

# Pravděpodobnost a Statistika 1

Poznámky z přednášek  
p. doc. Roberta Šámala

Letní semestr 2020/2021

**Viktor Soukup, Lukáš Salak**

# Obsah

<b>1 První přednáška</b>	<b>3</b>
1.1 Úvodem . . . . .	3
1.2 Základní definice . . . . .	4
<b>2 Druhá přednáška</b>	<b>6</b>
2.1 Opakování . . . . .	6
2.2 Podmíněná pravděpodobnost . . . . .	7
<b>3 Třetí přednáška</b>	<b>9</b>
3.1 Typy rozdělení . . . . .	10
3.2 Rozptyl a LOTUS . . . . .	11
<b>4 Čtvrtá přednáška</b>	<b>12</b>
4.1 Parametry rozdělení . . . . .	14
4.2 Náhodné vektory . . . . .	15
4.3 Marginální rozdělení . . . . .	15
<b>5 Pátá přednáška</b>	<b>16</b>
<b>6 Šestá přednáška</b>	<b>18</b>
<b>7 Sedmá přednáška</b>	<b>20</b>
<b>8 Osmá přednáška</b>	<b>23</b>
<b>9 Devátá přednáška</b>	<b>27</b>
9.1 Nerovnosti, které známe z minula . . . . .	27
9.2 Slabý zákon velkých čísel . . . . .	27
9.3 Centrální limitní věta . . . . .	27
9.4 Momentová vytvořující funkce . . . . .	28
9.5 Statistika . . . . .	28
9.6 Empirická distribuční funkce - Dvoretzky-Keifer-Wolfowitz (DKW) . . . . .	28
9.7 Intro - explorační analýza dat (exploratory data analysis) . . . . .	29
<b>10 Desátá přednáška</b>	<b>29</b>
10.1 náhodný výběr . . . . .	29
10.2 Statistika - model . . . . .	29
10.3 Zkoumané úlohy - cíle konfirmační analýzy . . . . .	29
10.4 Vlastnosti bodových odhadů . . . . .	30
10.5 Metoda momentů . . . . .	31
10.6 Metoda maximální věrohodnosti (maximal likelihood, ML) . . . . .	31
<b>11 Jedenáctá přednáška</b>	<b>32</b>
11.1 Intervalové odhady . . . . .	32
11.2 Testování hypotéz . . . . .	33
<b>12 Dvanáctá přednáška</b>	<b>33</b>
12.1 Testování hypotéz - ilustrace . . . . .	33
12.2 $p$ -hacking . . . . .	35
12.3 $\chi_k^2$ - rozdělení $\chi$ -kvadrát . . . . .	35
12.4 Multinomické a kategoriální rozdělení . . . . .	35
12.5 Test dobré shody (goodness of fit) . . . . .	36
<b>13 Třináctá přednáška</b>	<b>36</b>
13.1 Simpsonův paradox . . . . .	36
13.2 Permutační test . . . . .	37
13.3 Bootstrap . . . . .	37

13.4	Bayesovská statistika . . . . .	38
13.4.1	Frekventistický/klasický přístup . . . . .	38
13.4.2	Bayesovský přístup . . . . .	38
13.5	Generování náhodných veličin . . . . .	39

# 1 První přednáška

## 1.1 Úvodem

Modely náhody  $\rightarrow$  Pravděpodobnost  $\rightarrow$  Pozorovaná data  $\rightarrow$  Modely náhody

Model náhody např. kostka  $1, \dots, 6$ ,

Pozorovaná data :  $1, 5, 4, 3, 3$

otázka na pravděpodobnost: jaká je pravděpodobnost... hodně pozorovaných dat  $\rightarrow$  statistika na model náhody.

**Příklad** (Schwartz-Zippel algoritmus): Máme dány dva polynomy  $f(x), g(x)$  stupně  $d$ . Chceme zjistit, zda jsou stejné, a to co nejrychleji.

Problém:  $g(x)$  je součin několik polynomů stupně  $\leq \frac{d}{4}$ , dostáváme víc než lineární čas.

**Řešení:** Algoritmus: zvolíme náhodně  $x \in \{1, 2, \dots, 100d\}$ , ověříme, zda  $f(x_1) = g(x_1)$ . Když  $f \neq g$ , tak  $x_1$  je kořen polynomu  $f - g$ . ... takových  $x_1$  je  $\leq d$ .

$$P(f(x_1) = g(x_1) : f \neq g) \leq \frac{1}{100}$$

Pokud jsme spokojeni s 1%, končíme, když ne, volíme  $x_2, x_3, \dots \in \{1, 2, \dots, 100d\}$ , pak

$$P(\text{Pro } x_1, x_2, x_3 \dots f(x_i) = g(x_i) : f \neq g) \leq \left(\frac{1}{3}\right)^3 = 10^{-6}$$

... aproximační algoritmy

Některé jevy neumíme/nechceme popsat kauzálně

- hod kostkou
- tři hody kostkou, nekonečně mnoho hodů kostkou
- hod šípkou na terč
- počet emailů za den
- dobu běhu programu (v reálném počítači)
- a další ...

Důvody:

- fyzikální vlastnost přírody
- komplikovaný proces (počasí, medicína, molekuly plynu...)
- neznáme vlivy (působení dalších lidí, programů...)
- randomizované algoritmy (test prvočíselnosti, quicksort)
- náhodné grafy (Ramseyovy čísla)
- a další ...

Pro popis pomocí teorie pravděpodobnosti napřed vybereme množinu elementárních jevů  $\Omega$  (sample space)

$$\Omega = \{1, 2, \dots, 6\} = [6] \implies \text{hod kostkou}$$

$$\Omega = [6]^3 \implies \text{hod třemi kostkami}$$

## 1.2 Základní definice

**Definice** (Prostor jevů):  $\mathcal{F} \subseteq \mathcal{P}(\Omega)$

$\mathbb{F} \subseteq \mathbb{P}(\Omega)$  je prostor jevů (též  $\sigma$ -algebra), pokud

1.  $\emptyset \in \mathcal{F}$  a  $\Omega \in \mathcal{F}$
2.  $A \in \mathcal{F} \implies \Omega \setminus A \in \mathcal{F}$
3.  $A_1, A_2, \dots \in \mathcal{F} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Často  $\mathcal{F} = \mathcal{P}(\Omega)$ , to je možné vždy, když je  $\Omega$  spočetná, např. pro  $\Omega = \mathbb{R}$  to již nejde.

**Definice** (Pravděpodobnost):  $P : \mathcal{F} \rightarrow [0, 1]$  se nazývá pravděpodobnost (probability), pokud:

1.  $P(\emptyset) = 0, P(\Omega) = 1$ , a
2.  $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ , pro libovolnou posloupnost po dvou *disjunktních* jevů **TODOOT**

Šance (odds) jevu A je  $O(A) = \frac{P(A)}{P(A^c)}$ . Např. šance na výhru je 1 ku 2 znamená, že pravděpodobnost výhry je  $\frac{1}{3}$ ; šance, že na kostce padne šestka je 1 ku 5.

**Konvence:**

- „A je jistý jev“ znamená  $P(A) = 1$ . Také se říká, že A nastává skoro jistě (almost surely), zkráceně s.j. (a.s.).
- „A je nemožný jev“ znamená  $P(A) = 0$ .

$$P(A) = 0 \stackrel{?}{\implies} A = \emptyset$$

$\leftarrow$  axiom

$\rightarrow$  platí často, ne vždy

- Např.  $A =$  střed kruhu (házení šipek na terč)  $\implies P(A) = 0$  B spočetná (konečná, velká jako  $\mathbb{N}$ ) množina:

$$P(B) = 0 + 0 + 0 + \dots = 0$$

$B_i$  je  $i$ -tý bod,  $B = \bigcup B_i$

**Věta** (Vlastnosti pravděpodobnostního prostoru): V pravděpodobnostním prostoru  $(\Omega, \mathcal{F}, P)$  platí pro  $A, B \in \mathcal{F}$ :

1.  $P(A) + P(A^c) = 1$
2.  $A \subseteq B \implies P(A) \leq P(B)$
3.  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
4.  $P(A_1 \cup A_2 \cup \dots) \leq \sum_i P(A_i)$  (subaditiva, Booleova nerovnost) (nevyžadujeme disjunktnost, pak by platila rovnost)

**Důkaz:**

1.  $\Omega = A \cup A^c$ ;  $A, A^c$  disj.,  
 $1 = P(\Omega) = P(A) + P(A^c)$
2.  $P(B) = P(A) + P(B \setminus A) \geq P(A)$
3. Využíváme Princip Inkluze a Exkluze. (ale nevím jiste, jestli to staci)
4. trik zdisjunktnění: z  $A_1, A_2 \dots$  uděláme disjunktní množiny

$$B_1 = A_1, B_2 = A_2 \setminus A_1, B_3 = A_3 \setminus A_1 \cup A_2 \dots$$

$$B_i \subseteq A_i \implies P(B_i) \leq P(A_i)$$

$$B_i \cap B_j = \emptyset : j < i \dots B_i \cap B_j \subseteq B_i \cap A_j = \emptyset$$

$$\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i$$

$$\subseteq ok$$

opačná inkluze **TODOOT**

$$P(\bigcup A_i) = P(\bigcup B_i) = \sum P(B_i) \leq \sum P(A_i).$$

□

**Příklad** (Pravděpodobnostní prostory):

1. *Konečný s uniformní pravděpodobností*

$\Omega$  je libovolná konečná množina,  $\mathcal{F} = \mathcal{P}(\Omega)$ ,

$$P(A) = \frac{|A|}{|\Omega|}.$$

2. *Diskrétní*

$\Omega = \{\omega_1, \omega_2, \dots\}$  je libovolná spočetná množina. Jsou dány  $p_1, p_2, \dots \in [0, 1]$  se součtem 1.

$P(A) = \sum_{i: \omega_i \in A} p_i$  (cinknutá loterie, nějaké možnosti mají jiné procenta)

3. *Spojité*

$\Omega \subseteq \mathbb{R}^d$  pro vhodné  $d$  ( $\Omega$  např. uzavřená nebo otevřená)

$\mathcal{F}$  vhodná (obsahuje např. všechny otevřené množiny)

$f: \Omega \rightarrow [0, 1]$  je funkce taková, že  $\int_{\Omega} f(x) dx = 1$ .

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

Speciální případ:  $f(x) = 1/V_d(\Omega)$

$P(A) = \frac{V_d(A)}{V_d(\Omega)}$ , kde  $V_d(A) = \int_A 1$  je  $d$ -rozměrný objem  $A$ .

4. *Bernoulliho krychle - nekonečné opakování*

$\Omega = S^{\mathbb{N}}$ , kde  $S$  je diskrétní s pravděpodobností  $Q$ ,

$\mathcal{F}$  vhodná (obsahuje např. všechny množiny tvaru

$$A = A_1 \times \dots \times A_k \times S \times S \times \dots)$$

$$P(A) = Q(A_1) \dots Q(A_k)$$

### Příklad (Nepřiklady):

1. Náhodné přirozené číslo: můžeme si vybrat mnoha způsoby, Ale všechna přirozená čísla nemají stejnou pravděpodobnost.  
není možné, aby měly všechny stejnou nenulovou pravděpodobnost, protože pokud  $P(0) = P(1) = P(2) \cdots = P$  tak  $P(\mathbb{N}) = p + p + p \cdots = \infty$ .
2. Náhodné reálné číslo
3. Betranův paradox

**Definice** (Podmíněná pravděpodobnost): Pokud  $A, B \in \mathcal{F}$  a  $P(B) > 0$ , pak definujeme podmíněnou pravděpodobnost A při B (probability of A given B) jako

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

$Q(A) := P(A|B)$ . Pak  $(\Omega, \mathcal{F}, Q)$  je pravděpodobnostní prostor.

**Definice** (Zřetězené podmínování):  $P(A \cap B) = P(B)P(A|B)$

**Věta:** Pokud  $A_1, \dots, A_n \in \mathcal{F}$  a  $P(A_1 \cap \dots \cap A_n) > 0$ , tak

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2|A_1) \dots \text{TODOOT}$$

## 2 Druhá přednáška

### 2.1 Opakování

1. definice pravděpodobnostního prostoru  $(\Omega, \mathcal{F}, P)$ : dva axiomy,
2. **naivní** pravděpodobnostní prostor:  $\Omega$  konečná,  $\mathcal{F} = \mathcal{P}(\Omega)$   
 $P(A) := |A|/|\Omega|$
3. **diskrétní** pravděpodobnostní prostor:  $\Omega = \omega_1, \omega_2, \dots$ ,  
 $\mathcal{F} = \mathcal{P}(\Omega)$ ,  $\sum_i p_i = 1$   
 $P(A) := \sum_{i: \omega_i \in A} p_i$
4. **geometrický** pravděpodobnostní prostor:  
 $\omega \subseteq \mathbb{R}^d$  s konečným objemem,  
 $P(A) := V_d(A)/V_d(\Omega)$
5. pravděpodobnostní prostor **spojitý s hustotou**:  
 $\Omega \subseteq \mathbb{R}^d$  s funkcí  $f$ , kde  $\int_{\Omega} f = 1$ ,  
 $P(A) := \int_A f$

V pravděpodobnostním prostoru  $(\Omega, \mathcal{F}, P)$  platí pro  $A, B \in \mathcal{F}$

1.  $P(A^c) = 1 - P(A)$  ... ( $A^c = \Omega \setminus A$ )
2.  $A \subseteq B \implies P(A) \leq P(B)$
3.  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$  ... PIE
4.  $P(A_1 \cup A_2 \cup \dots) \leq \sum_i P(A_i)$  (subaditivita, Booleova nerovnost)
5. Definujeme podmíněnou pravděpodobnost (pro  $P(B) > 0$ ).

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

6.  $Q(A) = P(A|B)$  splňuje axiomy pro pravděpodobnost.

$$P(\emptyset|B) = 0$$

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

$$\begin{aligned} P(A_1 \circ A_2|B) &= \frac{P((A_1 \circ A_2) \cup B)}{P(B)} = \frac{P((A_1 \cap B) \cup (A_2 \cap B))}{P(B)} \\ &= \frac{P(A_1 \cap B)}{P(B) + \frac{P(A_2 \cap B)}{P(B)}} = P(A_1|B) + P(A_2|B) \end{aligned}$$

## 2.2 Podmíněná pravděpodobnost

**Definice** (Zřetězené podmínování):

$$P(A \cup B) = P(B)P(A|B)$$

**Věta:** Pokud  $A_1, \dots, A_n \in \mathcal{F}$  a  $P(A_1 \cap \dots \cap A_n) > 0$ , tak

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \dots P(A_n | \bigcap_{i=1}^{n-1} A_i)$$

**Důkaz:** indukcí

□

**Příklad:** Vytáhneme 3 karty z balíčku 52 karet. Jaká je  $P(\text{žádné srdce})$ ?

$A_i$  = i-tá karta není srdce

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(A_1) \times P(A_2|A_1)P(A_3|A_2 \cap A_1) = \frac{13 * 3}{52} \times \frac{13 * 3 - 1}{51} \times \frac{13 * 3 - 2}{50} \\ &= \frac{\# \text{dobrých}}{\# \text{všech}} = \frac{\binom{39}{3}}{\binom{52}{3}} \end{aligned}$$

**Definice:** Spočetný systém množin  $B_1, B_2, \dots \in \mathcal{F}$  je rozklad (partition)  $\Omega$ , Pokud

1.  $B_i \cap B_j = \emptyset$  pro  $i \neq j$  a
2.  $\bigcup_i B_i = \Omega$ .

**Věta** (Věta o úplné pravděpodobnosti): = Rozbor všech možností  
Pokud  $B_1, B_2, \dots$  je rozklad  $\Omega$  a  $A \in \mathcal{F}$ , tak

$$P(A) = \sum_i P(B_i)P(A|B_i)$$

(sčítance s  $P(B_i) = 0$  považujeme za 0).

$$A = (A \cap B_1) \cup (A \cap B_2) \cup \dots$$

(sjednocení disjunktních množin)

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(B_i)P(A|B_i)$$

**Příklad:** Máme tři mince: P+O, P+P, O+O. Jaká je pravděpodobnost, že padne orel?

Označíme  $M_1, M_2, M_3$  pro P+O, P+P, O+O.

$$\begin{aligned} P(O) &= P(M_1)P(O|M_1) + P(M_2)P(O|M_2) + P(M_3)P(O|M_3) \\ &= \frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times 0 + \frac{1}{3} \times 1 = \frac{1}{2} \end{aligned}$$

Rychlejší je vypsát si strom a pak posčítat výsledné jevy



**Příklad** (Gambler's ruin - zbankrotování hazardního hráče.): Máme  $a$  korun, náš protihráč  $b$  korun. Hrajeme opakovaně spravedlivou hru o 1kč, dokud někdo nepřijde o všechny peníze. Jaká je pravděpodobnost, že vyhraje?

**Důkaz:**

$$\begin{aligned}
 P_a &= P(\text{z této pozice vyhraje}) \\
 P_0 &= 0, P_n = 1 \dots (a + b = n) \\
 P(\text{výhra} | 1. \text{ kolo výhra}) &P(1. \text{ kolo výhra}) \\
 + P(\text{výhra} | 1. \text{ kolo prohra}) &P(1. \text{ kolo prohra}) \\
 \text{výhra} &\implies P_{a+1}, \text{prohra} \implies P_{a-1} \\
 P_a &= \frac{P_{a+1}}{2} + \frac{P_{a-1}}{2} \\
 &\Leftrightarrow \\
 P_a - P_{a-1} &= P_{a+1} - P_a = \Delta \\
 1 = P_n = P_0 + n * \Delta &\implies \Delta = \frac{1}{n} \\
 P_a &= \frac{a}{a+b} = \frac{a}{n}
 \end{aligned}$$

□

**Věta** (Bayesova Věta): Pokud  $B_1, B_2, \dots$  je rozklad  $\Omega$ ,  $A \in \mathcal{F}$ ,  $P(A) > 0$  a  $P(B_j) > 0$ , tak

$$P(B_j|A) = \frac{P(B_j)P(A|B_j)}{P(A)} = \frac{P(B_j)P(A|B_j)}{\sum_i P(B_i)P(A|B_i)}.$$

(sčítance s  $P(B_i) = 0$  považujeme za 0).

**Důkaz:**

$$\begin{aligned}
 P(B_j|A)P(A) &= P(B_j)P(A|B_j) \\
 P(A \cap B_j) &= P(B_j \cap A)
 \end{aligned}$$

□

**Příklad:**  $N$  = nemocný,  $T$  = testovaný, specif.  $P(N|T)$ , sens.  $P(T|N)$ .

$$P(N|T) = \frac{P(N)P(T|N)}{P(N)P(T|N) + P(N^c)P(T|N^c)} = \frac{p * 0.8}{p * 0.8 + (1 - p) * 0.01}$$

$$p = 0.001 \dots 7\%$$

$$p = 0.0016 \dots 56\% \dots \text{momentální stav testování}$$

$$p = 0.05 \dots 80\%$$

**Definice:** Jevy  $A, B \in \mathcal{F}$  jsou nezávislé (independent) pokud  $P(A \cap B) = P(A)P(B)$ . Pak také platí  $P(A|B) = P(A)$ , pokud  $P(B) > 0$ .

**Definice:** Jevy  $\{A_i : i \in I\}$  jsou (vzájemně) nezávislé, pokud pro každou konečnou množinu  $J \subseteq I$

$$P\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} P(A_i).$$

Pokud podmínka platí jen pro dvouprvkové množiny  $J$ , nazýváme jevy  $\{A_i\}$  po dvou nezávislé (pairwise independent).

**Definice:** Necht' pro množiny z prostoru jevů platí

$$A_1 \subseteq A_2 \subseteq A_3 \subseteq \dots$$

a  $A = \bigcup_{i=1}^{\infty} A_i$ . Pak platí

$$P(A) = \lim_{i \rightarrow \infty} P(A_i).$$

**Důkaz:**

$$A = A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus A_2) \cup \dots$$

$$P(A) = P(A_1) + P(A_2 \setminus A_1) + P(A_3 \setminus A_2) + \dots$$

$$\lim_{i \rightarrow \infty} (P(A_1) + \dots + P(A_i \setminus A_{i-1})) = \lim_{i \rightarrow \infty} P(A_i).$$

□

$A_n \subset P, O^{\mathbb{N}}, A_n =$  mezi prvními  $n$  hody padl aspoň jednou orel.

$$P(A) = P(\geq 1 \text{ orel v } \infty \text{ hodech}) = \lim_{n \rightarrow \infty} \dots = 1$$

**Definice** (Náhodná veličina): Mějme pravděpodobnostní prostor  $(\Omega, \mathcal{F}, P)$ . Funkci  $X : \Omega \rightarrow \mathbb{R}$  nazveme diskrétní náhodná veličina, pokud  $I_m(X)$  je spočetná množina a pokud pro všechna reálna  $x$  platí

$$\{\omega \in \Omega : X(\omega) = x\} \in \mathcal{F}.$$

**Definice:** Pravděpodobnostní funkce diskrétní náhodné veličiny  $X$  je funkce  $p_X : \mathbb{R} \rightarrow [0, 1]$  taková, že

$$p_X(x) = P(X = x) = P(\{\omega \in \Omega : X(\omega) = x\})$$

**Definice:**  $\sum_{x \in I_m(X)} p_X(x) = 1$

**Definice:**  $S := I_m(X)1, Q(A) := \sum_{x \in A} p_X(x)$   
 $(S, \mathcal{P}(S), Q)$  je diskrétní pravděpodobnostní prostor.

**Definice:** Pro  $S = \{s_i : i \in I\}$  spočetnou množinu reálných čísel a  $c_i \in [0, 1]$   $\sum_{i \in I} c_i = 1$  existuje pravděpodobnostní prostor a diskrétní n.v.  $X$  na něm taková, že  $p_X(s_i) = c_i$  pro  $i \in I$ .

### 3 Třetí přednáška

**Definice** (Distribuční funkce): Distribuční funkce (cumulative distribution function, CDF) n.v.  $X$  je funkce

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

1.  $F_X$  je neklesající funkce
2.  $\lim_{x \rightarrow -\infty} F_X(x) = 0$
3.  $\lim_{x \rightarrow +\infty} F_X(x) = 1$
4.  $F_X$  je zprava spojitá

**Příklad:**  $X = \{0 \text{ s pravděpodobností } \frac{1}{2}, 1 \text{ s pravděpodobností } \frac{1}{2}\}$

**Důkaz:**  $F_X$  je neklesající funkce

$x < y \implies P(X \leq x) \leq P(X \leq y)$  protože  $A = \{\omega : X(\omega) \leq x\}$  a  $B = \{\omega : X(\omega) \leq y\}$ , pak  $A \subseteq B \implies P(A) \leq P(B)$

□

**Důkaz:**  $\lim_{x \rightarrow +\infty} F_X(x) = 1$

$A_n = \{X \leq n\}$ ; platí  $A_1 \subseteq A_2 \subseteq \dots$

Takže  $\bigcup_{n=1}^{\infty} A_n = \Omega$ , podle věty o spojitosti pak

$$P(\Omega) = \lim_{n \rightarrow \infty} P(A_n) = \lim_{n \rightarrow \infty} F_X(n)$$

Obdobně postupujeme pro druhou limitu.

□

### 3.1 Typy rozdělení

**Definice** (Bernoulliho/alternativní rozdělení):

1.  $X$  = počet orlů při jednom hodu nespravedlivou mincí.
2. Značíme  $X \sim \text{Bern}(p)$ . Někdy se značí  $\text{Alt}(p)$ .

1. Dáno  $p \in [0, 1]$ .
2.  $p_X(1) = p$
3.  $p_X(0) = 1 - p$
4.  $p_X(k) = 0$  pro  $k \neq 0, 1$

1. Pro libovolný jev  $A \in \mathcal{F}$  definujeme *indikátorovou* n.v.  $I_A$ :
2.  $I_A(\omega) = 1$  pokud  $\omega \in A$ ,  $I_A(\omega) = 0$  jinak.
3.  $I_A \sim \text{Bern}(P(A))$ .

**Definice** (Binomické rozdělení):

1.  $X$  = počet orlů při  $n$  hodech nespravedlivou mincí.
2. Dáno  $p \in [0, 1]$  – pravděpodobnost orla při jednom hodu.
3. Značíme  $X \sim \text{Bin}(n, p)$ .
1.  $X = \sum_{i=1}^n X_i$  pro nezávislé n.v.  $X_1, \dots, X_n \sim \text{Bern}(p)$ .
2.  $p_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$  pro  $k \in 0, 1, \dots, n$ .

$$\sum_{k=1}^n \binom{n}{k} p^k (1-p)^{n-k} = 1$$
$$(p + (1-p))^n = 1^n = 1$$

**Definice** (Hypergeometrické rozdělení):

1.  $X$  = počet vytažených červených míčku při  $n$  tazích, v osudí je  $K$  červených z  $N$  celkových míčků
2. Dáno  $n, N, K$ .
3. Značíme  $X \sim \text{Hyper}(N, K, n)$ .
4.  $p_X(k) = P(X = k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}$

**Definice** (Poissonovo rozdělení (poasón)):

1. Značíme  $X \sim \text{Pois}(\lambda)$ .
2. Dáno reálné  $\lambda > 0$ .
3.  $p_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}$
4.  $\text{Pois}(\lambda)$  je limitou  $\text{Bin}(n, \lambda/n) \dots \sim X_n \dots \lambda$  pevné
5.  $X$  popisuje např. počet emailů, které dostaneme za jednu hodinu.

chceme  $\sum \frac{\lambda^k}{k!} e^{-1} = 1$

$$e^\lambda = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$$

$$P(X_n = k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} =$$

$$= \frac{n(n-1)\dots(n-k+1)}{k!} \frac{\lambda^k}{n^k} \left(1 - \frac{1}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} =$$

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

**Poznámka** (Poissonovo paradigma):  $A_1, \dots, A_n$  jsou (skoro) nezávislé jevy s  $P(A_i) = p_i$ ,  $\lambda = \sum_j p_j$ . Necht  $n$  je velké, každé z  $p_i$  malé. Pak přibližně platí

$$\sum_{i=1}^n I_{A_i} \sim Pois(\lambda)$$

**Definice** (Geometrické rozdělení):

1.  $X$  = kolikátým hodem mincí padl první orel.
2. Značíme  $X \sim Geom(p)$ .
3. Dáno  $p \in [0, 1]$ .
4.  $p_X(k) = (1-p)^{k-1}p$ , pro  $k = 1, 2, \dots$
5. Někdy se tomuto rozdělení říká posunuté geometrické, a za normální geometrické se považuje rozdělení  $X-1$ , t.j. počet neúspěšných hodů.

**Důkaz:** chceme  $\sum (1-p)^{k-1}p = 1$

$$= \frac{(1-p)^0 p}{1 - (1-p)} = \frac{p}{p} = 1$$

□

**Definice** (Střední hodnota): Pokud  $X$  je diskrétní n.v., tak její střední hodnota (expectation) je označovaná  $\mathbb{E}(X)$  a definovaná

$$\mathbb{E}(X) = \sum_{x \in Im(X)} xP(X=x),$$

pokud součet má smysl.

Necht  $X$  je definovaná na diskrétním prostoru  $(\Omega, \mathcal{F}, P)$ . Pak střední hodnotu lze také definovat jako vážený průměr

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega)P(\omega).$$

**Poznámka:** Obě definice spolu souhlasí.

**Důkaz:**

$$\sum_{x \in Im(X)} \sum_{\omega \in \Omega} X(\omega)P(\omega) = \sum_{x \in Im(X)} (x \times P(\omega \in \Omega : X(\omega) = x))$$

□

## 3.2 Rozptyl a LOTUS

**Definice** (Rozptyl): Rozptyl(variace) n.v.  $X$  nazveme číslo  $\mathbb{E}((X - \mathbb{E}X)^2)$ . Značíme jej  $var(X)$

**Věta:**

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

**Definice** (LOTUS (Law of The Unconscious Statistist)): Pro reálnou funkci  $g$  a diskrétní n.v.  $X$  je  $Y = g(X)$  také diskrétní n.v.

**Věta (LOTUS):** Pokud  $X$  je diskrétní n.v. a  $g$  reálná funkce, tak

$$\mathbb{E}(g(X)) = \sum_{x \in \text{Im}(X)} g(x)P(X = x)$$

pokud součet má smysl.

**Důkaz:**

$$Y = g(X)$$

$$\begin{aligned} \mathbb{E}Y &= \sum_{y \in Y} y \times P(Y = y) \dots \text{definice} \\ &= \sum_{y \in Y} \sum_{x \in \text{Im}(X)} g(x)P(X = x) \\ &= \sum_{x \in \text{Im}(X)} g(x)P(X = x) \end{aligned}$$

□

## 4 Čtvrtá přednáška

**Věta:** Necht'  $X, Y$  jsou diskrétní n.v. a  $a, b \in \mathbb{R}$ .

1. Pokud  $P(X \geq 0) = 1$  a  $\mathbb{E}(X) = 0$ , tak  $P(X = 0) = 1$ .
2. Pokud  $\mathbb{E}(X) \geq 0$  tak  $P(X \geq 0) > 0$ .
3.  $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$ .
4.  $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$

**Důkaz:**

1.

$$\begin{aligned} \mathbb{E}(X) &= \sum_{x \in X} xP(X = x) = 0P(X = 0) + \sum_{x > 0 \wedge x \in X} xP(X = x) = \sum_{x < 0 \wedge x \in X} xP(X = x) = 0 \\ &\implies \forall x > 0 : P(X = x) = 0 \implies P(X = 0) = 1 \end{aligned}$$

2.

$$\mathbb{E}(X) = \sum xP(X \geq x) = 0$$

kdyby ne:  $P(x \geq 0) = 0$ , všechny členy v sumě by byly záporné...spor

3.

$$\mathbb{E}(aX + b) = \sum_{x \in X} (ax + b)P(X = x) = a \sum xP(X = x) + b \sum P(X = x)$$

4.

$$\mathbb{E}(X + Y) = \sum_{\omega} (X(\omega) + Y(\omega))P(\omega) = \sum X(\omega)P(\omega) + \sum Y(\omega)P(\omega) = \mathbb{E}(X) + \mathbb{E}(Y)$$

□

**Věta:** Necht'  $X$  je diskrétní n.v. nabývající jen hodnot  $z \in \mathbb{N} = 0, 1, 2, \dots$ . Pak platí

$$\mathbb{E}(X) = \sum_{n=0}^{\infty} P(X > n).$$

**Důkaz:**

$$\begin{aligned}\mathbb{E}(X) &= \sum_{k=0}^{\infty} kP(X=k) = \sum_{k=0}^{\infty} \sum_{n=0}^{k-1} P(X=k) \\ &= \sum_{n=0}^{\infty} \sum_{k=n+1}^{\infty} P(X=k) = \sum_{n=0}^{\infty} \sum_{k=n+1}^{\infty} P(\omega \in \Omega : X(\omega) = k) \\ &= \sum_{n=0}^{\infty} P(\omega \in \Omega : X(\omega) > n) = \sum_{n=0}^{\infty} P(X > n)\end{aligned}$$

□

**Definice** (Rozptyl): Rozptyl (*variance*) n.v.  $X$  nazveme číslo  $\mathbb{E}((X - \mathbb{E}(X))^2)$ . Značíme jej  $\text{var}(X)$ . ... (kvadratické měření odchylky)

1. Směrodatná odchylka (standard deviation)  $\sigma_X = \sqrt{\text{var}(X)}$

**Poznámka:** "stejné jednotky jako  $X$ "

2. Měří, jak je daleko "typicky"  $X$  od  $\mathbb{E}(X)$ . Mohli bychom to měřit i jinak (např.  $\mathbb{E}(|X - \mathbb{E}(X)|)$ , ale rozptyl je výhodnější).

**Věta:**  $\text{var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$

**Důkaz:**

$$\begin{aligned}\mu &= \mathbb{E}(X) \\ \text{var}(X) &= \mathbb{E}((X - \mu)^2) = \mathbb{E}(X^2 - 2\mu X + \mu^2) = \mathbb{E}(X^2) - 2\mu\mathbb{E}(X) + \mu^2 \\ &= \end{aligned}$$

□

**Definice** (Podmíněná střední hodnota): Pokud  $X$  je diskrétní n.v. a  $P(B) > 0$ , tak podmíněná střední hodnota  $X$  za předpokladu  $B$  (conditional expectation of  $X$  given by  $B$ )

**Věta** (Věta o úplné střed. hodnotě): Pokud  $B_1, B_2, \dots$  je rozklad  $\Omega$  a  $X$  je d.n.v., tak

$$\mathbb{E}(X) = \sum_i \mathbb{E}(X|B_i)P(B_i)$$

kdykoliv má součet smysl. (Sčítance s  $P(B_i) = 0$  považujeme za 0.)

**Důkaz:**

$$\begin{aligned}\mathbb{E}(X) &= \sum_i P(B_i)\mathbb{E}(X|B_i) \\ &= \sum_i P(B_i) \sum_x xP(X=x|B_i) \\ &= \sum_x x \left( \sum_i P(B_i)P(X=x|B_i) \right)\end{aligned}$$

□

**Poznámka:**

Rozbor všech možností:  $X \sim \text{Geom}(p)$

$B_1 = S \dots$  první pokus úspěšný

$B_2 = B_1^C = F \dots$  první pokus neúspěšný

$$\begin{aligned}\mathbb{E}(X) &= P(S)\mathbb{E}(X|S) + P(F)\mathbb{E}(X|F) \\ &= p \cdot 1 + (1-p)(\mathbb{E}(X) + 1)\end{aligned}$$

$$p\mathbb{E}(X) = p + (1-p) = 1$$

$$\mathbb{E}(X) = \frac{1}{p}$$

## 4.1 Parametry rozdělení

**Věta** (Parametry rozdělení - Bernoulliho):

Pro  $X \sim \text{Bern}(p)$  je

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

**Důkaz:**  $\mathbb{E}(X) = 0P(X=0) + 1P(X=1) = P(X=1) = p$

$$\text{var}(X) = \mathbb{E}(X-p)^2 = (0-p)^2P(X=0) + (1-p)^2P(X=1) = p(1-p)(p + (1-p)) = p(1-p) \quad \square$$

**Věta** (Parametry rozdělení - binomické):

Pro  $X \sim \text{Bin}(n, p)$  je

1.  $\mathbb{E}(X) = np$
2.  $\text{var}(X) = np(1-p)$

**Důkaz:**

1. První postup:  $X = \sum_{i=1}^n X_i$ , kde  $X_i = [i\text{-tý hod úspěš}]$

$$\mathbb{E}(X_i) = P(X_i = 1) = p$$

2. Druhý postup:

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

$\square$

**Věta** (Parametry rozdělení - hypergeometrické): Pro  $X \sim \text{Hyper}(N, K, n)$  je

1.  $\mathbb{E}(X) = n \frac{K}{N}$
2.  $\text{var}(X) = n \frac{K}{N} (1 - \frac{K}{N}) \frac{N-n}{N-1}$

1. První postup:  $X = \sum_{i=1}^n X_i$ , kde  $X_i = [i\text{-tý míček červený}]$

$$\mathbb{E}(X_i) = P(X_i = 1) = \frac{K}{N}$$

2. Druhý postup:

$$\mathbb{E}(X) = \sum_{j=1}^K Y_j, \text{ kde } Y_j = [byl \text{ vytažen (z } n \text{ tahů) míček s číslem } j]$$

$$\begin{aligned} \mathbb{E}(Y_j) &= P(Y_j = 1) = \frac{n}{N} \\ &= \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N} \end{aligned}$$

**Věta** (Parametry rozdělení - geometrické): *Pro  $X \sim \text{Geom}(p)$  je*

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

**Věta** (Parametry rozdělení - hypergeometrické): *Pro  $X \sim \text{Hyper}(N, K, n)$  je*

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

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$
$$\mathbb{E}(X) = \sum k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda$$
$$\lambda \sum \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} = \lambda$$

## 4.2 Náhodné vektory

**Definice** (Základní popis náhodných vektorů):

1.  $X, Y$  - náhodné veličiny na stejném pravděpodobnostním prostoru  $(\Omega, \mathcal{F}, P)$ .
2. Budeme chtít uvažovat  $(X, Y)$  jako jeden objekt - náhodný vektor.
3. Jak to udělat?
4. Příklad: házíme dvakrát čtyřstěnnout kostkou,  $X$  = první hod,  $Y$  = druhý hod.

**Definice:** Pro diskrétní n.v.  $X, Y$  na pravděpodobnostním prostoru  $(\Omega, \mathcal{F}, P)$  definujeme jejich sdruženou pravděpodobnostní funkci (joint pmf)  $p_{X,Y} : \mathbb{R}^2 \rightarrow [0, 1]$  předpisem

$$p_{X,Y}(x, y) = P(\omega \in \Omega : X(\omega) = x \& Y(\omega) = y) = P(X = x \& Y = y)$$

## 4.3 Marginální rozdělení

Máme-li dáno  $p_{X,Y}$ , jak zjistit rozdělení jednotlivých složek, t.j.  $p_X$  a  $p_Y$  ?

**Věta:** *Nechť  $X, Y$  jsou diskrétní n.v. Pak:*

$$p_X(x) = P(X = x) = \sum_{Y \in \text{Im}(Y)} P(X = x \& Y = y) = \sum_{Y \in \text{Im}(Y)} p_{X,Y}(x, y)$$
$$p_Y(y) = P(Y = y) = \sum_{X \in \text{Im}(X)} P(X = x \& Y = y) = \sum_{X \in \text{Im}(X)} p_{X,Y}(x, y)$$

**Věta:** *Nechť  $X, Y$  jsou n.v. na  $(\Omega, \mathcal{F}, P)$ , nechť  $g : \mathbb{R}^2 \rightarrow \mathbb{R}$  je funkce.*

- Pak  $Z = g(X, Y)$  je n.v. na  $(\Omega, \mathcal{F}, P)$
- a platí pro ni

$$\mathbb{E}(g(X, Y)) = \sum_{x \in \text{Im}(X)} \sum_{y \in \text{Im}(Y)} g(x, y) P(X = x, Y = y).$$

**Věta:** *Pro  $X, Y$  n.v. a  $a, b \in \mathbb{R}$  platí*

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



**Definice** (Nezávislost náhodných veličin): Diskrétní n.v.  $X, Y$  jsou nezávislé (independent) pokud pro každé  $x, y \in \mathbb{R}$  jsou jevy  $\{X = x\}$  a  $\{Y = y\}$  nezávislé. To nastane právě když

$$P(X = x, Y = y) = P(X = x)P(Y = y).$$

**Věta** (Součin nezávislých n.v.): Pro nezávislé diskrétní n.v.  $X, Y$  platí

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

**Důkaz:**

$$\begin{aligned} \mathbb{E}(XY) &= \sum_{x \in \text{Im}(X), y \in \text{Im}(Y)} P(X = x, Y = y) \\ &= \sum_x x P(X = x) \sum_y y P(Y = y) = \mathbb{E}(X)\mathbb{E}(Y) \end{aligned}$$

□

## 5 Pátá přednáška

**Definice** (Coupling):

1.  $X = \sum_{i=1}^n X_i$  kde  $X_1, \dots, X_n$  jsou n.n.v.  $\dots \sim \text{Bern}(p)$
2.  $Y = \sum_{i=1}^n Y_i$  kde  $Y_1, \dots, Y_n$  jsou n.n.v.  $\dots \sim \text{Bern}(q) \dots o < q$
3. vztah  $X, Y$  není určen, můžou být jakékoliv.
4. Zařídíme, že nebudou nezávislé, dokonce bude vždy  $X \leq Y$ .
5. Stačí definovat:

pokud  $X_i = 1$  tak  $Y_i = 1$

pokud  $X_i = 0$  tak  $Y_i$  buď 1 nebo 0

$$\implies Y_1, \dots, Y_n \text{ jsou n.n.v. } \implies Y \sim \text{Bin}(n, q)$$

$$\implies X \leq Y \text{ vždy } (Y \leq k \implies X \leq k) \implies P(X \leq k) \geq P(Y \leq k)$$

**Věta** (Funkce náhodného vektoru):

Nechť  $X, Y$  jsou n.v. na  $(\Omega, \mathcal{F}, P)$ , nechť  $g: \mathbb{R}^2 \rightarrow \mathbb{R}$  je funkce.

1. Pak  $Z = g(X, Y)$  je n.v. na  $(\Omega, \mathcal{F}, P)$
2. platí pro ni

$$\mathbb{E}(g(X, Y)) = \sum_{x \in \text{Im}(X)} \sum_{y \in \text{Im}(Y)} g(x, y) P(X = x, Y = y)$$

**Věta** (Linearita střední hodnoty):

Pro  $X, Y$  n.v. a  $a, b \in \mathbb{R}$  platí

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

**Důkaz:**

$$g(x, y) = ax + by$$

$$\begin{aligned} \mathbb{E}(aX + bY) &= \mathbb{E}(g(X, Y)) = \sum_{x, y} g(x, y) P(X = x, Y = y) = \sum_{x, y} ax P(X = x, Y = y) \\ &+ \sum_{x, y} by P(X = x, Y = y) = \sum_x ax P(X = x) + \sum_y by P(Y = y) \end{aligned}$$

□

**Věta** (Konvoluce): Pokud  $X, Y$  jsou diskrétní náhodné veličiny, tak pro  $Z = X + Y$  platí

$$P(Z = z) = \sum_{x \in \text{Im}(X)} P(X = x, Y = z - x).$$

Pokud  $X, Y$  jsou navíc nezávislé, tak

$$P(Z = z) = \sum_{x \in \text{Im}(X)} P(X = x)P(Y = z - x).$$

**Důkaz:**

$$\begin{aligned} P_z &= \sum_x P_X(x)P_Y(z - x) \dots \text{konvoluce} \\ P(Z = z) &= \sum_k P(X = k \& Y = z - k) \\ &= \sum_{k=0}^m P(X = k)P(Y = z - k) \\ &= \sum_{k=0}^m \binom{m}{k} p^k (1-p)^{m-k} \binom{n}{z-k} p^{z-k} (1-p)^{n-(z-k)} \\ &= \sum_{k=0}^m p^z (1-p)^{m+n-z} \binom{m}{k} \binom{n}{z-k} \\ &= p^z (1-p)^{m+n-z} \sum \binom{m}{k} \binom{n}{z-k} \\ &= \text{Bin}(m+n, p) \end{aligned}$$

□

**Definice** (Podmíněné rozdělení):  $X, Y$  - diskrétní náhodné veličiny na  $(\Omega, \mathcal{F}, P)$ ,  $A \in \mathcal{F}$

1.  $p_{X|A}(x) := P(X = x|A)$  ... příklad:  $X$  je výsledek hodu kostkou,  $A$  = padlo sudé číslo
2.  $p_{X|Y}(x|y) := P(X = x|Y = y)$  ... příklad:  $X, Z$  jsou výsledky dvou nezávislých hodů kostkou,  $Y = X + Z$ .

**Definice** (Obecná náhodná veličina): Náhodná veličina (random variable) na  $(\Omega, \mathcal{F}, P)$  je zobrazení  $X : \Omega \rightarrow \mathbb{R}$ , které pro každé  $x \in \mathbb{R}$  splňuje

$$\omega \in \Omega : X(\omega) \leq x \in \mathcal{F}$$

...

$$F_X(x) = P(X \leq x)$$

**Definice** (Spojitá náhodná veličina): N.v.  $X$  se nazývá spojitá (continuous), pokud existuje nezáporná reálná funkce  $f_X$  tak, že

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(t) dt$$

Někdy se též používá pojem absolutně spojitá veličina.

Funkce  $f_X$  se nazývá hustota (probability density function) náhodné veličiny  $X$ .

Podmínka na hustotu:

$$\int_{-\infty}^{\infty} f_X(t) dt = 1 \dots \lim_{x \rightarrow \infty} F_X(x) = 1$$

## 6 Šestá přednáška

**Definice** (Kvantilová funkce): Pro náhodnou veličinu  $X$  definujeme *kvantilovou funkci*  $Q_X : [0, 1] \rightarrow \mathbb{R}$  pomocí

$$Q_X(p) := \min\{x \in \mathbb{R} : p \leq F_X(x)\}$$

1. Pokud  $F_X$  je spojitá, tak  $Q_X = F_X^{-1}$ .
2. Obecně platí:  $Q_X(p) \leq x \Leftrightarrow p \leq F_X(x)$ .
3.  $Q_X(\frac{1}{2}) = \text{medián}$  (pozor, když  $F_X$  není rostoucí)

**Definice** (Spojitá náhodná veličina): N.v.  $X$  se nazývá spojitá (*continuous*) pokud existuje nezáporná reálná funkce  $f_X$  tak, že

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(t) dt$$

1. Alternativně: máme zadanou funkci  $f \geq 0$  s  $\int_{-\infty}^{\infty} f = 1$ .
2. Vybereme náhodný bod pod grafem  $f$ .
3. Označíme jeho souřadnice  $(X, Y)$ .
4. Pak je  $X$  n.v. s hustotou  $f$ .

**Věta** (Práce s hustotou): *Nechť spojitá n.v.  $X$  má hustotu  $f_X$ . Pak*

1.  $P(X = x) = 0 \quad \forall x \in \mathbb{R}$ .
2.  $P(a \leq X \leq b) = \int_a^b f_X(t) dt \quad \forall a, b \in \mathbb{R}$ .
3. *V důsledku taky platí (pro rozumnou množinu  $A$ ):*

$$P(X \in A) = \int_A f_X(t) dt$$

**Důkaz:**

$$2 \implies 1 : P(x \leq X \leq x) = \int_x^x f = 0$$

$$2 : P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a) = \int_{-\infty}^b f - \int_{-\infty}^a f$$

$$P(a \leq X \leq b) = \lim_{n \rightarrow \infty} P(a - \frac{1}{n} < X \leq b) = \lim_{n \rightarrow \infty} \int_{a - \frac{1}{n}}^b f = \int_a^b f$$

□

**Definice** (Střední hodnota spojitě n.v.): Nechť spojitá n.v.  $X$  má hustotu  $f_X$ . Pak její střední hodnota (*expectation, expected value, mean*) je označována  $\mathbb{E}(X)$  a definována

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

pokud integrál má smysl, t.j. pokud se nejedná o typ  $\infty - \infty$ .

1. Analogie s výpočtem těžiště tyče ze znalosti hustoty
2. Diskretizace.

**Věta** (LOTUS): *Pokud  $X$  je spojitá n.v. s hustotou  $f_X$  a  $g$  reálná funkce, tak*

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

*pokud integrál má smysl. (Důkaz pomocí substituce v integrálu)*

**Věta** (Linearita střední hodnoty): Pro  $X_1, \dots, X_n$  diskrétní nebo spojitě n.v. platí

$$\mathbb{E}(X_1 + \dots + X_n) = \mathbb{E}(X_1) + \dots + \mathbb{E}(X_n)$$

**Definice** (Rozptyl spojitě n.v.):

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx$$
$$\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x) dx$$

Označíme-li  $\mu = \mathbb{E}(X)$ , tak

$$\text{var}(X) := \mathbb{E}((X - \mu)^2) = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx$$

**Věta:** Pro spojitě n.v. platí  $\text{var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$

**Důkaz:** (Důkaz jako pro diskrétní n.v.) □

**Věta** (Rozptyl součtu): Pro  $X_1, \dots, X_n$  nezávislé diskrétní nebo spojitě n.v. platí

$$\text{var}(X_1 + \dots + X_n) = \text{var}(X_1) + \dots + \text{var}(X_n).$$

**Důkaz:** Triviální. □

**Definice** (Uniformní rozdělení): N.v.  $X$  má uniformní rozdělení na intervalu  $[a, b]$ , píšeme  $X \sim U(a, b)$ , pokud  $f_X(x) = \frac{1}{b-a}$  pro  $x \in [a, b]$  a  $f_X(x) = 0$  jinak.

**Definice** (Exponenciální rozdělení):

$$F_X(x) = \begin{cases} 0 & \dots x \leq 0 \\ 1 - e^{-\lambda x} & \dots x \geq 0 \end{cases}$$

**Poznámka:**  $X$  modeluje např. čas před příchodem dalšího telefonního hovoru do callcentra, dotazu na webserver, čas do dalšího blesku v bouřce atd.

**Poznámka:** Souvislost  $X \sim \text{Exp}(\lambda)$  a  $Y \sim \text{Geom}(p)$

1.  $P(X > x) = e^{-\lambda x}$  pro  $x > 0$
2.  $P(Y > n) = (1 - p)^n$  pro  $n \in \mathbb{N}$

**Definice** (Standardní normální rozdělení):

1.  $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
2.  $\Phi(x)$  - primitivní funkce k  $\phi$
3. Standardní normální rozdělení  $N(0, 1)$  má hustotu  $\phi$  a distribuční funkci  $\Phi$ .
4. Pokud  $Z \sim N(0, 1)$ , tak  $\mathbb{E}(Z) = 0$  a  $\text{var}(Z) = 1$ .

**Definice** (Obecné normální rozdělení):

1. Pro  $\mu, \sigma \in \mathbb{R}, \sigma > 0$  položíme  $X = \mu + \sigma Z$ , kde  $Z \sim N(0, 1)$ .
2. Píšeme  $X \sim N(\mu, \sigma^2)$  - obecné normální rozdělení
3. Normální rozdělení  $N(\mu, \sigma^2)$  má hustotu  $\frac{1}{\sigma} \phi\left(\frac{x-\mu}{\sigma}\right)$

$$\Phi(z) = P(Z \leq z) = P(X \leq \mu + \sigma z) = F_X(\mu + \sigma z)$$

**Poznámka** (Odolnost vůči součtu): Pokud  $X_1, \dots, X_k$  jsou n.n.v., kde  $X_i \sim N(\mu_i, \sigma_i^2)$ , pak

$$X_1 + \dots + X_k \sim N(\mu, \sigma^2),$$

kde  $\mu = \mu_1 + \dots + \mu_n$ .

**Poznámka** (Normální rozdělení - klíčové vlastnosti):

1. Pravidlo  $3\sigma$  (68 – 95 – 99.7 rule)

$$X \sim N(\mu, \sigma^2)$$

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = 68\%$$

$$2\sigma = 95$$

$$3\sigma = 99.7$$

2. Centrální limitní věta

## 7 Sedmá přednáška

**Definice** (Cauchyho rozdělení): hustota  $f(x) = \frac{1}{\pi(1+x^2)}$  nemá střední hodnotu!

**Poznámka:**

$$\begin{aligned} \int_{-\infty}^{\infty} f &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{1}{1+x^2} = \frac{1}{\pi} [\arctg(x)]_{-\infty}^{\infty} = 1 \\ \mathbb{E}X &= \int_{-\infty}^{\infty} xf(x) = \int_0^{\infty} \frac{2x}{2\pi(1+x^2)} + \int_{-\infty}^0 \frac{x}{\pi(1+x^2)} \\ &\quad \left[ \frac{1}{2\pi} \log(1+x^2) \right]_0^{\infty} + \left[ \frac{1}{x\pi} \log(1+x^2) \right] \\ &\quad \infty - 0 + 0 - \infty = \infty - \infty?! \end{aligned}$$

**Definice** (Gamma rozdělení):  $Gamma(w, \lambda)$ , gamma rozdělení s parametry  $w > 0$  a  $\lambda > 0$  má hustotu

$$f(x) = 0 \text{ pro } x \leq 0 \quad \& \quad \frac{1}{\Gamma(w)} \lambda^w x^{w-1} e^{-\lambda x} \text{ pro } x \geq 0$$

kde  $\Gamma(w) = (w-1)! = \int_0^{\infty} x^{w-1} e^{-x} dx$

Pro  $w = 1$  dostáváme znovu exponenciální rozdělení ...  $\frac{1}{0!} \lambda^1 e^{-\lambda x}$

Pokud  $X_1, \dots, X_n$  jsou n.n.v s rozdělením  $Exp(\lambda)$ , tak  $X_1 + \dots + X_n \sim Gamma(n, \lambda)$ .

**Věta:** Nechť  $X$  je n.v. s distribuční funkcí  $F_X = F$ , nechť  $F$  je spojitá a rostoucí. Pak  $F(X) \sim U(0, 1)$ .

**Důkaz:**

$$\begin{aligned} F_Y(y) &= P(F(X) \leq y) = 0 \text{ pro } y < 0 \& 1 \text{ pro } y \geq 1 \\ \text{pro } y \in (0, 1) P(X \leq x) &\implies \text{stejně jevy } \dots = F(x) = y \end{aligned}$$

□

**Věta:** Nechť  $F$  je funkce "typu distribuční funkce": neklesající zprava spojitá funkce s  $\lim_{x \rightarrow -\infty} F(x) = 0$  a  $\lim_{x \rightarrow \infty} F(x) = 1$ .

Nechť  $Q$  je odpovídající kvantilová funkce. Nechť  $U \sim U(0, 1)$ ,  $X = Q(U)$ . Pak  $X$  má distribuční funkci  $F$ .

**Důkaz:**

$$F_X(x) = P(Q(U) \leq x)$$

**Poznámka:**

$$Q(p) = \inf\{x : F(X) \geq p\} \implies Q(p) \leq x \Leftrightarrow F(x) \geq p$$

$$F_X(x) = P(U \leq F(x)) = F(x)$$

□

**Příklad:**

$$\begin{aligned} F(x) &= 1 - e^{-\lambda x} \dots \text{Exp}(\lambda) \\ Q(p) &= \frac{\log(1-p)}{-\lambda} > 0 \\ U &\sim U(0,1) \dots \frac{\log(1-U)}{-\lambda} \sim \text{Exp}(\lambda) \end{aligned}$$

**Definice:** Sdružená distribuční funkce (Joint cdf)

Pro n.v.  $X, Y$  na pravděpodobnostním prostoru  $(\Omega, \mathcal{F}, P)$  definujeme jejich sdruženou distribuční funkci (joint cdf)  $F_{X,Y} : \mathbb{R}^2 \rightarrow [0, 1]$  předpisem

$$F_{X,Y}(x, y) = P(\{\omega \in \Omega : X(\omega) \leq x \& Y(\omega) \leq y\}).$$

1. Formální podmínka: potřebujeme  $\{X \leq x \& Y \leq y\} \in \mathcal{F}$ , jinak  $(X, Y)$  není náhodný vektor.
2. Mohli bychom definovat i pro více než dvě n.v.  $\dots F_{X_1, \dots, X_n}(x_1, \dots, x_n) = P(X_1 \leq x_1 \& \dots X_n \leq x_n)$ .
3. Můžeme odsud odvodit pravděpodobnost obdélníku:

$$P(X \in (a, b] \& Y \in (c, d]) = F(b, d) - F(b, c) - F(a, d) + F(a, c)$$

**Definice:** Sdružená hustota (Joint pdf)

Často můžeme sdruženou distribuční funkci psát jako integrál pomocí nezáporné funkce  $f_{X,Y}$

$$F_{X,Y}(x, y) = \int_{-\infty}^x \int_{-\infty}^y f_{X,Y}(s, t) ds dt.$$

Pak nazýváme n.v.  $X, Y$  sdruženě spojitě. Funkce  $f_{X,Y}$  je jejich sdružená hustota.

Jako u jednorozměrného případu může být  $f_{X,Y} > 1$ .

Stejně jako u jednorozměrného případu můžeme pak pomocí hustoty vyjádřit i další pravděpodobnosti, pro "rozumnou množinu  $A$ ".

$$\begin{aligned} P((X, Y) \in A) &= \int_A f_{X,Y}(x, y) dx dy \\ \int_{\mathbb{R}^2} f_{X,Y} &= 1 \end{aligned}$$

**Poznámka:**

$$\begin{aligned} f_{X,Y}(x, y) &= \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y) \\ f_{X,Y}(x) &\doteq \frac{P(x \leq X \leq x + \Delta_x \& y \leq Y \leq y + \Delta_y)}{\Delta_x \Delta_y} \\ P((X, Y) \in A) &= \int_A f = \int_x^{x+\Delta_x} \int_y^{y+\Delta_y} f_{X,Y}(s, t) ds dt = f_{X,Y}(x, y) \Delta_x \Delta_y \end{aligned}$$

**Definice:** LOTUS

Analogicky jako v diskrétním případě platí pro střední hodnotu funkce dvou n.v.

$$\mathbb{E}(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy.$$

A tak jako v diskretním případě odsud odvodíme

$$\begin{aligned}\mathbb{E}(aX + bY + c) &= a\mathbb{E}(X) + b\mathbb{E}(Y) + c \\ \mathbb{E}(g(X, Y)) &= \int \int g(x, y) f_{X,Y}(x, y) = \int \int axf(x, y) + \int \int byf(x, y) + c \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) = \\ &= a \int x \int f_{X,Y}(x, y) dy dx + b \int y \int f(x, y) dy dx + c = \\ &= a \int x f_X(x) + b \int y f_Y(y) + c = \\ &= a\mathbb{E}(X) + b\mathbb{E}(Y) + c\end{aligned}$$

**Definice:** Nezávislost spojitých náhodných veličin

Libovolné náhodné veličiny nazveme nezávislé (independent), pokud jevy  $\{X \leq x\}$  a  $\{Y \leq y\}$  jsou nezávislé pro libovolná  $x, y \in \mathbb{R}$ . Ekvivalentně:

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y),$$

$$F_{X,Y}(x, y) = F_X(x)F_Y(y)$$

**Věta:** Necht  $X, Y$  mají sdruženou hustotu  $f_{X,Y}$ . Následující tvrzení jsou ekvivalentní:

1.  $X, Y$  jsou nezávislé
2.  $f_{X,Y}(x, y) = f_X(x)f_Y(y)$

**Důkaz:**

$$\implies : f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y) =$$

$$F'_X F'_Y = f_X(x)f_Y(y)$$

*doplň druhú implikáciu (nestíhol som, zo slídov)*

□

Vícerozměrné normální rozdělení

1.  $\varphi(t) = \frac{e^{-t^2/2}}{\sqrt{(2\pi)}}$
2.  $f(t_1, \dots, t_n) = \varphi(t_1)\varphi(t_2) \dots \varphi(t_n) = \frac{e^{-\frac{t_1^2 + \dots + t_n^2}{2}}}{\sqrt{2\pi}}$
3.  $f(t_1, \dots, t_n) = (2\pi)^{-\frac{n}{2}} e^{-\frac{r^2}{2}}$ , kde  $r^2 = t_1^2 + \dots + t_n^2$  je radiálně symetrická funkce.
4. Necht  $Z = (Z_1, \dots, Z_n)$  má hustotu  $f$ .
5.  $Z_1, \dots, Z_n$  jsou n.n.v,  $Z_i \sim N(0, 1)$
6.  $Z/\|Z\|$  je uniformně náhodný bod na  $n$ -rozměrné sféře
7. skalární součin  $Z$  s libovolným jednotkovým vektorem je  $N(0, 1)$
8.  $\langle u, Z \rangle = \sum_{i=1}^n u_i Z_i$  má také rozdělení  $N(0, 1)$

Vícerozměrné normální rozdělení obecné

1. Obecněji můžeme vzít náhodný vektor s hustotou  $ce^{Q(t)}$ , kde  $c > 0$  je vhodná konstanta a  $Q(t)$  je obecná kvadratická funkce.
2. Používá se ve strojovém učení.
3. Souřadnice nejsou nezávislé.

## 8 Osmá přednáška

**Definice:** Podmínování

zúžení náhodné veličiny na množinu:  $X$  je n.v. na  $(\Omega, \mathcal{F}, P)$ ,  $B \in \mathcal{F}$ , t. ž.  $P(B) > 0$ .

$$F_{X|B}(x) := P(X \leq x|B)$$

K tomu příslušná hustotní funkce  $f_{X|B}$ :

Pokud  $B = \{X \in S\}$ , tak

$$f_{X|B}(x) = \begin{cases} \frac{f_X(x)}{P(X \in S)} \dots & \text{pokud } x \in S \\ 0 \dots & \text{jinak} \end{cases}$$

**Věta:** *Věta o rozkladu hustoty*

*Nechť  $X$  je spojitá n.v., nechť  $B_1, B_2, \dots$  je rozklad  $\Omega$ . Pak*

$$F_X(x) = \sum_i P(B_i) F_{X|B_i}(x),$$

$$f_X(x) = \sum_i P(B_i) f_{X|B_i}(x).$$

**Důkaz:** věta o úplné pravděpodobnosti.

$$P(X \leq x) = \sum P(\dots)$$

□

**Věta:** *Marginální hustota*

$$f_X(x) = \int_{y \in \mathbb{R}} f_{X,Y}(x, y) dy$$

$$f_Y(y) = \int_{x \in \mathbb{R}} f_{X,Y}(x, y) dx$$

**Důkaz:** TODO

□

**Definice:** Podmíněná hustota

Pro spojitě n.v.  $X, Y$  definujeme podmíněnou hustotu předpisem

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

pokud je  $f_Y(y) > 0$ , jinak ji nedefinujeme.

1. připomeňme, že  $f_Y(y) = \int_{x \in \mathbb{R}} f_{X,Y}(x, y) dx$
2. pro fixované  $y$  je  $f_{X|Y}(x|y)$  hustota.

**Věta:** *Podmíněná, sdružená a marginální hustota*

$$f_{X,Y}(x, y) = f_Y(y) f_{X|Y}(x|y)$$

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x|y) f_Y(y) dy$$



**Věta:** Součet spojitých n.v.

Nechť spojitě  $X, Y$  jsou n.n.v. Pak  $Z = X + Y$  je také spojitá n.v. a její hustotu dostaneme jako konvoluci funkcí  $f_X, f_Y$ , neboli

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx$$

**Důkaz:** Náhled:

$$P(Z = z|X = x) = P(Y = z - x)$$

$$f_{Z|X}(z|x) = f_Y(z - x)$$

(n.v.  $Z|X = x$  je stejná jako  $Y + x$ )

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_{Z|X}(z|x)f_X(x) = \\ &= \int f_Y(z - x)f_X(x) \end{aligned}$$

□

**Příklad:**  $X, Y \sim N(0, 1)$  nezávislé. n.v. ...  $f_X = f_Y = \varphi \dots \varphi(t) = \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$

$$Z = X + Y$$

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx = \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \frac{1}{\sqrt{(2\pi)}}e^{-\frac{(z-x)^2}{2}}dx \\ &= \frac{1}{2\pi}e^{-\frac{z^2}{2}} \int_{-\infty}^{\infty} e^{-x^2+zx}dx \\ &= \frac{1}{2\pi}e^{-\frac{z^2}{2}+\frac{z^2}{4}} \int e^{-(x-\frac{z}{2})^2}dx \\ &= \frac{1}{\sqrt{2}\sqrt{2\pi}}e^{-\frac{z^2}{4}} \dots \text{ hustota } N(0, 2) \end{aligned}$$

**Definice:** Podmíněná hustota a střední hodnota

1.  $\mathbb{E}(X|B) := \int_{-\infty}^{\infty} xf_{X|B}(x)dx$
2.  $\mathbb{E}(g(X)|B) := \int_{-\infty}^{\infty} g(x)f_{X|B}(x)dx$

**Věta:** Věta o úplné střední hodnotě

Nechť  $X$  je spojitá n.v.. Pokud  $B_1, B_2, \dots$  je rozklad, tak

$$\mathbb{E}(X) = \sum_i P(B_i)\mathbb{E}(X|B_i).$$

**Důkaz:** pomocí rozkladu hustoty:

$$\int xf_X(x) = \int_{-\infty}^{\infty} x \sum_i P(B_i)f_{X|B}(x) = \sum_i P(B_i) \int xf_{X|B_i}(x)$$

**Definice:** Podmíněná hustota a střední hodnota

1.  $f_{X|Y}(x|y) := \frac{f_{X,Y}(x,y)}{f_Y(y)}$  je hustota n.v.  $X$ , pokud  $Y = y$
2.  $\mathbb{E}(X|Y = y) := \int xf_{X|Y}(x,y)dx$  je střední hodnota této veličiny

$$3. \mathbb{E}(g(X)|Y=y) = \int g(x)f_{X|Y}(x,y)dx$$

4. Analogie věty o úplné střední hodnotě:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} \mathbb{E}(X|Y=y)f_Y(y)dy$$

$$5. \mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y))$$

**Definice:** Kovariance

Pro n.v.  $X, Y$  definujeme jejich kovarianci předpisem

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

Věta:

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

1.  $var(X) = cov(X, X)$
2.  $cov(X, \alpha Y + \beta Z + c) = \alpha cov(X, Y) + \beta cov(X, Z)$
3.  $cov(X, Y) = 0$  pokud  $X, Y$  jsou nezávislé
4. ale nejen tehdy

**Definice:** Korelace

Korelace náhodných veličin  $X, Y$  je definovaná předpisem

$$\varrho(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}}.$$

1. je to "přenormovaná"kovariance
2.  $-1 \leq \varrho(X, Y) \leq 1$ .
3. Korelace neznamená příčinnou souvislost! (Např. korelace je symetrická, kauzalita nikoli!)
4. Naopak, nekorelace neznamená nezávislost. (Př.  $X$  libovolná,  $Y = +X$  nebo  $Y = -X$ , obojí se stejnou pravděpodobností).

**Věta:** Rozptyl součtu

Necht  $X = \sum_{i=1}^n X_i$ . Pak

$$var(X) = \sum_{i=1}^n \sum_{j=1}^n cov(X_i, X_j) = \sum_{i=1}^n var(X_i) + \sum_{i \neq j} cov(X_i, X_j).$$

Sec. jsou  $X_1, \dots, X_n$  nezávislé, pak

$$var(X) = \sum_{i=1}^n var(X_i)$$

**Důkaz:**

$$\begin{aligned} var(X) &= \mathbb{E}(\sum X_i \times \sum X_j) - (\sum \mathbb{E}X_i)(\sum \mathbb{E}X_j) \\ &= \mathbb{E}(\sum X_i X_j) - \sum \mathbb{E}(X_i)\mathbb{E}(X_j) \end{aligned}$$

□

**Věta:** Cauchyho nerovnost

$$\mathbb{E}(XY) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}$$

**Důkaz:** jako v LA, součin norem

□

**Poznámka:** Důsledek pro korelaci:  $-1 \leq \varrho(X, Y) \leq 1$

**Věta:** Jensenova věta

Nechť  $X$  má konečnou střední hodnotu a nechť  $g$  je konvexní reálná funkce. Pak

$$\mathbb{E}(g(X)) \geq g(\mathbb{E}(X)).$$

**Důkaz:**

$$\mu = \mathbb{E}(X)$$

$$L(\mu) = g(\mu)$$

$\forall t L(t) \leq g(t) \dots L(t)$  je tečna  $g(t)$  v bodě  $\mu$

$$L(X) \leq g(X)$$

$$\mathbb{E}L(X) \leq \mathbb{E}g(X)$$

*z linearity  $L$*

$$L(\mathbb{E}X) = g(\mathbb{E}(X))$$

□

**Věta:** Markovova nerovnost

Nechť náhodná veličina  $X$  splňuje  $X \geq 0$ . Pak

$$P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

**Důkaz:**

$$\mathbb{E}(X) = P(X \geq a)\mathbb{E}(X|X \geq a) + P(X < a)\mathbb{E}(X|X < a)$$

$$\mathbb{E}(X) \geq P(X \geq a)a$$

□

**Věta:** Čebyševova nerovnost

Nechť  $X$  má konečnou střední hodnotu  $\mu$  a rozptyl  $\sigma^2$ . Pak

$$P(|X - \mu| \geq a\sigma) \leq \frac{1}{a^2}$$

**Důkaz:**

$$Y = (X - \mu)^2$$

$$P(Y \geq a^2\sigma^2) \leq \frac{\mathbb{E}(Y)}{a^2\sigma^2} = \frac{\text{var}(X)}{a^2\sigma^2}$$

□

**Věta:** Černovova nerovnost

Nechť  $X = \sum_{i=1}^n X_i$ , kde  $X_i$  jsou n.n.v. nabývající hodnot  $\pm 1$  s pravděpodobností  $1/2$ . Pak pro  $t > 0$  platí:

$$P(X \leq -t) = P(X \geq t) \leq e^{-\frac{t^2}{2\sigma^2}},$$

kde  $\sigma = \sigma_X = \sqrt{n}$

## 9 Devátá přednáška

### 9.1 Nerovnosti, které známe z minula

- Markovova

$$X \geq 0 \implies P(X \geq a\mathbb{E}(X)) \leq \frac{1}{a}$$

- Čebyševova

$$P(|X - \mathbb{E}(X)| \geq a\sigma_X) \leq \frac{1}{a^2}$$

- Chernoffova ( $\sigma_X = \sqrt{n}$ )

$$X = \sum_{i=1}^n X_i, X_i = \pm 1 \implies P(|X - \mathbb{E}(X)| \geq a\sigma_X) \leq 2e^{-a^2/2}$$

### 9.2 Slabý zákon velkých čísel

**Věta:** *Nechť  $X_1, \dots, X_n$  jsou stejné rozdělené n.n.v. se střední hodnotou  $\mu$  a rozptylem  $\sigma^2$ . Označme  $S_n = (X_1 + \dots + X_n)/n$ . Pak pro každé  $\varepsilon > 0$  platí*

$$\lim_{n \rightarrow \infty} P(|S_n - \mu| \geq \varepsilon) = 0.$$

Říkáme, že posloupnost  $S_n$  konverguje k  $\mu$  v pravděpodobnosti, píšeme  $S_n \xrightarrow{P} \mu$ .

**Důkaz:**

$$\mathbb{E}S_n = \mathbb{E}\frac{X_1 + \dots + X_n}{n} = \frac{\mathbb{E}X_1 + \dots + \mathbb{E}X_n}{n} = \frac{\mu + \dots + \mu}{n} = \mu$$

$$\text{var}(S_n) = \text{var}\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{\text{var}(X_1) + \dots + \text{var}(X_n)}{n^2} = \frac{\sigma^2 + \dots + \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$P(|S_n - \mathbb{E}S_n| \geq a\sigma_{S_n}) \leq \frac{1}{a^2} = \frac{1}{\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)^2} = \frac{\sigma^2}{\varepsilon^2 n} \xrightarrow{n \rightarrow \infty} 0$$

□

### 9.3 Centrální limitní věta

**Věta** (Centrální limitní věta): *Nechť  $X_1, \dots, X_n$  jsou stejně rozdělené n.n.v se střední hodnotou  $\mu$  a rozptylem  $\sigma^2$ . Označme*

$$Y_n = ((X_1 + \dots + X_n) - n\mu)/(\sqrt{n} \cdot \sigma).$$

*Pak  $Y_n \xrightarrow{d} N(0, 1)$ . Neboli, pokud  $F_n$  je distribuční funkce  $Y_n$ , tak*

$$\lim_{n \rightarrow \infty} F_n(x) = \Phi(x) \quad \forall x \in \mathbb{R}.$$

Říkáme, že posloupnost  $Y_n$  konverguje k  $N(0, 1)$  v distribuci.

**Doplňit tři grafy z prezentace**

## 9.4 Momentová vytvořující funkce

**Definice** (Momentová vytvořující funkce): Pro náhodnou veličinu  $X$  označíme

$$M_X(t) = \mathbb{E}(e^{tx}).$$

Funkci  $M_X(t) \dots$  **DOPLNIT**

## 9.5 Statistika

**Příklad** (1. Počet leváků): •  $\#L = 6 = 14\%$

- $\#P = 37 = 87\%$
- spolu:  $43 = 100\%$

Tipujeme, že je 4 - 12% leváků v ČR.

**Poznámka:** otázky statistiky  $\rightarrow$  co můžeme z výsledků v malém vzorku odvodit o výsledcích v celé skupině

- bodové odhady  $\dots 14\%$
- intervalové odhady  $\dots (10\%, 20\%)$

Obtíže statistiky  $\rightarrow$  otázky typu

- máme reprezentativní vzorek?
- je otázka dobře formulovaná?

**Příklad** (2. Doba běhu programu):

- $X_1, \dots, X_n \sim F$  n.n.v.,  $F$  je jejich distribuční funkce

**Definice:** Empirická distribuční funkce (empirical CDF) je definována

$$\hat{F}_n(x) = \frac{\sum_{i=1}^n I(X_i \leq x)}{n},$$

kde  $I(X_i \leq x) = 1$  pokud  $X_i \leq x$  a 0 jinak.

**Věta:** Pro pevné  $x$  platí

- $\mathbb{E}(\hat{F}_n(x)) = F(x)$
- $\text{var}(\hat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}$
- $\hat{F}_n(x)$  konverguje k  $F(x)$  v pravděpodobnosti, píšeme  $\hat{F}_n(x) \rightarrow^P F(x)$ .

**Důkaz:** Slabý zákon velkých čísel:

$$\mathbb{E}\hat{F}_n(x) = \mathbb{E}S_n = \mathbb{E}I(X_i \leq x) = P(X_i \leq x) = F(x)$$

$$\text{var}(\hat{F}_n(x)) = \frac{\text{var}(X'_1)}{n}$$

$$X'_i \sim \text{Bern}(p) \dots p = F(x)$$

□

## 9.6 Empirická distribuční funkce - Dvoretzky-Keifer-Wolfowitz (DKW)

**Věta** (Empirická distribuční funkce): Necht  $X_1, \dots, X_n \sim F$  jsou n.n.v.,  $\hat{F}_n$  jejich empirická distribuční funkce. Necht  $\mathbb{E}(X_i)$  je konečná. Zvolme  $\alpha \in (0, 1)$  (pravděpodobnost chyby) a označme  $\varepsilon = \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}$ . Pak platí:

$$P(\hat{F}_n(x) - \varepsilon \leq F(x) \leq \hat{F}_n(x) + \varepsilon) \geq 1 - \alpha$$

## 9.7 Intro - explorační analýza dat (exploratory data analysis)

- posbíráme data ( a dáme pozor na systémové chyby - nezávislost, nezaujatost. . . )
- různé tabulky (třeba v Excelu a spol.)
- vhodné obrázky: histogram, krabicový diagram (boxplot) atd.

## 10 Desátá přednáška

### 10.1 náhodný výběr

- bez vracení  
 $\Omega =$  všechny  $n$  - tice obyvatel ČR  
Pro  $\omega = (\omega_1, \dots, \omega_n)$  zvolíme  $X_i = I(\omega_i \text{ je levák})$ .
- s vracením  
 $\Omega = \{\text{všechny } n\text{-tice obyvatel ČR, mohou se opakovat}\}$   
Pro  $\omega = (\omega_1, \dots, \omega_n)$  zvolíme  $X_i = I(\omega_i \text{ je levák})$ .
- varianty (stratifikovaný výběr)  
Chceme adekvátně reprezentovat různé podmnožiny (dané věkem, bydlištěm, . . .). Nebudeme dále zkoumat.

### 10.2 Statistika - model

- nezávislá měření - hodnoty n.n.v.  $X_1, \dots, X_n \sim F$  náhodný výběr s distribuční funkcí  $F$  s rozsahem  $n$
- neparametrické modely: povolujeme velkou třídu  $F$ .
- parametrické modely:  $F \in \{F_\vartheta : \vartheta \in \Theta\}$
- příklady:
  - $Pois(\lambda)$  (parametr  $\vartheta$ )
  - $U(a, b)$  (parametr  $\vartheta = (a, b)$ ,  $\Theta = \mathbb{R}^2$ )
  - $N(\mu, \sigma^2)$  (parametr  $\vartheta = (\mu, \sigma)$ ,  $\Theta = \mathbb{R} \times \mathbb{R}^+$ )

### 10.3 Zkoumané úlohy - cíle konfirmační analýzy

1. bodové odhady
2. intervalové odhady
3. testování hypotéz
4. (linární) regrese

**Definice:** statistika - libovolná funkce náhodného výběru, tj. např. aritmetický průměr, medián, maximum atd.

$$T = T(X_1, \dots, X_n)$$

**Příklad** (Výběrový průměr a rozptyl):

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

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

$$\widehat{S}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2$$

**Definice:** Odhad je libovolná statistika.

## 10.4 Vlastnosti bodových odhadů

**Definice** (Vlastnosti bodových odhadů): Odhad  $\widehat{\Theta}_n = \widehat{\Theta}_n(X_1, \dots, X_n)$  parametrů  $\vartheta$  je

- nestranný (unbiased), pokud  $\vartheta = \mathbb{E}(\widehat{\Theta}_n)$  (pro každé  $\vartheta$ )
- asymptoticky nestranný (asymptotically unbiased) - pokud  $\vartheta = \lim_{n \rightarrow \infty} \mathbb{E}(\widehat{\Theta}_n)$
- konzistentní (consistent) - pokud  $\widehat{\Theta}_n \xrightarrow{P} \vartheta$
- vychýlení (bias)  $bias_{\vartheta}(\widehat{\Theta}_n) := \mathbb{E}(\widehat{\Theta}_n) - \vartheta$
- střední kvadratická chyba je  $MSE := \mathbb{E}((\widehat{\Theta}_n - \vartheta)^2)$

**Věta:**

$$MSE = bias_{\vartheta}(\widehat{\Theta}_n)^2 + var_{\vartheta}(\widehat{\Theta}_n)$$

**Důkaz: TODO**

□

**Věta** (Parametry výběrového momentu a rozptylu):

1.  $\overline{X}_n$  je konzistentní nestranný odhad  $\mu = \mathbb{E}X_1 = \mathbb{E}X_2 = \dots$
2.  $\overline{S}_n^2$  je konzistentní asymptoticky nestranný odhad  $\sigma^2$
3.  $\widehat{S}_n^2$  je konzistentní nestranný odhad  $\sigma^2$

**Důkaz:**

1.

$$\overline{X}_n = \frac{1}{n} X_1 + X_2 + \dots + X_n$$

$\overline{X}_i$  je nestranný, t. j.  $\mathbb{E}(\overline{X}_n) = \mu$

$$= \frac{1}{n} \mathbb{E}X_1 + \mathbb{E}X_2 + \dots + \mathbb{E}X_n = \frac{1}{n} \mu + \mu + \dots + \mu = \mu$$

$\overline{X}_n$  je konzistentní, t. j.  $\overline{X}_n \xrightarrow{P} \mu$  (slabý zákon velkých čísel)

$var(\overline{X}_n) = \frac{\sigma^2}{n} \dots$  Čebyšev

2.

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

$$\mathbb{E}\overline{S}_n = \mathbb{E} \frac{1}{n} \sum_{i=1}^n ((X_i - \mu) - (\overline{X}_n - \mu))^2$$

$$= \mathbb{E} \frac{1}{n} \sum [(X_i - \mu)^2 - 2(X_i - \mu)(\overline{X}_n - \mu) + (\overline{X}_n - \mu)^2]$$

$$= \mathbb{E} \frac{1}{n} \sum (X_i - \mu)^2 - \mathbb{E} \frac{2}{n} \sum (X_i - \mu)(\overline{X}_n - \mu) + \mathbb{E}(\overline{X}_n - \mu)^2$$

$$= \frac{1}{n} \sum \mathbb{E}(X_i - \mu)^2 - (\overline{X}_n - \mu)^2$$

$$= \sigma^2 - var(\overline{X}_n) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n} \sigma^2$$

3.

$$\widehat{S}_n^2 = \frac{n}{n-1} \overline{S}_n^2$$

$$\mathbb{E}\widehat{S}_n^2 = \frac{n}{n-1} \mathbb{E}\overline{S}_n^2 = \sigma^2 \gg \widehat{S}_n^2 \text{ je nestranný odhad.}$$

□

Je lepší  $\widehat{S}_n^2$  nebo  $\overline{S}_n^2$ ?

$\rightarrow \widehat{S}_n^2$  je nestranný,  $\overline{S}_n^2$  ne.

## 10.5 Metoda momentů

- $m_r(\vartheta) := \mathbb{E}(X^r)$  pro  $X \sim F_\vartheta \dots r$ -tý momentu
- $\widehat{m_r(\vartheta)} := \frac{1}{n} \sum_{i=1}^n X_i^r$  pro náhodný výběr  $X_1, \dots, X_n$  z  $F_\vartheta \dots r$ -tý výběrový moment

**Věta:**  $\widehat{m_r(\vartheta)}$  je nestranný konzistentní odhad pro  $m_r(\vartheta)$ .

**Důkaz:**

$$\mathbb{E}\widehat{m_r(\vartheta)} = \frac{1}{n} \sum \mathbb{E}(X_i^r) = \frac{1}{n} \sum \mathbb{E}(X_i^r) = m_r(\vartheta)$$

□

**Příklad:**  $X_1, \dots, X_n \sim \text{Bern}(p) \dots X_i = "i\text{-tý člověk je levák}"$

$\vartheta = p \in [0, 1]$

$m_1(\vartheta) = \mathbb{E}X_1 = \vartheta$

$\widehat{m_r(\vartheta)} = \frac{1}{n}(X_1 + \dots + X_n) = \overline{X}_n$

## 10.6 Metoda maximální věrohodnosti (maximal likelihood, ML)

- náh. výběr  $X = (X_1, \dots, X_n)$  z modelu s parametrem  $\vartheta$
- možný výsledek  $x = (x_1, \dots, x_n)$
- ...sdružená pravděpodobnostní funkce  $p_X(x; \vartheta)$
- ...sdružená hustota  $f_X(x; \vartheta)$
- věrohodnost (likelihood)  $L(x; \vartheta)$  značí  $p_X$  nebo  $f_X$
- normálně: máme pevné  $\vartheta$ , a  $L(x; \vartheta)$  je funkce  $x$
- teď: máme pevné  $x$  a  $L(x; \vartheta)$  je funkce  $\vartheta$
- Metoda MV (ML): volíme takové  $\vartheta$ , pro které je  $L(x; \vartheta)$  maximální
- definujeme také  $\ell(x; \vartheta) = \log(L(x; \vartheta))$
- díky nezávislosti je

$$L(x; \vartheta) = p(x_1; \vartheta)p(x_2; \vartheta) \dots p(x_n; \vartheta)$$

$$\ell(x; \vartheta) = \sum_{i=1}^n \log p(x_i; \vartheta)$$

$$0 = \ell'(x; \vartheta) = \sum_{i=1}^n \frac{1}{p(x_i; \vartheta)} \cdot p'(x_i; \vartheta)$$



# 11 Jedenáctá přednáška

## 11.1 Intervalové odhady

- místo jednoho čísla s nejistým významem vypočítáme z dat interval  $[\hat{\Theta}^-, \hat{\Theta}^+]$

**Definice** (Konfidenční interval): Necht  $\hat{\Theta}^-, \hat{\Theta}^+$  jsou n.v., které závisí na náhodném výběru  $X = (X_1, \dots, X_n)$  z distribuce  $F_\vartheta$ . Tyto n.v. určují intervalový odhad, též konfidenční interval o spolehlivosti  $1 - \alpha$  (*confidence interval*) pokud

$$P(\hat{\Theta}^- \leq \vartheta \leq \hat{\Theta}^+) \geq 1 - \alpha$$

- tohle jsou tzv. oboustranné odhady
- jednostranný odhad:  $[\hat{\Theta}^-, \infty)$  nebo  $(-\infty, \hat{\Theta}^+]$

**Věta:**  $X_1, \dots, X_n$  je náhodný výběr z  $N(\vartheta, \sigma^2)$ .

$\sigma$  známe,  $\vartheta$  chceme určit,  $\alpha \in (0, 1)$ .

Necht  $\Phi(z_{\alpha/2}) = 1 - \alpha/2$ . Zvolíme  $\hat{\Theta}_n := \hat{X}_n$ .

$$C_n := \left[ \hat{\Theta}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \hat{\Theta}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

Pak  $P(C_n \ni \vartheta) = 1 - \alpha$ .

**Důkaz:**

$$C_n \ni \vartheta \Leftrightarrow |\hat{\Theta}_n - \vartheta| \leq z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$\Leftrightarrow \left| \frac{\hat{\Theta}_n - \vartheta}{\sigma/\sqrt{n}} \right| \leq z_{\alpha/2}$$

$$\frac{\hat{\Theta}_n - \vartheta}{\sigma/\sqrt{n}} = Z \sim N(0, 1)$$

$$\begin{aligned} P(C_n \ni \vartheta) &= P(|Z| \leq z_{\alpha/2}) = \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) \\ &= (1 - \alpha/2) - (+\alpha/2) = 1 - \alpha \end{aligned}$$

□

**Věta:**  $X_1, \dots, X_n$  je náhodný výběr z rozdělení se střední hodnotou  $\vartheta$ , rozptylem  $\sigma^2$ .

$\sigma$  známe,  $\vartheta$  chceme určit,  $\alpha \in (0, 1)$ .

Necht  $\Phi(z_{\alpha/2}) = 1 - \alpha/2$ . Zvolíme  $\hat{\Theta}_n := \hat{X}_n$ .

$$C_n := \left[ \hat{\Theta}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \hat{\Theta}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

Pak  $\lim_{n \rightarrow \infty} P(C_n \ni \vartheta) = 1 - \alpha$ .

**Důkaz:** Centrální limitní věta.

□

**Definice** (Studentovo rozdělení):

- $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \dots$  výběrový průměr
- $\hat{S}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \dots$  výběrový rozptyl
- Necht  $X_1, \dots, X_n \sim N(\mu, \sigma^2)$
- Pak  $\frac{\bar{X}_n - \mu}{\hat{S}_n^2 / \sqrt{n}} \sim N(0, 1)$
- Studentovo  $t$ -rozdělení s  $n - 1$  stupni volnosti je rozdělení n.v.  $\frac{\bar{X}_n - \mu}{\hat{S}_n^2 / \sqrt{n}}$

- Distribuční funkci budeme značit  $\Psi_{n-1}$ . Je v tabulkách, v  $R : pt(x, n-1)$  **TODO**

**Věta:**  $X_1, \dots, X_n$  je náhodný výběr z  $N(\vartheta, \sigma^2)$ .  
 $\vartheta$  chceme určit,  $\sigma$  neznáme,  $\alpha \in (0, 1)$ . Nechť

$$\Psi_{n-1}(z_{\alpha/2}) = 1 - \alpha/2, \quad \hat{\Theta}_n = \hat{X}_n, \quad \hat{S}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

$$C_n := \left[ \hat{\Theta}_n - z_{\alpha/2} \frac{\hat{S}_n}{\sqrt{n}}, \hat{\Theta}_n + z_{\alpha/2} \frac{\hat{S}_n}{\sqrt{n}} \right]$$

Pak  $P(C_n \ni \vartheta) = 1 - \alpha$

**Důkaz:**

$$P(C_n \ni \vartheta) = P(|Z| \leq z_{\alpha/2}) = \Psi_{n-1}(z_{\alpha/2}) - \Psi_{n-1}(-z_{\alpha/2}) = 1 - \alpha/2 - \alpha/2 = 1 - \alpha$$

\* $Z =$  — st.  $t$  – rozdělení s  $n - 1$ .

□

## 11.2 Testování hypotéz

- Je naše mince spravedlivá?
- Je naše kostka spravedlivá?
- Má vylepšený program kratší dobu běhu než původní?
- Je léčba nemoci metodou X dobrá? (Lepší než placebo, lepší než metoda Y, ...)
- Jsou leváci lepší boxeři?
- dvě hypotézy:  $H_0, H_1$
- $H_0$  - nulová hypotéza - značí defaultní, konzervativní model (léčba, mince je spravedlivá)
- $H_1$  - alternativní hypotéza - značí alternativní model "pozoruhodnost"

**Příklad** (Testování hypotéz):

- Chceme testovat, zda je mince spravedlivá.
- Hodíme  $n$ -krát mincí, orel padne  $S$ -krát.
- Pokud je  $|S - n/2|$  moc velké, tak mince není spravedlivá.

## 12 Dvanáctá přednáška

### 12.1 Testování hypotéz - ilustrace

- Chceme testovat, zda je mince spravedlivá.
- $H_0$  : je spravedlivá - očekávaný stav světa
- $H_1$  : není spravedlivá - překvapivé zjištění
- Výsledky. zamítneme  $H_0$  / nezamítneme  $H_0$
- Chyba 1. druhu: chybné zamítnutí. Zamítneme  $H_0$ , i když platí. Trapas.
- Chyba 2. druhu: chybné přijetí. Nezamítneme  $H_0$ , ale ona neplatí. Promarněná příležitost.
- Potřebujeme určit  $k$  takové, že budeme zamítat  $H_0$  pokud **DOPLNIT**

- Vybereme vhodný statistický model.
- Volíme hladinu významnosti (significance level)  $\alpha$ : pravd. chybného zamítnutí  $H_0$ . Typicky  $\alpha = 0.05$ .
- Určíme testovou statistiku  $T = h(X_1, \dots, X_n)$ , kterou budeme určovat z naměřených dat.
- Určíme kritický obor (rejection region) - množinu  $W$ .
- Naměříme hodnoty  $x_1, \dots, x_n$  náh. veličin  $X_1, \dots, X_n$ .
- Rozhodovací pravidlo: zamítneme  $H_0$  pokud  $h(x_1, \dots, x_n) \in W$ .
- $\alpha = P(h(X) \in W; H_0)$
- $\beta = P(h(X) \notin W; H_1) \dots 1 - \beta$  je tzv. síla testu
- často  $\alpha$  nevolíme předem, ale spočítáme tzv.  $p$ -hodnotu: minimální  $\alpha$ , pro které bychom  $H_0$  zamítli.

**Příklad:** Měříme teplotu, chceme  $\mu = 5^\circ\text{C}$

- $X_1, \dots, X_n$  náhodný výběr z  $H(\vartheta, \sigma^2)$
- $\sigma^2$  známe,  $\mu$  dáno
- $H_0 : \vartheta = \mu, H_1 : \vartheta \neq \mu$

$$T = \frac{X_1 + \dots + X_n}{n} = \bar{X}_n \sim N(\vartheta, \sigma^2/n)$$

... víme ze vzorce pro rozptyl

$$S = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

vezmeme množinu:

$$W := \{s \in R : |s| > Z_{\alpha/2}\}$$

kde  $z_{\alpha/2} = \Phi^{-1}(1 - \frac{\alpha}{2})$ ... pro  $\alpha = 0.05$  dostaneme 1.96

**Příklad** (příklad dvojvýběrového testu):

- $X_1, \dots, X_{n_1}$  náhodný výběr z  $Ber(\vartheta_X)$
- $Y_1, \dots, Y_{n_2}$  náhodný výběr z  $Ber(\vartheta_Y)$
- $H_0 : \vartheta_X = \vartheta_Y \dots H_1 : \vartheta_X \neq \vartheta_Y$

Máme  $n_1$  lidí a jiných  $n_2$  lidí které léčíme různými metodami ( $H_0$  vs  $H_1$ )

$$\hat{\Theta}_X = \frac{X_1 + \dots + X_{n_1}}{n_1} \dots \text{odhad } \vartheta_X$$

$$\hat{\Theta}_Y = \frac{Y_1 + \dots + Y_{n_2}}{n_2} \dots \text{odhad } \vartheta_Y$$

$$Z := \hat{\Theta}_X - \hat{\Theta}_Y$$

$\hat{\Theta}_X, \hat{\Theta}_Y$  mají přibližně normální rozdělení (Centrální limitní věta)

Předpokládáme, že platí  $H_0$ :

$$\mathbb{E}\hat{\Theta}_X = \mathbb{E}\hat{\Theta}_Y \implies \mathbb{E}Z = 0$$

Víme, že  $Z$  je přibližně  $N(0, \sigma^2)$ ,  $\sigma^2$  neznáme

$$\sigma^2 = \text{var}(Z) = \text{var}(\hat{\Theta}_X) - \text{var}(\hat{\Theta}_Y) = \frac{\text{var}X_1}{n_1} + \frac{\text{var}Y_1}{n_2} = \frac{\vartheta_X(1 - \vartheta_X)}{n_1} + \frac{\vartheta_Y(1 - \vartheta_Y)}{n_1} = \vartheta \left( \frac{1}{n_1} + \frac{1}{n_2} \right)$$

$$\hat{\Theta} = \frac{\sum X_i + \sum Y_j}{n_1 + n_2} \dots \text{odhad } \vartheta \implies \hat{\sigma}^2 := \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \hat{\Theta}(1 - \hat{\Theta}) \implies T := \frac{\hat{\Theta}_X - \hat{\Theta}_Y}{\hat{\sigma}}$$

## 12.2 $p$ -hacking

- napřed získáme data, pak v nich hledáme zajímavosti
- když máme dost dat, tak tam nějaké budou "shodou okolností"
- reprodukovatelnost - po explorační analýze dat uděláme nezávislý sběr dat a ten analyzujeme konfirmačně
- nebo dopředu náhodně rozdělíme data na část pro tvorbu hypotéz a část pro jejich potvrzení ...jednoduchý případ křížové validace (cross validation)

## 12.3 $\chi_k^2$ - rozdělení $\chi$ -kvadrát

**Definice** ( $\chi_k^2$  - rozdělení  $\chi$ -kvadrát):  $Z_1, \dots, Z_k \sim N(0, 1)$  n.n.v. Rozdělení náhodné veličiny

$$Q = Z_1^2 + \dots + Z_k^2$$

se nazývá  $\chi$ -kvadrát s  $k$  stupni volnosti.

- $\mathbb{E}(Q) = k$
- $\text{var}(Q) = 2k$
- hustota jde napsat vzorcem, lze najít např. na Wikipedii
- $Q \doteq N(k, 2k)$  pro velká  $k$  (CLV)

## 12.4 Multinomické a kategoriální rozdělení

**Definice:** Dána  $p_1, \dots, p_k \geq 0$  a tak, že  $p_1 + p_2 + \dots + p_k = 1$ .

$n$ -krát zopakují pokus, kde může nastat jedna z  $k$  možností,  $i$ -tá má pravděpodobnost  $p_i$ .

$X_i :=$  kolikrát nastala  $i$ -tá možnost ( $X_1, \dots, X_k$ ) má multinomické rozdělení s parametry  $n, (p_1, \dots, p_k)$ .

- triviální případ:  $X_i =$  počet hodů kostkou, kdy padlo  $i$
- důležitý případ:  $X_i =$  počet výskytů  $i$ -tého písmene,  $i$ -tého slovního druhu, ...
- $P(X_1 = x_1, \dots, X_k = x_k) = \binom{n}{x_1, \dots, x_k} p_1^{x_1} \dots p_k^{x_k}$

**Definice** (Pearsonova  $\chi^2$  Statistika):

- $(X_1, \dots, X_k)$  - multinomické rozdělení s parametry  $n, (p_1, \dots, p_k)$  jako minule
- $\mathbb{E}_i = \mathbb{E}(X_i) = np_i$
- Pearsonova  $\chi^2$  statistika je funkce

$$\chi^2 = T := \sum_{i=1}^k \frac{(X_i - E_i)^2}{E_i}$$

**Věta:**  $T \xrightarrow{d} \chi_{k-1}^2$

**Důkaz:** pro  $k = 2$

$$\begin{aligned} X_1 + X_2 &= n, p_1 + p_2 = 1, E_i = np_i \\ T &= \frac{(X_1 - E_1)^2}{E_1} + \dots = \frac{(X_1 - np_1)^2 + (p_1 + p_2)}{np_1 p_2} \\ &= \left( \frac{X_1 - np_1}{\sqrt{np - (1 - p_1)}} \right) \end{aligned}$$

□

## 12.5 Test dobré shody (goodness of fit)

- $(X_1, \dots, X_k)$  - multinomické rozdělení s parametry  $n, \vartheta = (\vartheta_1, \dots, \vartheta_k)$  jako minule
- $n$  známe,  $\varphi$  neznáme.
- Hypotéza  $H_0 : \vartheta = \vartheta^*$
- $E_i := n\vartheta_i^*$  pro všechna  $i$
- Použijeme statistiku  $\chi^2 = T := \sum_{i=1}^k \frac{(X_i - E_i)^2}{E_i}$
- Hypotézu  $H_0$  zamítneme, pokud  $T > \gamma$
- $\gamma := F_Q^{-1}(1 - \alpha)$ , kde  $Q \sim \chi_{k-1}^2$
- $P(\text{chyba prvního druhu}) = P(T > \gamma; H_0) \rightarrow P(Q > \gamma) = \alpha$

**Příklad** (Test dobré shody - házíme kostkou):

- Házíme opakovaně kostkou. Jednotlivá čísla padla s četností 92, 120, 88, 98, 95 a 107.
- Je kostka spravedlivá?

$$n = 92 + 120 + \dots = 600$$

$$\vartheta^* = \left(\frac{1}{6}, \dots, \frac{1}{6}\right), \quad E_i = n\frac{1}{6} = 100$$

$$T = \sum_{i=1}^6 \frac{(X_i - 100)^2}{100} = \dots = \frac{(80^2 + 20^2 + 12^2 + 2^2 + 5^2 + 7^2)}{100} = 6.86$$

$$Q \sim \chi_5^2 \dots F_Q^{-1}(1 - \alpha = 0.95) = 11.1$$

kdyby nám  $T$  vyšlo víc než 11.1, můžeme říct, že kostka je nespravedlivá.

$$p - \text{hodnota} : 1 - F_Q(6.86) \doteq 1 - 0.77 = 0.23$$

zhruba ve čtvrtině hodů najdeme extremnější odchylku.

Další rozšíření

- Pro zkoumání rozdělení libovolné n.v.  $Y$  můžeme vybrat "příhrádky"  $B_1, \dots, B_k$  (rozklad  $\mathcal{R}$ ) a zkoumat, kolikrát je  $Y \in B_i$
- Obdobný test pro nezávislost (diskrétních) náhodných veličin

**Definice** (Lineární regrese):

- data:  $(x_i, y_i)$  pro  $i = 1, \dots, n$
- TODO

## 13 Třináctá přednáška

### 13.1 Simpsonův paradox

**DOPLNIT (tabulka + graf)** Problém s tím, jestli jsou naměřená data (skupiny dat) dostatečně homogenní

## 13.2 Permutační test

**Příklad:**

- Máme k dispozici dvě sady nezávislých náhodných veličin (náhodné výběry):
- $X_1, \dots, X_n \sim F_X$  a  $Y_1, \dots, Y_m \sim F_Y$
- Chceme rozhodnout, zda platí  $H_0 : F_X = F_Y$  nebo  $H_1 : F_X \neq F_Y$
- Příklady: doba běhu programu před/po vylepšení, hladina cholesterolu u lidí co jedí/nejedí Zá-zrčnou Superpotravu™, frekvenci
- **TODO**

**Postup:**

- Zvolíme vhodnou statistiku, například.

$$T(X_1, \dots, X_n, Y_1, \dots, Y_m) = |\bar{X}_n - \bar{Y}_m|$$

- $t_{\text{obs}} := T(X_1, \dots, X_n, Y_1, \dots, Y_m)$
- Za předpokladu  $H_0$  jsou "všechny permutace stejné":  $X_i$  i  $Y_j$  se generovaly ze stejného rozdělení.
- Náhodně zpermutujeme zadaných  $m + n$  čísel a pro každou permutaci vyčíslíme  $T$  - dostaneme čísla  $T_1, \dots, T_{(m+n)!}$  (každé stejně pravděpodobné).
- Jako  $p$ -hodnotu vezmeme pravděpodobnost, že  $T > t_{\text{obs}}$ , neboli

$$p = \frac{1}{(m+n)!} \sum_j I(T_j > t_{\text{obs}}).$$

- To je pravděpodobnost chyby 1. druhu, neboli  $H_0$  zamítneme, pokud je  $p < \alpha$  (pro naši zvolenou hodnotu  $\alpha$ , např.  $\alpha = 0.05$ ).

**Vylepšení:**

- Zkoušet všechny permutace může trvat moc dlouho. Vezmeme tedy jen vhodný počet  $B$  nezávisle náhodně vygenerovaných permutací a spočítáme jenom  $B$  hodnot  $T_1, \dots, T_B$ .
- Jako  $p$ -hodnotu vezmeme odhad pravděpodobnosti, že  $T > t_{\text{obs}}$
- **DOPLNIT**

## 13.3 Bootstrap

**Příklad:** Základní idea

- z naměřených dat  $X_1 = x_1, \dots, X_n = x_n \sim F$  vytvoříme  $\hat{F}_n$
- další data můžeme samplovat z  $\hat{F}_n$
- to se dělá tak, že vybereme uniformně náhodné **DOPLNIT**

**Základní použití**

- $T_n = g(X_1, \dots, X_n)$  nějaká statistika (funkce dat)
- chceme odhadnout  $\text{var} T_n$
- nasamplujeme  $X_1^*, \dots, X_n^* \sim \hat{F}_n$  (viz minulá strana)
- spočteme  $T_n^* = g(X_1^*, \dots, X_n^*)$
- opakujeme  $B$ -krát, dostaneme  $T_{n,1}^*, \dots, T_{n,B}^*$

- odhad rozptylu:

$$\frac{1}{B} \sum_{b=1}^B \left( T_{n,b}^* - \frac{1}{B} \sum_{k=1}^B T_{n,k}^* \right)^2$$

## 13.4 Bayesovská statistika

### 13.4.1 Frekventistický/klasický přístup

- Pravděpodobnost je dlouhodobá frekvence (z 6000 hodů kostkou padla šestka 1026krát). Je to objektivní vlastnost reálného světa.
- Parametry jsou pevné, neznáme konstanty. Nelze o nich říkat smysluplné pravděpodobnostní výroky.
- Navrhujeme statistické procedury tak, aby měly žádané dlouhodobé vlastnosti. Např. 95% z našich intervalových odhadů pokryje neznámý parametr.

### 13.4.2 Bayesovský přístup

- Pravděpodobnost popisuje, jak moc věříme nějakému jevu, jak moc jsme ochotní se vsadit. (Pravděpodobnost, že Thomas Bayes měl 18. prosince 1760 šálek čaje je 90%.)
- Můžeme vyslovovat pravděpodobnostní výroky i o parametrech (třebaže jsou to pevné konstanty).
- Spočítáme distribuci  $\vartheta$  a z ní tvoříme bodové a intervalové odhady, atd.

Bayesovská metoda - základní popis

- neznámý parametr považujeme za náhodnou veličinu  $\Theta$ .
- zvolíme apriorní distribuci (prior distribution), neboli hustotu pravděpodobnosti  $f_{\Theta}(\vartheta)$  nezávislou na datech.
- zvolíme statistický model  $F_{X|\Theta}(x|\vartheta)$ , který popisuje co naměříme (s jakou pravděpodobností), v závislosti na hodnotě parametru
- poté, co pozorujeme hodnotu  $X = x$ , spočítáme posteriorní distribuci (posterior distribution)  $f_{\Theta|X}(\vartheta|x)$
- z té pak odvodíme co potřebujeme, např. najdeme  $a, b$  tak, aby

$$\int_a^b f_{\Theta|X}(\vartheta|x) d\vartheta \geq 1 - \alpha$$

- $\vartheta = \theta$  malá théta,  $\Theta$  je velká théta

**Věta** (Bayesova věta pro diskrétní náhodné veličiny):  $X, \Theta$  jsou diskrétní n.v.

$$p_{\Theta|X}(\vartheta|x) = \frac{p_{X|\Theta}(x|\vartheta)p_{\Theta}(\vartheta)}{\sum_{\vartheta' \in I_{m\Theta}} p_{X|\Theta}(x|\vartheta')p_{\Theta}(\vartheta')}.$$

(sčítance s  $p_{\Theta}(\vartheta') = 0$  považujeme za 0).

**Věta** (Bayesova věta pro spojité náhodné veličiny):  $X, \Theta$  jsou spojité n.v., které mají hustotu  $f_X, f_{\Theta}$  i sdruženou hustotu  $f_{X,\Theta}$

$$p_{\Theta|X}(\vartheta|x) = \frac{f_{X|\Theta}(x|\vartheta)f_{\Theta}(\vartheta)}{\int_{\vartheta' \in I_{m\Theta}} f_{X|\Theta}(x|\vartheta')f_{\Theta}(\vartheta')d\vartheta'}.$$

Bayesovské bodové odhady - MAP a LMS

- MAP - Maximum A-Posteriori  
Volíme  $\hat{\vartheta}$  tak, aby maximalizovalo
  - $p_{\Theta|X}(\vartheta|x)$  v diskrétním případě
  - $f_{\Theta|X}(\vartheta|x)$  v spojitém případě
  - Podobné metodě  $ML$  v klasickém přístupu, pokud bychom volili "flat prior"- uniformní  $p_{\Theta}(\vartheta)$ .
- LMS - Least Mean Square  
Těž metoda podmíněné střední hodnoty
  - Volíme  $\hat{\vartheta} = \mathbb{E}(\Theta|X = x)$ .
  - Nestranný bodový odhad, má nejmenší možnou hodnotu LMS:  $\mathbb{E}((\Theta - \hat{\vartheta})^2|X = x)$ .

**Příklad** (Bayesovský klasifikátor spamů):

- vytvoříme seznam podezřelých slov (money, win, pharmacy, ...)
- n.v.  $X_i$  (0 nebo 1) popisuje, zda email obsahuje podezřelé slovo  $w_i$ .
- n.v.  $\Theta$  popisuje, zda email je spam  $\Theta = 1$  nebo ne  $\Theta = 0$ .
- Z předchozích emailů získáme odhady  $p_{X\Theta}, p_{\Theta}$
- Použijeme Bayesovu větu na výpočet  $p_{\Theta|X}$

**Příklad:** Romeo a Julie se mají sejít přesně v poledne. Julie ale přijde pozdě o dobu popsanou náhodnou veličinou  $X \sim U(0, \vartheta)$ . Parametr  $\vartheta$  modelujeme náhodnou veličinou  $\Theta \sim U(0, 1)$ . Co z naměřené hodnoty  $X = x$  usoudíme o  $\vartheta$ ?

**Doplnit řešení**

## 13.5 Generování náhodných veličin

The End