Matematika IV – 7. přednáška
Pravděpodobnost – opakování a zobecnění pojmů

Michal Bulant

Masarykova univerzita
Fakulta informatiky

3. 4. 2013

Obsah přednášky

Doporučené zdroje

- Martin Panák, Jan Slovák, **Drsná matematika**, *e*-text.
- Karel Zvára, Josef Štěpán, **Pravděpodobnost a matematická statistika**, Matfyzpress, 4. vydání, 2006, 230 stran, ISBN 80-867-3271-1.
- Marie Budíková, Štěpán Mikoláš, Pavel Osecký, **Teorie pravděpodobnosti a matematická statistika (sbírka příkladů)**, Masarykova univerzita, 3. vydání, 2004, 117 stran, ISBN 80-210-3313-4.
- Marie Budíková, **Statistika**, Masarykova univerzita, 2004, distanční studijní opora ESF, `http://www.math.muni.cz/~budikova/esf/Statistika.zip`.
- Marie Budíková, Tomáš Lerch, Štěpán Mikoláš, **Základní statistické metody**, Masarykova univerzita, 2005, 170 stran, ISBN 80-210-3886-1.

Motto:

42,35 procenta všech statistik je nesmyslných.

Statistika v širším slova smyslu je jakékoliv zpracování číselných dat o nějakém souboru objektů a jejich více či méně přehledná prezentace.

Podstatou **matematické statistiky** je pro daná data zjišťovat, jaké vlastnosti mají objekty, které jsou daty popisovány. Zpravidla jde o sběr dat o části souboru objektů, jejich následnou analýzu a konečně o vyslovení důsledků pozorování pro celý soubor.
Výsledkem práce matematického statistika je sdělení o velkém souboru objektů na základě studia malé (zpravidla náhodně vybrané) části z nich **společně s kvalitativním odhadem věrohodnosti výsledného sdělení.**
Teorie pravděpodobnosti studuje modely popisující chování abstraktních souborů (pravděpodobnost jevů z jevového pole), statistika studuje skutečné náhodné výběry z nějakého základního souboru a zdůvodňuje výběr teoretického pravděpodobnostního modelu, resp. kvalitativní informace o jeho parametrech.

Příklad

Za soubor objektů vezměme všechny studenty přednášky Matematika III (podzim 2007), jako číselný údaj můžeme uvažovat

1. průměrné bodové hodnocení studenta u zkoušky,
2. průměrnou známku u zkoušky z tohoto (2,92) a z jiných pevně vybraných předmětů (`IB000` – 2,95; `IB102` – 2,89) ,
3. nejčastější známku (resp. úspěšnou známku) z tohoto předmětu (`F` – 92 krát, `E` – 91 krát), nejméně častou známku (`B` – 15 krát),
4. průměrný počet bodů dosažených na jednotlivých termínech zkoušky (1. – 16,8; 2. – 8,9; 3. – 8,1; příklad, za nějž bylo uděleno nejvíce (nejméně) procent možných bodů – min. kostra (1B, 82,5%), resp. rekurence (2A, 3,6%)
5. počet pracovních hodin týdně odpracovaných mimo fakultu,
6. číselná data vypovídající o historii dřívějšího studia

a mnoho dalších údajů.

Zastavme se u prvního údaje. Samotný aritmetický průměr bodů nám mnoho neřekne nejen o kvalitě přednášky a o kvalitě přednášejícího, ale ani o samotném hodnocení. Zajímá nás také hodnota, která bude „uprostřed souboru“, tj. počet bodů, pro které je stejně studentů pod ní a nad ní.
Obdobně první a poslední čtvrtina, desetina apod. Všem takovým údajům říkáme **statistiky** posuzované veličiny. V uvedených příkladech se jim říká **medián**, **kvartil**, **decil** apod.
Z obecné zkušenosti nebo jako výsledek úvah mimo matematiku víme, že rozumné hodnocení by mělo mít tzv. **normální rozdělení** (odpovídá tzv. *Gaussově křivce*). Tento pojem patří do teorie pravděpodobnosti a k jeho zavedení budeme potřebovat poměrně dost matematiky.
Porovnáním výsledku třeba i docela malého náhodného výběru studentů s teoretickým modelem můžeme zjistit odhad parametrů takového rozdělení a činit závěry, zda je hodnocení „rozumné“. Zároveň lze popsat věrohodnost našich závěrů.

Daleko zajímavější vývody ovšem můžeme činit, když porovnáním statistik pro různé veličiny budeme moci dovozovat informace o souvislostech (*korelace*, *závislost*). Pokud např. neexistuje žádná doložitelná souvislost mezi historií předchozího studia a výsledky v dané přednášce, je jedním z možných vysvětlení závěr, že je přednáška (nebo její hodnocení) prostě špatná.

Závěr úvodních úvah:

- V matematice pracujeme s abstraktním matematickým popisem pravděpodobnosti.
- Vývody pro konktrétní soubory dat, pro které je zvolený model relevantní, dává matematická statistika.
- To, zda je takový popis adekvátní pro konkrétní výběr dat, je také možné podpořit nebo zavrhnout pomocí metod matematické statistiky.

Připomeneme (a trochu zobecníme) pojmy a výsledky z prvního semestru.

Definice (Náhodné jevy)

Budeme pracovat s neprázdnou pevně zvolenou množinou Ω všech možných výsledků, kterou nazýváme **základní prostor**.
Prvky $\omega \in \Omega$ představují jednotlivé **možné výsledky**.
Systém (ne nutně všech) podmnožin $\mathcal{A}$ základního prostoru se nazývá **jevové pole** a jeho prvky se nazývají **jevy**, jestliže

- $\Omega \in \mathcal{A}$, tj. základní prostor, je jevem,
- je-li $A, B \in \mathcal{A}$, pak $A \setminus B \in \mathcal{A}$, tj. pro každé dva jevy je jevem i jejich množinový rozdíl,
- je-li $A_i \in \mathcal{A}$, $i \in I$ nejvýše spočetný systém jevů, pak také jejich sjednocení je jevem, tj. $\cup_{i \in I} A_i \in \mathcal{A}$.

Důsledek

- *Komplement* $A^c = \Omega \setminus A$ *jevu* A *je jevem, který nazýváme* opačný jev *k jevu* A.
- *Průnik dvou jevů opět jevem, protože pro každé dvě podmnožiny* $A, B \subset \Omega$ *platí*

$$A \setminus (\Omega \setminus B) = A \cap B.$$

Takový systém množin $\mathcal{A}$ se pak nazývá σ-algebra.
Jevové pole je tedy systém podmnožin základního prostoru uzavřený na konečné průniky, spočetná sjednocení a množinové rozdíly. Jednotlivé množiny $A \in \mathcal{A}$ nazýváme **náhodné jevy** (vzhledem k $\mathcal{A}$).

Terminologie připomíná souvislosti s popisem skutečných jevů a jejich statistickým popisem:

- celý základní prostor Ω se nazývá **jistý jev**, prázdná podmnožina $\emptyset \in \mathcal{A}$ se nazývá **nemožný jev**,
- jednoprvkové podmnožiny $\{\omega\} \in \Omega$ se nazývají **elementární jevy**,
- **společné nastoupení jevů** A_i, $i \in I$, odpovídá jevu $\cap_{i \in I} A_i$, **nastoupení alespoň jednoho z jevů** A_i, $i \in I$, odpovídá jevu $\cup_{i \in I} A_i$,
- $A, B \in \mathcal{A}$ jsou **neslučitelné jevy**, je-li $A \cap B = \emptyset$,
- jev A má za **důsledek** jev B, když $A \subset B$,
- je-li $A \in \mathcal{A}$, pak se jev $B = \Omega \setminus A$ nazývá **opačný jev k jevu** A, píšeme $B = A^c$.

Definice (Kolmogorovova definice pravděpodobnosti)

Pravděpodobnostní prostor je jevové pole $\mathcal{A}$ podmnožin (konečného) základního prostoru Ω, na kterém je definována funkce $P : \mathcal{A} \to \mathbb{R}$ s následujícími vlastnosti:

- je nezáporná, tj. $P(A) \geq 0$ pro všechny jevy A,
- je aditivní, tj. $P(\cup_{i \in I} A_i) = \sum_{i \in I} P(A_i)$, pro každý nejvýše spočetný systém po dvou neslučitelných jevů,
- pravděpodobnost jistého jevu je 1.

Funkci P nazýváme **pravděpodobností** na jevovém poli $(\Omega, \mathcal{A})$.

Důsledek

Pro všechny jevy $A, B \in \mathcal{A}$ platí

- $P(\emptyset) = 0,\ 0 \leq P(A) \leq 1,$
- $P(A^c) = 1 - P(A),$
- $A \subseteq B \implies P(A) \leq P(B),\ P(B \setminus A) = P(B) - P(A),$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Podobná tvrzení platí i pro nekonečné posloupnosti jevů:

Tvrzení

Pro libovolnou nejvýše spočetnou množinu jevů $(A_i)_{i=1}^{\infty}$ *platí:*

- *Je-li* $A_1 \subseteq A_2 \subseteq \cdots$, *pak*

$$P(\bigcup_{i=1}^{\infty} A_i) = \lim_{i\to\infty} P(A_i),$$

- *Je-li* $A_1 \supseteq A_2 \supseteq \cdots$, *pak*

$$P(\bigcap_{i=1}^{\infty} A_i) = \lim_{i\to\infty} P(A_i),$$

- $P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$,
- $P(\bigcap_{i=1}^{\infty} A_i) \geq 1 - \sum_{i=1}^{\infty}(1 - P(A_i))$.

Klasická pravděpodobnost

Připomeňme si klasickou konečnou pravděpodobnost.

Definice

Nechť Ω je konečný základní prostor a nechť jevové pole $\mathcal{A}$ je právě systém všech podmnožin v Ω. **Klasická pravděpodobnost** je pravděpodobnostní prostor $(\Omega, \mathcal{A}, P)$ s pravděpodobnostní funkcí $P : \mathcal{A} \to \mathbb{R}$,

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

Zjevně takto zadaná funkce skutečně definuje pravděpodobnost, kdy všem elementárním jevům přiřazujeme stejnou pravděpodobnost.

Že s klasickou pravděpodobností nevystačíme, ukazují následující příklady:

Příklad

- Cestou z Kotlářské na Botanickou jsem ztratil zadání písemky. Určete pravděpodobnost jevu ω_X slovně vyjádřeného: *ztracená písemka se nachází nejblíže k zastávce trolejbusu X*.
- Určete pravděpodobnost, jevu ω_k: *při opakovaném hodu mincí padne hlava poprvé při k-tém pokusu*.

V prvním případě je třeba pracovat s nekonečně mnoha stejně pravděpodobnými elementárními jevy: *písemku jsem ztratil v bodě* (x, y) , ve druhém pak musíme připustit teoretickou možnost, že hlava nepadne nikdy, a prostorem jevů tedy bude $\mathbb{N} \cup \{\infty\}$.

Peterburgský „paradox" (Bernoulli, 1738)

Typický příklad klasické pravděpodobnosti jsou jevy související s házením mincí. Představme si následující pravidla kasina:

Casino rules

Návštěvník zaplatí vklad C a poté hází mincí. V banku je na začátku dolar a při každém hodu se bank zdvojnásobí. Padne-li hlava, hráč získá obsah banku. Je-li tedy T počet hodů potřebných k první hlavě, hráč obdrží výhru 2^T. Jaká je „fér hodnota" pro vklad C?

A co vy? Zaplatili byste za možnost zahrát si tuto hru třeba 20$?

Odvození

Pravděpodobnost, že padne hlava je u férové mince $1/2$, je proto $P(T = k) = 2^{-k}$. Sečteme-li všechny pravděpodobnosti výsledků vynásobených výhrami 2^k, dostaneme očekávanou výhru

$$\frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 2^2 + \cdots = \sum_{1}^{\infty} 1 = \infty.$$

Zdá se proto, že se vyplatí vložit i velký vklad, protože libovolný vklad C se nám „časem“ vrátí.
Ve skutečnosti simulací hry zjistíme, že nezávisle na počtu pokusů se prakticky všechny výhry budou pohybovat v rozmezí malých hodnot. Důvodem je, že vysoké výhry jsou velice nepravděpodobné a proto je při reálných úvahách nelze brát vážně.
Tento paradox je vysvětlován nelinearitou funkce *užitečnosti peněz* (utility function), případně nezbytností diskontování jejich hodnoty.

Podmíněná pravděpodobnost a nezávislost

Motto:

Je dokázáno, že slavení narozenin je zdraví prospěšné. Statistika ukazuje, že lidé, kteří oslavili nejvíce narozenin, se dožívají nejvyššího věku.

Obvyklé je také klást dotazy s dodatečnou podmínkou. Např.

- Jaká je pravděpodobnost, že při hodu dvěma kostkami padly dvě pětky, je-li součet hodnot deset?
- Mějme urnu s 10 koulemi. Desetkrát jsem vytáhl kouli, zkontroloval její barvu a vrátil do urny. Jestliže byla vždy bílé barvy, s jakou pravděpodobností jsou všechny koule v urně bílé?
- Na dostizích jsou známy pravděpodobnosti vítězství jednotlivých koní. Jak se tyto pravděpodobnosti změní, pokud uprostřed závodu spadne jezdec jednoho z koní ze sedla?

Připomeňme, že formalizovat takové úvahy umíme následovně.

Definice

Nechť H je jev s nenulovou pravděpodobností v jevovém poli $\mathcal{A}$ v pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$. **Podmíněná pravděpodobnost** $P(A|H)$ jevu $A \in \mathcal{A}$ vzhledem k jevu H je definována vztahem

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

Přirozená definice nezávislosti je, že hypotéza H a jev A jsou nezávislé tehdy, je-li $P(A) = P(A|H)$.
Z výše uvedeného snadno vyplývá *symetričtější* definice:

Definice

Říkáme, že jevy A a B jsou nezávislé, jestliže

$$P(A \cap B) = P(A)P(B).$$

Definice

Říkáme, že jevy $A_1, A_2, \ldots$ jsou nezávislé, jestliže pro každou k-tici $A_{i_1}, \cdots, A_{i_k}$ z nich platí

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P(A_{i_j}).$$

Příklad

V urně jsou 4 lístky označené 000, 110, 101, 011. Uvažujme pro $i = 1, 2, 3$ náhodné jevy
$A_i = \{\text{náhodně vytažený lístek má na } i\text{-tém místě } 1\}$.
Snadno se vidí, že $P(A_1) = P(A_2) = P(A_3) = \frac{1}{2}$, dále, že $P(A_1 \cap A_2) = P(A_1 \cap A_3) = P(A_2 \cap A_3) = \frac{1}{4}$ a že $P(A_1 \cap A_2 \cap A_3) = 0$. Jevy A_1, A_2, A_3 jsou tedy po dvou nezávislé, ale nejsou nezávislé.

Bayesovy věty

Přepsáním formule pro podmíněnou pravděpodobnost dostáváme

$$P(A \cap B) = P(B \cap A) = P(A)P(B|A) = P(B)P(A|B).$$

Věta (Bayesovy věty)

Pro pravděpodobnost jevů A a B platí

1. $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$.
2. $P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A)+P(A^c)P(B|A^c)}$.

Důkaz.

První tvrzení je přepsáním předchozí formule, druhé z prvého plyne dosazením $P(B) = P(A)P(B|A) + P(A^c)P(B|A^c)$. □

Specifičnost a senzitivita (citlivost) testu

	Pozitivní skutečnost	Negativní skutečnost
Test pozitivní	True positive	**False positive**
Test negativní	**False negative**	True negative
	Senzitivita	Specifičnost

Příklad – preventivní screening

Předpokládejme, že krevní test na HIV pozitivní osoby má 99% správnost v případě osoby skutečně HIV pozitivní (*vysoká citilivost – sensitivity*). Zároveň předpokládejme, že u HIV negativní osoby dopadně test pozitivně v 0,2% případů (*relativně vysoká specifičnost – specificity*).
Náhodně z populace vybereme osobu a otestujeme pozitivně.
S jakou pravděpodobností je skutečně HIV pozitivní, jestliže četnost výskytu HIV v populaci je p promile (tj. p osob z tisíce je skutečně HIV pozitivní).
Označme A jev, že je daná osoba HIV pozitivní, a B jev, že daná osoba má pozitivní test. Dle druhé Bayesovy věty je hledaná pravděpodobnost

$$P(A|B) = \frac{p/1000 \cdot 99/100}{p/1000 \cdot 99/100 + (1000 - p)/1000 \cdot 2/1000}$$

Příklad – preventivní screening, pokr.

Jestliže zvolíme za p nějaké konkrétní četnosti, dostaneme příslušné očekávatelné spolehlivosti testu. V následující tabulce je spočten výsledek pro několik p:

p	100	10	1	0,1
$P(A\|B)$	0,982	0,8333	0,3313	0,0471

Výsledek asi neodpovídá naší intuici a může se zdát šokující ve vztahu k použití takovýchto testů.

Poznámka

Sami si můžete podobný výpočet udělat pro tzv. triple test na Downův syndrom, prováděný ve 2. trimestru těhotenství s 70% citlivostí a 5% „false-positive rate“ či pro statistiky svého oblíbeného spamfilteru (např. SpamAssassin s někde udávanou citlivostí 99,64% a specifičností 98.23%).

Triple test a jeho výsledky

Triple test je vyšetření krevního séra na hodnoty choriogonadotropinu, estriolu a alfa-fetoproteinu. Provádí se v druhém trimestru těhotenství a má sloužit k detekci rizik genetických poruch a poruch vývoje nervové trubice. Detekuje poruchy s úspěšností **70%** a naopak **5%** zdravých případů rozpozná jako porušené. Budoucím matkám, u kterých triple test ukáže zvýšené riziko vad plodu, je obvykle doporučeno nějaké další zpřesňující vyšetření, například amniocentéza (odběr plodové vody). Uvádí se, že u těhotné ženy ve věku 20–24 let je pravděpodobnost narození dítěte s Downovým syndromem cca **1:1500**, u těhotné ženy ve věku 35–39 let je pravděpodobnost narození dítěte s Downovým syndromem cca **1:200**. Prozkoumejme (alespoň z matematického hlediska) význam provádění tohoto testu za uvedených předpokladů, kdy se rodí cca 100 tis. dětí ročně, z toho cca 10% ženám ve věku 35–39 let a cca 12% ženám ve věku 20-24 let.

Specifičnost a senzitivita (citlivost) testu

	Pozitivní skutečnost	Negativní skutečnost
Test pozitivní	True positive	**False positive**
Test negativní	**False negative**	True negative
	Senzitivita	Specifičnost

Triple test	Pozitivní skutečnost	Negativní skutečnost
Test pozitivní	70%	5%
Test negativní	30%	95%
	Senzitivita	Specifičnost

Za dříve uvedených předpokladů snadno vypočteme, že pravděpodobnost, že dítě „starší“ matky bude skutečně postiženo Downovým syndromem, pokud vyšel pozitivní test, je pouhých cca 6,6%. U mladých žen se pak tato pravděpodobnost pohybuje kolem 0,9% a je tedy na zváženou, zda toto plošné testování v dané věkové skupině provádět, pokud navíc uváděné riziko potratu při případné amniocentéze se pohybuje kolem jednoho promile.

Výpočet

Uvažujme (reálný) vzorek deseti tisíc žen ve věku 35–39 let:

Starší ženy	Pozitivní skutečnost	Negativní skutečnost	
Test pozitivní	35	497,5	532,5
Test negativní	15	9452,5	9467,5
	50	9950	

Proto lze pravděpodobnost, že dítě „starší“ matky bude skutečně postiženo Downovým syndromem, pokud vyšel pozitivní test, spočítat jako $\frac{35}{532{,}5} \approx 6{,}6\%$. A pro 12 tis. žen ve věku 20–24 let dostaneme:

Mladší ženy	Pozitivní skutečnost	Negativní skutečnost	
Test pozitivní	5,6	599,6	605,2
Test negativní	2,4	11392,4	11394,8
	8	11992	

Pravděpodobnost, že dítě „mladší“ matky bude skutečně postiženo

Evidentně prostý výběr náhodné osoby a použití jediného testu, byť velmi citlivého a specifického, nejsou vhodné ani na otestování skutečného stavu populace, ani na preventivní vyšetření jednotlivců, pokud nemáme další podpůrné informace a lepší nástroje.
Právě matematická statistika dává nástroje na kvalifikovanější postupy v medicínské i průmyslové diagnostice, ekonomických modelech, vyhodnocování experimentálních dat atd.

Vraťme se k jednoduchému a názornému příkladu statistik kolem výsledků studentů v daném předmětu, který je a není podobný klasické pravděpodobnosti a s ní související statistice při házení kostkou.
Na jedné straně jsme připustili pouze konečný počet možných bodových hodnocení (v tomto případě celá čísla od 0 do 30), zároveň ale není patrně vhodné představovat si výsledky jednotlivých studentů jako analogii nezávislého házení kostkou (to by byla skutečně divně vedená přednáška).
Místo toho máme na základním prostoru Ω všech studentů definovánu funkci bodového ohodnocení $X : \Omega \to \mathbb{R}$. Je to typický příklad **náhodné veličiny**.
U každé náhodné veličiny potřebujeme umět pracovat s vhodnou množinou jevů. Zpravidla požadujeme, abychom mohli pracovat s pravděpodobnostmi příslušnosti hodnoty X do předem zadaného intervalu.

Přirozenější interpretací výsledku pokusu je totiž často spíše než zjištění, zda náhodný jev *nastal či nenastal*, nějaká hodnota:

- součet bodů na dvou kostkách,
- počet bakterií v daném množství roztoku nebo
- počet studentů, kteří uspěli u zkoušky nebo kteří získali alespoň 5 bodů z konkrétního příkladu.

Od pravděpodobnostního prostoru $(\Omega, \mathcal{A}, P)$ tedy potřebujeme přejít k obdobné dvojici $(\mathbb{R}, \mathcal{B})$ tak, abychom podmnožinám $\mathbb{R}$, ležícím v σ-algebře $\mathcal{B}$ byli schopni přiřadit pravděpodobnost odvozenou z $(\Omega, \mathcal{A}, P)$.