Anglický matematik *Philip Beecroft* zvolil roku 1842 k odvození vztahu (11) pozoruhodný a podstatně jiný přístup. Uvedeme jej v příštím, závěrečném dílu tohoto seriálu.

Literatura

[1] *Leischner, P.:* Polibky kružnic: Jakob Steiner, 1.část. MFI, roč. 26 (2017), č. 1, s. 6–13.

[2] *Steiner, J.:* Einige geometrische Betrachtungen (1826), Verlag von Wilhelm Engelmann, Leipzig, 1901.

Od rodiny se třemi dětmi ke spolehlivosti testů

ŠÁRKA GERGELITSOVÁ – TOMÁŠ HOLAN

Matematicko-fyzikální fakulta UK, Praha

V tomto článku rozvineme úvahy o podmíněné pravděpodobnosti. Nejprve si zopakujme základní pojmy: budeme předpokládat, že máme nějaký *náhodný pokus*, u kterého jsme schopni předem vyjmenovat všechny jeho možné *výsledky*, které se navzájem vylučují a jeden z nich nastane vždy. *Jevem* nazveme podmnožinu množiny všech možných výsledků náhodného pokusu. Výsledky patřící jevu A nazýváme *výsledky příznivé jevu* A.

Pravděpodobnost jevu je součtem pravděpodobností výsledků, které jsou tomuto jevu příznivé. Pokud jsou všechny možné výsledky náhodného pokusu stejně pravděpodobné, je pravděpodobnost jevu rovna podílu

$$\frac{\textit{počet výsledků příznivých danému jevu}}{\textit{počet všech možných výsledků}}.$$

Úloha 1

Jaká je pravděpodobnost, že rodina se třemi dětmi má samé dívky?

Řešení. Pravděpodobnost narození chlapce považujeme za rovnu pravděpodobnosti narození dívky, $P(h) = P(d) = \frac{1}{2}$. Pohlaví jednotlivých dětí považujeme za navzájem nezávislé. Proto pravděpodobnost, že v rodině jsou tři dívky, je $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$.

Úloha 2

Jaká je pravděpodobnost, že rodina se třemi dětmi má samé dívky, jestliže víme, že jednu dívku mají? (tj. jestliže známe jednu dívku z rodiny, ale nevíme nic o jejích dvou sourozencích)

Řešení. Podle výše uvedené úvahy mohou být v rodině se stejnou pravděpodobností trojice dětí (od nejmladšího po nejstarší): (h, h, h); (h, h, d); (h, d, h); (d, h, h); (h, d, d); (d, h, d); (d, d, h); (d, d, d). Víme, že prvá možnost nenastává (nejsou tam tři bratři). Ze zbývajících sedmi stejně pravděpodobných možností vyhovuje jediná. Pravděpodobnost, že v rodině jsou tři sestry, je proto 1/7.

Úloha 3

Anička je nejstarší ze tří sourozenců. Jaká je pravděpodobnost, že má dvě sestry?

Řešení. Dívka Anička je v rodině nejstarším dítětem. Řešíme proto pouze pravděpodobnost toho, že následující dvě děti jsou dívky, tedy úlohu o dvou dětech, ne třech. Pravděpodobnost, že mladší z nich je chlapec, je rovna pravděpodobnosti, že to je dívka, je tedy rovna $\frac{1}{2}$. Pravděpodobnost, že starší sourozenec je dívka, je také $\frac{1}{2}$. Pohlaví mladšího a staršího sourozence je nezávislé, proto pravděpodobnost, že dívka Anička má dvě sestry, je $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$.

Jiné řešení. Řešme úlohu jako úlohu o třech dětech: Je-li první dítě dívka, mohou být v rodině pouze ty trojice, které mají na prvém místě dívku, tedy: (d, h, h); (d, h, d); (d, d, h); (d, d, d). Všechny mají stejnou pravděpodobnost, navzájem se vylučují a jejich sjednocení dává jistý jev. Příznivá je z nich jediná, proto je hledaná pravděpodobnost rovna $\frac{1}{4}$.

Proč je *úloha 3* jiná než *úloha 2*? Stručně řečeno: protože v *úloze 2* nevíme, kolikátým dítětem ona dívka je, a to způsobí změny pravděpodobnosti nahlížených jevů. Při řešení úloh o pravděpodobnosti rozebereme zadání úlohy a snažíme se najít odpovídající „model“, který umíme vyjádřit pomocí známých pravděpodobností. A mnohdy dojdeme ke správnému řešení různými úvahami. *Úloha 2* bývá někdy označována pro svou záměnu s *úlohou 3* jako paradox. Ukažme si několik „bludných“ úvah, jejichž vinou to opravdu jako paradox vypadá.

Různé chybné úvahy o druhé úloze

Bludná úvaha 1. Vezmeme v úvahu všechny možnosti (a pojmenujeme zmíněnou dívku Anička): Anička může být nejmladší, prostřední nebo nejstarší. *Tyto možnosti se navzájem vylučují* a dohromady dávají jev jistý. Pravděpodobnost, že Anička je nejmladší, prostřední či nejstarší, je stejná, tedy $\frac{1}{3}$. V každém ze zmíněných případů je pravděpodobnost, že Anička má dvě sestry, rovna $\frac{1}{4}$ (viz úloha 3). Celková pravděpodobnost je tedy $\frac{1}{3} \cdot \frac{1}{4} + \frac{1}{3} \cdot \frac{1}{4} + \frac{1}{3} \cdot \frac{1}{4} = \frac{1}{4}$.

Bludná úvaha 2. Dívka, o které už víme, má tedy dva sourozence. *Pravděpodobnost, že mladší z nich je chlapec, je rovna pravděpodobnosti, že to je dívka, je tedy $\frac{1}{2}$. Pravděpodobnost, že starší sourozenec je dívka, je také $\frac{1}{2}$. Pohlaví mladšího a staršího sourozence je nezávislé, proto pravděpodobnost, že dívka má dvě sestry je $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$.*

Kde je chyba?

Ukažme si, kde se v jednotlivých úvahách stala chyba. (Pozorný čtenář si zajisté všiml, že v každé z úvah je část textu vyznačena kurzivou.)

Co nás mate v úvaze 1? Jméno. Chceme-li počítat pravděpodobnost jevu jako

$$\frac{\textit{počet výsledků příznivých danému jevu}}{\textit{počet všech možných výsledků}},$$

musí mít ony počítané výsledky stejnou pravděpodobnost.

Často ale posuzujeme jevy, jejichž různé výsledky stejnou pravděpodobnost nemají. Příkladem výsledků pokusu, které stejnou pravděpodobnost nemají, jsou například součty ok na dvou vržených kostkách: výsledek „padne součet 2“ má jistě menší pravděpodobnost než výsledek „padne součet 7“. Takové výsledky se pro výpočet pravděpodobnosti podle zmíněného schématu použít nedají.

Dalším typickým příkladem pokusů, kde výsledky stejnou pravděpodobnost nemají, je vznik skupin prvků, v nichž se prvky opakují („kombinace s opakováním“). Proto při řešení úloh o pravděpodobnosti prvky rozlišujeme (hrnečky číslujeme, dětem dáváme jména…). A právě to nás teď zmátlo. Kdyby Anička v úvaze 1 byla opět jen „dívka“, asi bychom nenapsali: „Vezmeme v úvahu všechny možnosti: dívka může být nejmladší, prostřední nebo nejstarší. Tyto možnosti se navzájem vylučují…“ Je-li v rodině více dívek než jedna, uvedené možnosti se nevylučují.

- Tomu, že dívka je v rodině nejmladší, odpovídají trojice: (d, h, h); (d, h, d); (d, d, h); (d, d, d).
- Tomu, že dívka je v rodině prostřední, odpovídají trojice: (h, d, h); (h, d, d); (d, d, h); (d, d, d).
- Tomu, že dívka je v rodině nejstarší, odpovídají trojice: (h, h, d); (h, d, d); (d, h, d); (d, d, d).

Trojice se dvěma dívkami jsme započítali dvakrát, trojici samých dívek dokonce třikrát. O stejné pravděpodobnosti nemůže být řeč.

Budiž, úvaha 1 není v pořádku. Ale co lze vytknout úvaze 2? Kurziva napovídá, že je to špatně celé…?

Již z minulých úvah vidíme, že pravděpodobnost, že jsou v rodině právě dva chlapci (trojice (h, h, d); (h, d, h); (d, h, h)), a pravděpodobnost, že další dvě děti v rodině jsou různého pohlaví (trojice (h, d, d); (d, d, h); (d, h, d)), je stejná a rovná trojnásobku pravděpodobnosti, že v rodině jsou jen dívky (trojice (d, d, d)), což je, jak víme $\frac{1}{7}$.

Řešit proto pravděpodobnost toho, že v rodině je další dívka, když tam již jedna dívka je, je matoucí. Úvahy nás přivádí zpět k předchozím argumentům. Nevíme-li, kolikátým dítětem ona dívka je, nemůžeme předpokládat, že to, že sourozenec je chlapec či dívka, má stejnou pravděpodobnost, a už vůbec nejsou tyto jevy nezávislé. Nejde o narození nového jedince (tam by pravděpodobnost opravdu $\frac{1}{2}$ byla), ale o zkoumání rodin, kde již 3 děti jsou. (Kdyby šlo o ještě nenarozené sourozence, byla by ona dívka nejstarší – a ten případ jsme již vyřešili.)

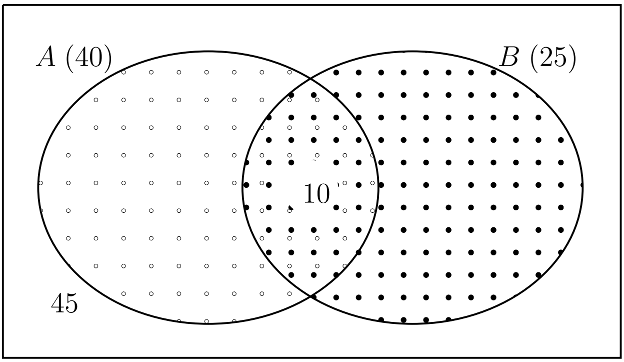
Situaci si můžeme představit jako losování. S pravděpodobností $\frac{3}{7}$ v něm vybereme rodinu, která má další děti dva hochy, z ní tedy sourozenec dívka nebude. S pravděpodobností $\frac{3}{7}$ vybereme rodinu, která má další děti hocha a dívku, z ní tedy (mladší) sourozenec bude dívka s pravděpodobností $\frac{1}{2}$. S pravděpodobností $\frac{1}{7}$ vybereme rodinu, která má další děti dvě dívky, z ní tedy sourozenec dívka je jev jistý, pravděpodobnost je 1. Jenže tento výběr zároveň odpověděl na otázku, jaké pohlaví má starší ze sourozenců, a tedy oba sourozenci. Dívka to byla právě v $\frac{1}{7}$ rodin.

Podmíněná pravděpodobnost

Veškeré pochyby si ušetříme, podíváme-li se na problém tak, jak je formulován: jako na podmíněnou pravděpodobnost. Připomeňme nejprve, co rozumíme podmíněnou pravděpodobností, a uveďme několik pravidel pro její určení.

Máme dva různé jevy A, B. Předpokládejme, že známe pravděpodobnosti toho, že nastane jev A, že nastane jev B a že nastanou oba jevy zároveň. Budeme dále zkoumat, jak je pravděpodobnost jednoho jevu v náhodném pokusu ovlivněna tím, že již známe výsledek jevu druhého. Označme pravděpodobnosti toho, že nastane jev A, jev B a oba jevy $A \cap B$, po řadě $P(A)$, $P(B)$, $P(A \cap B)$. Zápisem $P(A|B)$ budeme rozumět pravděpodobnost, že nastal jev A za předpokladu, že nastal jev B.

Víme-li, že nastal jev B, pak to, že nastal jev A znamená, že nastaly oba jevy současně. Protože pravděpodobnost, že nastal jev B, má hodnotu $P(B)$, platí $P(A|B) \cdot P(B) = P(A \cap B)$.



Obr. 1

Ilustrujme si uvedenou úvahu na příkladu relativních četností provedených pokusů z pohledu jevů A, B na obr. 1. Zde platí:

$$P(A) = \frac{40}{100}, \quad P(B) = \frac{25}{100}, \quad P(A \cap B) = \frac{10}{100}.$$

Pokud nastal jev B, je množinou všech možných výsledků celá množina B, příznivými výsledky jevu A zůstávají jen ty, které leží zároveň v množině B, a tedy $P(A|B) = \frac{10}{25}$. Potom

$$P(A|B) \cdot P(B) = \frac{10}{25} \cdot \frac{25}{100} = \frac{10}{100} = P(A \cap B).$$

Analogicky platí $P(B|A) \cdot P(A) = P(B \cap A)$. Z rovnosti pravděpodobností $P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$ odvodíme pro $P(A) > 0$ a $P(B) > 0$ vztahy pro obě podmíněné pravděpodobnosti:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}, \quad P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}.$$

Tyto vztahy jsou známy jako *Bayesův vzorec.*

Vraťme se k naší úloze. Otázka „Jaká je pravděpodobnost, že rodina se třemi dětmi má samé dívky, jestliže víme, že jednu dívku mají?“ je zřejmou formulací podmíněné pravděpodobnosti.

Označme D tvrzení „v rodině je dívka“, $D3$ tvrzení „v rodině jsou tři dívky“. Dále $D3|D$ je zkoumané tvrzení a $D|D3$ je triviálně pravdivé tvrzení „v rodině je dívka, jsou-li tam tři dívky“.

Potom naše otázka je otázkou na pravděpodobnost $P(D3|D)$ a podle vzorce pro podmíněnou pravděpodobnost platí:

$$P(D3|D) = \frac{P(D|D3) \cdot P(D3)}{P(D)}.$$

V naší úloze platí $P(D) = \frac{7}{8}$, $P(D3) = \frac{1}{8}$, $P(D|D3) = 1$. Proto

$$P(D3|D) = \frac{1 \cdot \frac{1}{8}}{\frac{7}{8}} = \frac{1}{7}.$$

Podmíněná pravděpodobnost tak, jak je užita ve výše uvedeném příkladu, je velmi jednoduchá. Jeden jev je zde podjevem (tj. podmnožinou) druhého. V obecnějších případech je podmíněná pravděpodobnost velmi důležitá, například při zkoumání pravděpodobností, jak nějaký test určí pravděpodobnost výskytu poruchy, odchylky, onemocnění…

Bayesův vzorec pro dílčí jevy

Představme si dva různé jevy A, B, přičemž jev B je sjednocením navzájem neslučitelných jevů $B1$, $B2$, $B3$. Potom

$$P(B) = P(B1) + P(B2) + P(B3).$$

Proto

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B1) + P(B2) + P(B3)}.$$

Ukažme si jeho použití na příkladu.

Z dětí starších 10 let, které žijí na vesnici, umí sekat trávu kosou 50 %. Z těch, které na vesnici tráví jen prázdniny, to umí 20 %. Z dětí, které opustí město jen na výlety, tábory a podobně, to umí jen 2 %. Jarda umí sekat trávu kosou. Jardu jsme náhodně vylosovali z 1000 dětí, z nichž je 300 z vesnice, 550 z města, ale jezdí na prázdniny na vesnici, a 150 tráví většinu času ve městě. Jaká je pravděpodobnost, že Jarda je z města?

U je jev „umí sekat kosou“, přičemž $U1$ znamená „umí sekat kosou a je z města“, $U2$ „umí sekat kosou a jezdí na prázdniny“ a $U3$ znamená „umí sekat kosou a je z vesnice“. M je jev „je z města“, L je jev „jezdí na prázdniny“, V je jev „je z vesnice“.

$$P(M|U) = \frac{P(U|M) \cdot P(M)}{P(U1) + P(U2) + P(U3)} =$$

$$= \frac{P(U|M) \cdot P(M)}{P(U|M) \cdot P(M) + P(U|L) \cdot P(L) + P(U|V) \cdot P(V)} =$$

$$= \frac{0{,}02 \cdot 0{,}15}{0{,}02 \cdot 0{,}15 + 0{,}2 \cdot 0{,}55 + 0{,}5 \cdot 0{,}3} = \frac{0{,}003}{0{,}003 + 0{,}11 + 0{,}15} = \frac{3}{263} \doteq 0{,}014.$$

Chceme-li si vzorec lépe představit, pak při uvedených počtech dětí a uvedených pravděpodobnostech, že umí sekat kosou, máme z 1 000 dětí 150 „sekáčů“ z vesnice, 110 z prázdninových a 3 z města. Jarda je z města, pokud je jedním z oněch tří, přičemž všech dětí, které umí sekat kosou, je 263. Proto je hledaná pravděpodobnost $\frac{3}{263} \doteq 0,014$.

Na závěr si ukažme jeden často uváděný příklad (podobný najdete například v [2]).

Screeningové programy

Odborníci na lidské zdraví nepochybují o tom, že včasné podchycení rizika vážných onemocnění významně zvyšuje šanci na dlouhý život bez vážných komplikací. Proto existují tzv. screeningové programy, při nichž jsou vyšetřeni všichni jedinci z určité (skupiny) populace. Jenže žádný test není stoprocentně spolehlivý. Dobrý test bude mít jednak vysokou schopnost podchytit danou vadu (tj. vysokou *senzitivitu* – procento zachycených výskytů), jednak nebude vyvolávat falešné poplachy (bude mít vysokou *specificitu* – procento nálezů, které jsou „oprávněné“).

To, že daným testem projdete jako pozitivní, tedy signalizuje danou vadu pouze s nějakou pravděpodobností, která podle výše uvedených úvah závisí také na tom, jak často se zmíněná diagnóza v populaci vyskytuje. Uveďme (smyšlený) příklad.

Ve sledované skupině obyvatel se vada V vyskytuje u každého tisícího člověka. Dostupný test má senzitivitu 98 % (tj. pouhá 2 % případů neodhalí) a specificitu 95 % (tj. 5 % zdravých lidí má pozitivní test).

Otázka: Jaká je pravděpodobnost, že člověk A trpí danou vadou, pokud byl jeho test pozitivní?

Odpověď: Označme $P(V)$ pravděpodobnost, že člověk A trpí vadou V,
$P(T)$ pravděpodobnost, že byl pozitivně testován,
$P(V|T)$ pravděpodobnost, že trpí vadou, byl-li pozitivně testován,
$P(T|V)$ pravděpodobnost, že byl pozitivně testován, trpí-li vadou,
$P(T|V')$ pravděpodobnost, že byl pozitivně testován, netrpí-li vadou.

Tedy v našem příkladu $P(V) = 0{,}001$, $P(T|V) = 0{,}98$, $P(T|V') = 0{,}05$. Pak podle Bayesova vzorce platí

$$P(V|T) = \frac{P(T|V) \cdot P(V)}{P(T)} = \frac{P(T|V) \cdot P(V)}{P(T|V) \cdot P(V) + P(T|V') \cdot P(V')} =$$

$$= \frac{0{,}98 \cdot 0{,}001}{0{,}98 \cdot 0{,}001 + 0{,}05 \cdot 0{,}999} \doteq 0{,}019.$$

Pravděpodobnost, že pozitivně testovaný člověk má opravdu problém, tedy není oněch obávaných 95 %, ale necelá dvě procenta.

Poznámka: Ověřte, že pokud je výskyt vady v populaci desetkrát vyšší, tj. $P(V) = 0{,}01$, pak při nezměněné senzitivitě a specificitě bude zkoumaná pravděpodobnost $P(V|T) \doteq 0{,}165$.

V článku jsme uvedli několik příkladů určení pravděpodobnosti podmíněných jevů. Do této skupiny spadají také mnohé úlohy uváděné v různých kvízech či sadách úloh pro volné chvíle. Klíčem ke správnému řešení zde bývá nejen znalost metod výpočtu pravděpodobnosti, ale hlavně schopnost správně zformulovat problém. Jednu podobnou úlohu najdete i v učebnici [1] (úloha *2.96).

Literatura

[1] *Calda, E.–Dupač, V.:* Matematika – Kombinatorika, pravděpodobnost, statistika, vyd. 4., Prometheus, Praha, 2002.

[2] *Gonick, L.*: The Cartoon Guide to Statistics, HarperCollins, 1993.

[3] *Poskitt, K.*: Děsivá věda: Vrtkavé štěstí, překlad Pipek, J., EGMONT ČR, Praha, 2005.