Matematika III – 10. přednáška
Náhodné veličiny – základní vlastnosti a typy

Michal Bulant

Masarykova univerzita
Fakulta informatiky

20. 11. 2013

Obsah přednášky

Doporučené zdroje

- Jan Slovák, Martin Panák, Michal Bulant, *Matematika drsně a svižně*, MU Brno, 2013, 774 s. (též jako 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.

Podmíněná pravděpodobnost

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 $P(B) = P(A)P(B|A) + P(A^c)P(B|A^c)$. □

Příklad

a) Z urny, v níž je a bílých a b černých koulí, vybereme postupně (bez vracení) dvě koule. Jaká je pravděpodobnost. že druhá koule je bílá, za předpokladu, že první byla bílá.

b) Ze skupiny 100 výrobků, která obsahuje 10 zmetků, vybereme náhodně bez vracení 3 výrobky. Určete pravděpodobnost, že:
   - třetí je zmetek za podmínky, že první 2 byly kvalitní.
   - první 2 jsou kvalitní a třetí zmetek.

Příklad

Dva střelci vystřelí nezávisle na sobě do téhož terče každý jednu ránu. Po střelbě byl v teči nalezen 1 zásah. Určete pravděpodobnost, že zásah patří 1. střelci, pokud tento trefuje terče s pravděpodobností 0,8, zatímco druhý střelec s pravděpodobností 0,4.

Příklady k procvičení

Příklad

V testu jsou u každé otázky 4 možné odpovědi. Pokud student nezná odpověď, tak hádá (uhodne s pravděpodobností $\frac{1}{4}$). Dobrý student zná 75% odpovědí, slabý 30%. Jestliže byla určitá otázka zodpovězena správně, určete pravděpodobnost, že student jen hádal, jde-li o:

- dobrého studenta,
- špatného studenta,
- náhodného studenta, kdy navíc víme, že dobrých studentů jsou 2/3.

Příklady k procvičení

Příklad

Turistický oddíl si předává zprávy Morseovou abecedou s těmito vlastnostmi: pokud je odvysílána *tečka*, pak ve 40% případů je přijata čárka (jinak tečka), pokud je odvysílána *čárka*, je v 1/3 případů přijata tečka (jinak čárka). Zpráva obsahuje tečky a čárky v poměru 5 : 3. Určete pravděpodobnost,

- že byla vyslaná tečka, pokud je přijatá čárka,
- že byla vyslaná tečka, pokud je přijatá tečka.

Příklad

Osoby X a Y přijdou na smluvené místo kdykoliv mezi 9.00 a 10.00[a]. Určete pravděpodobnost, že:

1. první z příchozích nebude muset na druhého čekat déle než 10 minut,
2. osoba Y přijde až jako druhá, jestliže přijde po 9.30.

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 $\frac{35}{532,5} \approx 6,6\%$. Pro 12 tis. žen ve věku 20–24 let:

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 Downovým syndromem, pokud vyšel pozitivní test, lze nyní spočítat jako $\frac{5,6}{605,2} \approx 0,9\%$.

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.

Martingale betting strategy

Letmý pohled na internet nám nabídne hned několik zaručených tipů, jak sázet v různých loteriích a neprohrát. Např. v ruletě sázíme na barvu dokud nevyhrajeme vždy dvojnásobek předchozí sázky (strategie známá jako *Martingale betting strategy)*. Viz návod již z roku 1882 (František Bačkovský, pseud. Vlastimil Benátský, Jak sázeti do loterie, bychom zcela jistě vyhráli).

Příklad

Alešovi zbylo 2500 Kč z pořádání tábora. Aleš není žádný ňouma: 50 Kč přidal z kasičky a rozhodl se jít hrát ruletu na automaty. Aleš sází pouze na barvu. Pravděpodobnost výhry při sázce na barvu je 18/37. Začíná sázet na 10 Kč a pokud prohraje, v další sázce vsadí dvojnásobek toho, co v předchozí (pokud na to ještě má, pokud ne, tak končí s hrou – byť by měl ještě peníze na nějakou menší sázku). Pokud nějakou sázku vyhraje, v následující sázce hraje opět o 10 Kč. Jaká je pravděpodobnost, že při tomto postupu vyhraje dalších 2550 Kč? (jakmile bude 2550 Kč v plusu, tak končí).

Řešení

Nejprve spočítejme, kolikrát po sobě může Aleš prohrát. Začíná-li s 10 Kč, tak na n vsazení potřebuje

$$10+20+\cdots+10\cdot 2^{n-1} = 10\cdot\left(\sum_{i=0}^{n-1} 2^i\right) = 10\cdot\left(\frac{2^n-1}{2-1}\right) = 10\cdot(2^n-1).$$

Jak snadno nahlédneme, číslo 2550 je tvaru $10(2^n-1)$ a to pro $n=8$. Aleš tedy může sázet osmkrát po sobě bez ohledu na výsledek sázky, na devět sázek by potřeboval již $10(2^9-1)=5110$ Kč a to v průběhu hry nikdy mít nebude (jakmile bude mít 5100 Kč, tak končí). Aby tedy jeho hra skončila neúspěchem, musel by prohrát osmkrát v řadě. Pravděpodobnost prohry při jedné sázce je $19/37$, pravděpodobnost prohry v osmi po sobě následujících (nezávislých) sázkách je tedy $(19/37)^8 \doteq 0{,}0048$, tedy docela mizivá.

Řešení

Pravděpodobnost, že v těchto osmi hrách vyhraje 10 Kč (při daném postupu), je tedy $1 - (19/37)^8$. Na to, aby vyhrál 2550 Kč, potřebuje 255 krát vyhrát po desetikoruně. Tedy opět podle pravidla součinu je pravděpodobnost výhry

$$\left(1 - \left(\frac{19}{37}\right)^8\right)^{255} \doteq 0{,}29.$$

Tedy pravděpodobnost výhry je nižší, než kdyby vsadil rovnou vše na jednu barvu.

Náhodné veličiny

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 40), 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 $(\Omega, \mathcal{A}, P)$.

Na prostoru $\mathbb{R}^k$ uvažujme nejmenší jevové pole $\mathcal{B}$ obsahující všechny k–rozměrné intervaly. Množinám v $\mathcal{B}$ říkáme **borelovské množiny** (nebo také měřitelné množiny) na $\mathbb{R}^k$.
Speciálně pro $k = 1$ jde o množiny, které obdržíme **z intervalů konečnými průniky a nejvýše spočetnými sjednoceními**.

Definice

Náhodná veličina X na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$ je taková funkce $X : \Omega \to \mathbb{R}$, že vzor $X^{-1}(B)$ patří do $\mathcal{A}$ pro každou Borelovskou množinu $B \in \mathcal{B}$ na $\mathbb{R}$ (tj. $X : \Omega \to \mathbb{R}$ je tzv. borelovsky měřitelná). Množinová funkce

$$P_X(B) = P(X^{-1}(B)) = P(\{\omega \in \Omega; X(\omega) \in B\})$$

se nazývá **rozdělení pravděpodobnosti** náhodné veličiny X.
Náhodný vektor $(X_1, \ldots, X_k)$ na $(\Omega, \mathcal{A}, P)$ je k–tice náhodných veličin.

Příklady k procvičení

Příklad

Hodíme jedenkrát kostkou, množina elementárních jevů je $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$. Jevovým polem nechť je $\mathcal{A} = \{\emptyset, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4, \omega_5, \omega_6\}, \Omega\}$.
Zjistěte jestli zobrazení $X : \Omega \to \mathbb{R}$ dané předpisem

a) $X(\omega_i) = i$ pro každé $i \in \{1, 2, 3, 4, 5, 6\}$,

b) $X(\omega_1) = X(\omega_2) = -2, X(\omega_3) = X(\omega_4) = X(\omega_5) = X(\omega_6) = 3$,

je náhodnou veličinou vzhledem k $\mathcal{A}$.

Příklad

Je dáno jevové pole $(\Omega, \mathcal{A})$, kde $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$ a

$$\mathcal{A} = \{\emptyset, \{\omega_1, \omega_2\}, \{\omega_3\}, \{\omega_4, \omega_5\}, \{\omega_1, \omega_2, \omega_3\}, \{\omega_1, \omega_2, \omega_4, \omega_5\}, \{\omega_3, \omega_4, \omega_5\}, \Omega\}.$$

Najděte nějaké (co nejobecnější) zobrazení $X : \Omega \to \mathbb{R}$, které bude náhodnou veličinou vzhledem k $\mathcal{A}$.

Definice náhodné veličiny zajišťuje, že pro všechny $-\infty \leq a \leq b \leq \infty$ existuje pravděpodobnost $P(a < X \leq b)$, kde používáme stručné značení pro jev $A = (\omega \in \Omega;\ a < X(\omega) \leq b))$.

Definice

Distribuční funkcí (*distribution, cumulative density function*) náhodné veličiny X je funkce $F : \mathbb{R} \to \mathbb{R}$ definovaná pro všechny $x \in \mathbb{R}$ vztahem

$$F(x) = P(X \leq x).$$

Distribuční funkcí náhodného vektoru $(X_1, \ldots, X_k)$ je funkce $F : \mathbb{R}^k \to \mathbb{R}$ definovaná pro všechny $(x_1, \ldots, x_k) \in \mathbb{R}^k$ vztahem

$$F(x) = P(X_1 \leq x_1 \wedge \cdots \wedge X_k \leq x_k).$$

Diskrétní náhodné veličiny

Předpokládejme, že náhodná veličina X na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$ nabývá jen konečně mnoha hodnot $x_1, x_2, \ldots, x_n \in \mathbb{R}$. Pak existuje tzv. **pravděpodobnostní funkce** $f(x)$ taková, že

$$f(x) = \begin{cases} P(X = x_i) & \text{pro } x = x_i \\ 0 & \text{jinak.} \end{cases}$$

Evidentně $\sum_{i=1}^{n} f(x_i) = 1$.
Takové náhodné veličině se říká **diskrétní**.
Každá náhodná veličina definovaná pro klasickou pravděpodobnost je diskrétní. Obdobně lze definici pravděpodobnostní funkce rozšířit na veličiny se spočetně mnoha hodnotami (pracujeme pak s nekonečnými řadami)

Příklady k procvičení

Příklad

Nechť $\Omega = \{\omega_1, \omega_2, \omega_3\}$ a $\mathcal{A} = \{\Omega, \emptyset, \{\omega_3\}, \{\omega_1, \omega_2\}\}$. Určete všechny pravděpodobnostní funkce zobrazující $\mathcal{A}$ do množiny $\{0, 1, \theta, 1 - \theta\}$.

Příklad

Třikrát nezávisle na sobě hodíme mincí. Náhodná veličina X udává počet hlav, které padnou při těchto hodech. Určete pravděpodobnostní a distribuční funkci náhodné veličiny X.

Příklad

Pravděpodobnost, že výrobek bude vyhovovat všem technickým požadavkům, je 0,9. Popište rozdělení náhodné veličiny udávající počet nevyhovujících výrobků mezi 3 výrobky.