$ . Střední (současná) hodnota toho, co musí pojišťovna vyplatit, je tudíž

$$10^6 \cdot \left( \sum_{k=0}^4 v^{k+1} \cdot {}_k p_{40} \cdot q_{40+k} \right) = 10^6 \cdot \frac{d_{40}v + d_{41}v^2 + \dots + d_{44}v^5}{l_{40}}.$$

V tomto příkladě se počítá pouze s pětiletým pojištěním. To však může být (a v praxi většinou bývá) dlouhodobější, takže suma má mnoho členů a každý z nich je navíc diskontován příslušnou mocninou diskontního faktoru. Aby byl tento výpočet jednodušší a přehlednější, zavádí se tzv. komutační čísla, a to:

- komutační čísla multého řádu:

$$\begin{aligned} D_x &= l_x v^x && \text{(diskontovaný počet dožívajících se věku } x) \\ C_x &= d_x v^{x+1} && \text{(diskontovaný počet zemřelých ve věku } x) \end{aligned}$$

- komutační čísla prvního řádu:

$$N_x = \sum_{j=0}^{\infty} D_{x+j} = D_x + D_{x+1} + D_{x+2} + \dots$$

$$M_x = \sum_{j=0}^{\infty} C_{x+j} = C_x + C_{x+1} + C_{x+2} + \dots$$

- komutační čísla druhého řádu:

$$S_x = \sum_{j=0}^{\infty} N_{x+j} = N_x + N_{x+1} + N_{x+2} + \dots$$

$$R_x = \sum_{j=0}^{\infty} M_{x+j} = M_x + M_{x+1} + M_{x+2} + \dots$$

Vrátíme-li se nyní k našemu příkladu, zjistíme, že se nám výpočet zjednoduší do tvaru

$$10^6 \cdot \frac{d_{40}v + d_{41}v^2 + \dots + d_{44}v^5}{l_{40}} = 10^6 \cdot \frac{d_{40}v^{41} + d_{41}v^{42} + \dots + d_{44}v^{45}}{l_{40}v^{40}} =$$

$$10^6 \cdot \frac{C_{40} + C_{41} + \dots + C_{44}}{D_{40}} = 10^6 \cdot \frac{M_{40} - M_{45}}{D_{40}},$$

přičemž hodnoty  $M_{40}$ ,  $M_{45}$  a  $D_{40}$  najdeme v úmrtnostních tabulkách.

### 5.2.3 Druhy pojištění

Základním dělením životního pojištění je dělení na

- kapitálové pojištění - jednorázová výplata částky v případě úmrtí nebo dožití se daného věku
- důchodové pojištění - pravidelné výplaty částek v případě dožití se daného věku

Oba tyto druhy pak mají spoustu typů, z nichž si zde uvedeme ty nejběžnější.

**Poznámka 5.5** Jelikož pro střední hodnotu platí  $\mathbb{E}(aZ) = a\mathbb{E}Z$ , budeme vždy, pokud nebude řečeno jinak, počítat jednorázové netto pojistné (JNP) pro výplatu jednotkové částky. Pokud by pojištění bylo sjednáno na částku  $c$ , bylo by výsledné JNP obyčejným  $c$ -násobkem námi vypočteného JNP.



## Kapitálová životní pojištění

### Pojištění pro případ dožití

spočívá ve výplatě předem sjednané částky na konci roku  $n$ , pokud se osoba pojištěná ve věku  $x$  dožije věku  $x + n$ , jinak pojištění zaniká bez náhrady. Pro (jednorázové) netto pojistné platí

$$JNP = \mathbb{E}Z = {}_n p_x \cdot v^n = \frac{D_{x+n}}{D_x}.$$

### Pojištění pro případ smrti

spočívá ve výplatě předem sjednané částky na konci roku, v němž osoba pojištěná ve věku  $x$  zemře, jinak pojištění zaniká bez náhrady. Pro (jednorázové) netto pojistné platí

$$JNP = \mathbb{E}Z = \sum_{k=0}^{\infty} {}_k p_x \cdot q_{x+k} \cdot v^{k+1} = \frac{M_x}{D_x}.$$

### Dočasné pojištění pro případ smrti

spočívá ve výplatě předem sjednané částky na konci roku, v němž osoba pojištěná ve věku  $x$  zemře, pokud k tomuto úmrtí dojde během  $n$  let, jinak pojištění zaniká bez náhrady. Pro (jednorázové) netto pojistné platí

$$JNP = \mathbb{E}Z = \sum_{k=0}^{n-1} {}_k p_x \cdot q_{x+k} \cdot v^{k+1} = \frac{M_x - M_{x+n}}{D_x}.$$

### Směšené pojištění

spočívá ve výplatě předem sjednané částky  $a$  na konci roku, v němž osoba pojištěná ve věku  $x$  zemře, pokud k tomuto úmrtí dojde během  $n$  let, jinak vyplatí částku  $b$ . Pro (jednorázové) netto pojistné platí

$$JNP = \mathbb{E}Z = a \cdot \sum_{k=0}^{n-1} {}_k p_x \cdot q_{x+k} \cdot v^{k+1} + b \cdot {}_n p_x \cdot v^n = \frac{a(M_x - M_{x+n}) + bD_{x+n}}{D_x}.$$

## Důchodová životní pojištění

### Pojištění doživotního důchodu

spočívá v pravidelné výplatě předem sjednaných částek vždy na začátku roku, pokud osoba pojištěná ve věku  $x$  žije. Pro (jednorázové) netto pojistné platí

$$JNP = \mathbb{E}Z = \sum_{k=0}^{\infty} \cdot_k p_x \cdot v^k = \frac{N_x}{D_x}.$$

### Pojištění odloženého doživotního důchodu

spočívá v pravidelné výplatě předem sjednaných částek vždy na začátku roku, pokud osoba pojištěná ve věku  $x$  žije, avšak tyto výplaty začnou až po  $j$  letech od uzavření tohoto pojištění. Pro (jednorázové) netto pojistné platí

$$JNP = \mathbb{E}Z = \sum_{k=j}^{\infty} \cdot_k p_x \cdot v^k = \frac{N_{x+j}}{D_x}.$$

### Pojištění dočasného doživotního důchodu

spočívá v pravidelné výplatě předem sjednaných částek vždy na začátku roku, pokud osoba pojištěná ve věku  $x$  žije a neuplynulo ještě  $n$  let od začátku pojištění. Pro (jednorázové) netto pojistné platí

$$JNP = \mathbb{E}Z = \sum_{k=0}^{n-1} \cdot_k p_x \cdot v^k = \frac{N_x - N_{x+n}}{D_x}.$$

# Kapitola 6

## Shluková analýza dat

Základním cílem shlukové analýzy je zařadit objekty z nějakého souboru objektů do skupin (shluků) tak, aby si dva objekty v jedné skupině byly podobnější než dva objekty z různých skupin. Metod pro toto zařazení je spousta, stejně tak struktur shluků může být více (kromě rozdělení do několika skupin můžeme řadit do vzájemně vnořených podskupin). Popis některých z nich je předmětem této kapitoly.

### 6.1 Vstupní data

Soubor objektů, které dostaneme, je tvořen  $n$  prvky (objekty), které se mají shlukovat. U každého z nich pak pozorujeme  $m$  různých znaků (proměnných). To znamená, že vstupní údaje můžeme seřadit do matice rozměru  $n \times m$ . Její prvky pak budeme značit  $x_{il}$ ,  $i = 1, \dots, n$ ,  $l = 1, \dots, m$ .

#### 6.1.1 Typy proměnných

Znaky, které pozorujeme, mohou být různě porovnatelné. Podle toho rozlišujeme proměnné

- poměrové - u jejich hodnot můžeme určit, o kolik i kolikrát je jedna hodnota větší než druhá (např. věk, cena, ...),
- intervalové - u jejich hodnot můžeme určit, o kolik, ne však už kolikrát, je jedna hodnota větší než druhá (např. teplota, ...),

- ordinální - u jejich hodnot můžeme určit pořadí hodnot (např. základní, střední a vysoká škola, ...),
- nominální - u jejich hodnot můžeme určit pouze, zda jsou stejné nebo různé (např. barva očí).

Zatímco u prvních dvou typů můžeme pracovat přímo s hodnotami těchto proměnných, zbylé dvě proměnné musíme ohodnotit. Ordinálním proměnným můžeme třeba přiřadit hodnoty jejich pořadí, popř. přidat větší váhy nějakým hodnotám. Pro nominální proměnné lze pak použít metodu rozdělení proměnné na více binárních (tj. nabývajících hodnot 0 nebo 1) proměnných, kde 1 znamená, že objekt splňuje danou vlastnost, 0 opak. Např. příslušnost k univerzitě (ČVUT, UK nebo VŠE) lze zapsat takto:

Univerzita	$X_1$	$X_2$	$X_3$
ČVUT	1	0	0
UK	0	1	0
VŠE	0	0	1

**Poznámka 6.1** Stejným způsobem se dají na binární proměnné převést i proměnné poměrové a intervalové, kde hodnota 1 bude značit příslušnost do nějakého intervalu hodnot. Proměnné ordinální lze pak pomocí binárních proměnných zapsat např. takto:

Vzdělání	$X_1$	$X_2$
základní	0	0
střední	1	0
vysokoškolské	1	1

## 6.1.2 Normování hodnot proměnných

Jelikož chceme porovnávat objekty na základě naměřených znaků, je většinou třeba, aby neměly různě velké a rozptýlené hodnoty (např. věk a počet dětí). Proto je třeba hodnoty vhodně vynormovat.

Jedním ze způsobů je např. převedení všech proměnných na proměnné binární (viz výše). Tím bychom ale dostali do vstupní matice zbytečně moc proměnných, proto je výhodnější použít jiných metod normování, např.

- vydělení směrodatnou odchylkou proměnné  $l$ :

$$z_{il} = \frac{x_{il}}{s_l},$$

- vydělení variančním rozpětím proměnné  $l$ , tj. hodnotou  $R_l = \max_i(x_{il}) - \min_i(x_{il})$ :

$$z_{il} = \frac{x_{il}}{R_l},$$

- převedení na hodnoty z intervalu  $\langle 0, 1 \rangle$ :

$$z_{il} = \frac{x_{il} - \min_i(x_{il})}{R_l},$$

- převedení na hodnoty z intervalu  $\langle 0, 1 \rangle$ , jejichž součet je roven 1:

$$z_{il} = \frac{x_{il}}{\sum_{i=1}^n x_{il}}.$$

## 6.2 Měření podobnosti

Na základě předchozího lze každý objekt  $i$  vyjádřit jako číselný vektor  $\mathbf{x}_i$  o složkách  $x_{il}, l = 1, \dots, m$  (popř. pracovat s jeho normovanými hodnotami  $\mathbf{z}_i$  o složkách  $z_{il}, l = 1, \dots, m$ ). Dva objekty pak můžeme považovat za podobnější než jiné dva, pokud jsou si v  $m$ -dimenzionálním prostoru blíží. K tomu, abychom mohli posuzovat, zda jsou si prvky blíží, potřebujeme definovat vzdálenost. Obecně se vzdáleností myslí funkce  $D_{ij}$  dvou prvků  $i$  a  $j$ , která splňuje následující:

1.  $D_{ij} \geq 0$ ,
2.  $D_{ii} = 0$ ,
3.  $D_{ij} = D_{ji}$

**Poznámka 6.2** *Občas se vyžaduje ještě čtvrtá vlastnost  $D_{ij} + D_{jk} \geq D_{ik}$ . Pak se mluví místo o vzdálenosti o metrice.*

Nejčastěji používanými vzdálenostmi jsou:

- eukleidovská:

$$D_{ij} = D(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{l=1}^m (x_{il} - x_{jl})^2},$$

- městských bloků (manhattanská):

$$D_{ij} = D(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l=1}^m |x_{il} - x_{jl}|,$$

- maximová (Čebyševova):

$$D_{ij} = D(\mathbf{x}_i, \mathbf{x}_j) = \max_l |x_{il} - x_{jl}|.$$

## 6.3 Metody shlukové analýzy

Většinou se v literatuře uvádí dělení těchto metod na dvě základní skupiny podle toho, co má být výsledkem shlukování, a to:

- metody rozkladu (nehierarchické) - výsledkem je rozdělení souboru do  $k$  shluků, kde počet shluků je předem daný,
- metody hierarchické - výsledkem je posloupnost do sebe vnořených skupin objektů.

### 6.3.1 Metody rozkladu

Metody rozkladu lze dále rozdělit, a to na

- metody jednoznačného přiřazení - výsledkem je jednoznačná příslušnost každého objektu do nějakého shluku,
- fuzzy shluková analýza - výsledkem jsou míry příslušnosti  $u_{ip}$  každého objektu  $i$  do  $p$ -tého shluku, pro které platí

1.  $0 \leq u_{ip} \leq 1$ ,
2.  $\sum_{p=1}^k u_{ip} = 1$ .

V dalším textu se budeme zabývat už jen metodami jednoznačného přiřazení, které se v praxi vyskytují častěji.

## Metoda $k$ -průměrů

Na začátku se vybere  $k$  počátečních centroidů (např. prvních  $k$  objektů v souboru). Pro každý prvek souboru se spočte jeho vzdálenost k jednotlivým centroidům a prvek se přiřadí do shluku k centroidu, ke kterému má nejbliž. Po přiřazení všech prvků se spočte nový centroid shluku (např. bod v prostoru, jehož souřadnicemi jsou průměry hodnot jednotlivých proměnných) a celá procedura se opakuje. Končí se ve chvíli, kdy už se žádný prvek během celé procedury nikam nepřesune.

## Metoda $k$ -medoidů

Jedná se o metodu podobnou metodě  $k$ -průměrů s tím rozdílem, že místo centroidu, což může být libovolný bod v prostoru, se prvky přiřazují medoidu, což je konkrétní objekt ze shluku. Ten se určí tak, aby součet vzdáleností od tohoto objektu byl minimální.

### 6.3.2 Hierarchické metody

Hierarchické metody lze stejně jako nehierarchické metody dále dělit, a to buď podle toho, zda shlukujeme podle jedné či podle více proměnných, na metody

- monotetické - shluky se vytvářejí postupně podle jednotlivých proměnných
- polytetické - v každém kroku jsou uvažovány všechny proměnné najednou

nebo podle toho, zda shluky postupně rozkládáme nebo slučujeme, na metody

- aglomerativní - na počátku je každý objekt samostatným shlukem a postupně dochází ke spojování shluků
- divizivní - na počátku je celý soubor jedním shlukem a postupně dochází k dělení shluků

#### Monotetické shlukování

Monotetické shlukování je výhodnější pro divizivní přístup.

Je zde potřeba, aby všechny proměnné byly binární, neboť princip tohoto shlukování je v postupném dělení shluků na dva další shluky podle toho, zda nabývají pro danou proměnnou hodnoty 0 nebo 1. Máme-li přitom  $m$  proměnných, je možností výběru první proměnné  $m$ , druhé  $m - 1$  atd., čímž ale dostáváme nejednoznačnost rozkladu. Existuje však kritérium výběru proměnné, podle které v daném kroku dělíme, a tou je intenzita závislosti. Uvažujme dvě proměnné, a to  $k$ -tou a  $l$ -tou, a k nim kontingenční tabulku jejich výskytů

$k \setminus l$	0	1
0	$a_{kl}$	$b_{kl}$
1	$c_{kl}$	$d_{kl}$

Pro každou dvojici se spočte koeficient

$$q_{kl} = |a_{kl}d_{kl} - b_{kl}c_{kl}|$$

a za proměnnou, podle které budeme shluky dělit, je proměnná s nejvyšší hodnotou

$$q_l = \sum_{k \neq l} q_{kl}, \quad k = 1, 2, \dots, m.$$

## Polytetické shlukování

Polytetické shlukování se využívá zase spíše pro aglomerativní přístup. Spočívá v tom, že na počátku je každý objekt samostatným shlukem a v každém kroku se vyberou dva shluky, které jsou si nejpodobnější, a ty se spojí.

Podobnost objektů už byla zmíněná výše, zbývá tedy zmínit, jak určit podobnost shluků. Uvědomme si, že k tomu stačí určit vzdálenost mezi  $g$ -tým shlukem a sjednocením shluků  $h$  a  $h'$ . Shluková analýza nám v tomto případě opět nabízí spoustu možností měření této podobnosti, uveďme si z nich např.

- metoda průměrné vazby:

$$D_{g<h,h'>} = \frac{n_h}{n_h + n_{h'}} D_{gh} + \frac{n_{h'}}{n_h + n_{h'}} D_{gh'},$$

kde  $n_h$  a  $n_{h'}$  jsou počty prvků ve shlucích  $h$ , resp.  $h'$ ,

- mediánová metoda:

$$D_{g<h,h'>} = \frac{1}{2} D_{gh} + \frac{1}{2} D_{gh'} - \frac{1}{4} D_{hh'},$$



- metoda nejbližšího souseda:

$$D_{g\langle h,h'\rangle} = \frac{1}{2}(D_{gh} + D_{gh'} - |D_{gh} - D_{gh'}|),$$

- atd.