Podmíněná pravděpodobnost

Definice: Pokud B je jev splňující $P(B) > 0$, pak *podmíněnou pravděpodobnost* jevu A za podmínky jevu B definujeme jako

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

Zadání: V osudí je b bílých a c černých míčků. Předpokládejme, že jsme v prvním tahu vytáhli bílý míček. Do osudí ho nevracíme. Jaká je za této podmínky pravděpodobnost, že druhý vytažený míček bude rovněž bílý?

Řešení: Vytažení bílého míčku v prvním tahu se podle klasické pravděpodobnosti stane s pravděpodobností $P(B_1) = b/(b + c)$. Při výběru bez vracení máme pro druhý tah jen $b+c-1$ míčků, z nichž $b-1$ je bílých, proto $P(B_2 \mid B_1) = (b-1)/(b+c-1)$. To odpovídá definici podmíněné pravděpodobnosti, protože $P(B_1 \cap B_2) = b(b-1)/[(b+c)(b+c-1)]$.

Pro nezávislé jevy A a B je $P(A \cap B) = P(A)P(B)$, a tudíž $P(A \mid B) = P(A)$. Podobně dostaneme, že $P(B \mid A) = P(B)$, pokud $P(A) > 0$. Obdobně také platí $P(A \mid B^c) = P(A)$ a $P(B \mid A^c) = P(B)$ pro $P(A), P(B) < 1$, protože $P(A \cap B^c) = P(A) - P(A \cap B) = P(A)P(B^c)$.

Zadání: Jak se řešení předchozí úlohy změní, když uvažujeme výběr s vracením?

Řešení: Pro druhý tah je nyní v osudí $b+c$ míčků, z nichž je b bílých. Proto $P(B_2 \mid B_1) = P(B_2) = b/(b+c)$. Jevy B_1 a B_2 jsou nezávislé. Obdobně $P(B_2 \mid B_1^c) = P(B_2) = b/(b+c)$, neboli jevy B_1^c a B_2 jsou nezávislé, což je odlišné od případu bez vracení, kde máme $P(B_2 \mid B_1^c) = b/(b+c-1)$.

Poznámka: Vlastnosti podmíněné pravděpodobnosti:

1. $0 \leq P(A \mid B) \leq 1$
2. $P(A_1 \cup A_2 \mid B) = P(A_1 \mid B) + P(A_2 \mid B)$, pokud $A_1 \cap A_2 = \emptyset$
3. $P(A^c \mid B) = 1 - P(A \mid B)$
   Důkaz: $P(A^c \mid B) = \frac{P(A^c \cap B)}{P(B)} = \frac{P(B) - P(A \cap B)}{P(B)} = 1 - \frac{P(A \cap B)}{P(B)}$
4. $P(A \mid B)P(B) = P(B \mid A)P(A) = P(A \cap B)$
5. $P(A \mid B) > P(A) \iff P(B \mid A) > P(B)$

Zadání: Tenista má prvé podání úspěšné s pravděpodobností 0,6 a druhé s pravděpodobností 0,8. S jakou pravděpodobností se dopustí dvojchyby?

Řešení: Úspěšné podání označíme U a neúspěšné N. Přidáme index podle toho, jestli se jedná o první nebo druhé podání. Ze zadání plyne $P(U_1) = 0{,}6$ a $P(U_2 \mid N_1) = 0{,}8$. Odtud díky třetí vlastnosti z předchozí poznámky $P(N_1) = 0{,}4$ a $P(N_2 \mid N_1) = 0{,}2$. Celkově dostaneme pravděpodobnost dvojchyby $P(N_1 \cap N_2) = P(N_2 \mid N_1)P(N_1) = 0{,}2 \cdot 0{,}4 = 0{,}08$.

Poznámka: Nesprávné úvahy s podmíněnými pravděpodobnostmi jsou často zdrojem chybných závěrů. Ukážeme to na některých situacích.

Zadání: U viníků dopravních nehod bylo v 10 % případů zjištěno požití alkoholických nápojů. Znamená to, že střízliví řidiči jsou více nebezpeční?

Řešení: Když N bude značit jev, že došlo k nehodě, a O jev, že řidič pil alkohol, pak zadání říká $P(O \mid N) = 0{,}1$, neboli $P(O^c \mid N) = 0{,}9$. To ale nic nevypovídá o $P(N \mid O)$ nebo $P(N \mid O^c)$. K tomu bychom potřebovali mít nějakou dodatečnou znalost. Budeme-li předpokládat, že $P(O) = 0{,}005$, potom dostáváme $P(N \mid O) = P(O \mid N)P(N)/P(O) = 20P(N)$ a $P(N \mid O^c) = P(O^c \mid N)P(N)/P(O^c) = \frac{0{,}9}{0{,}995}P(N) \doteq 0{,}905P(N)$, tedy požití alkoholu je daleko nebezpečnější.

Poznámka: K podobným dezinterpretacím statistických informací dochází často, zde je pár konkrétních příkladů ze sdělovacích prostředků:

- nebezpečnější je řídit večer než ráno (čtyřikrát víc nehod na dálnici bylo v 7 večer než v 7 ráno),
- pozor na české turisty v Tatrách (nejvíc zahraničních turistů zapletených do nehody pochází z Česka),
- chlapci jsou ve větším ohrožení při jízdě na kole (většina dětí při nehodách na kole jsou chlapci),
- fotbal je nejnebezpečnější sport (průzkum nehod při sportu),
- domov je nebezpečné místo (ke třetině nehod dochází ve vlastním bydlení),
- manželství je nebezpečný stav (nejvíc vražd žen spáchali jejich manželé).

Ve všech případech je z toho, že podmíněná pravděpodobnost $P(A \mid B)$ je velká, nesprávně usuzováno, že $P(B \mid A) > P(B)$. To ale nelze učinit bez informace o $P(A)$. Řekněme, že jev A je příznivě nakloněný jevu B, když $P(B \mid A) > P(B)$. Díky symetrické vlastnosti (viz 5. vlastnost z poznámky výše) je to ekvivalentní tomu, že jev B je příznivě nakloněný jevu A. Problém je v tom, že nemůžeme prohlásit, že A a B jsou příznivě nakloněny, jenom na základě toho, že $P(A \mid B)$ je velká. V případě s opilými řidiči jsme viděli, že střízlivost není příznivě nakloněná nehodě, i když 90 % nehod zaviní střízlivý člověk.

Poznámka: Když B je příznivě nakloněno A, pak B není příznivě nakloněno A^c. Nicméně může být příznivě nakloněno podmnožinám A^c.

Příklad: (Kaighův paradox) Předpokládejme, že firma najala 160 z 1 000 uchazečů (z nich bylo 200 černých, 200 Hispánců a 600 bílých). Z přijatých uchazečů bylo 120 bílých, což znamená, že bylo přijato 20 % bílých uchazečů a pouze 10 % nebílých. Být bělochem je tak příznivě nakloněno tomu být přijat: $P(\text{přijat} \mid \text{bílý}) = 0{,}2 > P(\text{přijat}) = 0{,}16$. Neimplikuje to ale, že příslušnost k nějaké nebílé populaci je nutně nepříznivě nakloněna k tomu být přijat. Předpokládejme, že 36 ze 40 přijatých nebílých jsou Hispánci. Potom $P(\text{přijat} \mid \text{Hispánec}) = 0{,}18 > P(\text{přijat}) = 0{,}16$. Neboli být Hispánec je také příznivě nakloněno přijetí.

Obecně platí, že uvažujeme-li místo rozkladu na dvě kategorie (A a A^c) disjunktní rozklad na víc kategorií, např. tři (A_1, A_2 a A_3), které jsou seřazené tak, že $P(B \mid A_1) \geq P(B \mid A_2) \geq P(B \mid A_3)$, pak B je příznivě nakloněno A_1, nepříznivě nakloněno A_3 a nemůžeme rozhodnout u A_2.

Poznámka: Další chybný závěr na základě podmíněných pravděpodobností je založen na faktu, že z nerovnosti $P(A \mid B) > P(A \mid C)$ neplyne $P(A \mid B \cap D) > P(A \mid C \cap D)$.

Příklad: V době španělsko–americké války (1898) byla v námořnictvu USA úmrtnost zhruba 9 promile, zatímco u civilistů v New Yorku byla 16 promile. Pracovníci armády použili tyto statistiky při náboru nováčků. Ve skutečnosti nelze tato čísla srovnávat. V armádě jsou zdraví a silní jedinci, zatímco mezi civilisty jsou zahrnuti také všichni staří a nemocní. Označíme-li jako A úmrtí daného jedince, jev B, že jde o obyvatele New Yorku, a jev C, že jde o člena námořnictva. Pak data udávají $P(A \mid B) > P(A \mid C)$. Ovšem, kdybychom uvažovali stejně staré jedince, např. ve věku 20–30 let (jev D), pak můžeme čekat, že $P(A \mid B \cap D) < P(A \mid C \cap D)$.

Poznámka: Obráceně z nerovnosti $P(A \mid B \cap D) > P(A \mid C \cap D)$ neplyne $P(A \mid B) > P(A \mid C)$.

Příklad: V jednom novinovém článku se objevilo tvrzení, že lidé se stávají šťastnější tím, jak stárnou. Zdůvodnění bylo následující: z lidí, kteří zemřou v rozmezí let 20 až 25, kolem 25 % spáchá sebevraždu. S věkem tato hladina klesá, takže například z lidí, kteří zemřou ve věku nad 70, spáchá sebevraždu pouze 2 procenta. Formálně to znamená $P(\text{sebevražda} \mid \text{věk 20–25 a smrt}) > P(\text{sebevražda} \mid \text{věk} \geq 70 \text{ a smrt})$, což je pravda, ale neimplikuje to, že $P(\text{sebevražda} \mid \text{věk 20–25}) > P(\text{sebevražda} \mid \text{věk} \geq 70)$.

Poznámka: Může se stát, že platí $P(A \mid B \cap C) > P(A \mid B^c \cap C)$ a $P(A \mid B \cap C^c) > P(A \mid B^c \cap C^c)$ a zároveň $P(A \mid B) < P(A \mid B^c)$.

Příklad: (Simpsonův paradox) Sledujme úspěšnost studentů u státních maturit z matematiky. Celkový podíl propadlých u maturit klesl z 23,9 % v roce 2015 na 23,2 % v roce 2016, což Cermat dokumentoval jako mírné zlepšení maturantů v matematice. Přitom podíly propadlých u studentů gymnázií a také u studentů ostatních škol se zvýšily.

	2015	2016
gymnázia	$\frac{249}{7\,095} \doteq 0{,}035$	$\frac{254}{6\,916} \doteq 0{,}037$
ostatní	$\frac{4\,366}{12\,209} \doteq 0{,}358$	$\frac{3\,686}{10\,072} \doteq 0{,}366$
celkem	$\frac{4\,615}{19\,304} \doteq 0{,}239$	$\frac{3\,940}{16\,988} \doteq 0{,}232$

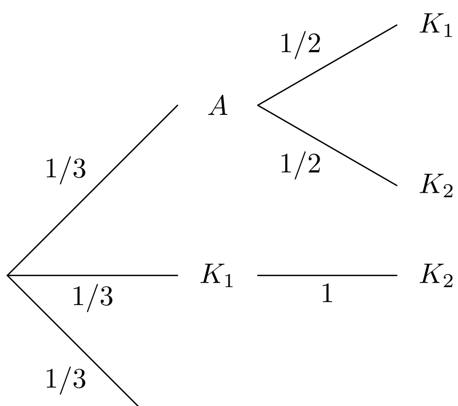
Poznámka: Chyby v úvahách s podmíněnými pravděpodobnostmi dělají i profesionální matematici. O tom svědčí slavný pravděpodobnostní problém, který přitáhl a stále přitahuje pozornost mnoha lidí. Je založen na jedné americké televizní show, kterou moderoval Monty Hall.

Zadání: (Monty Hallův problém) V televizní show jsou hráči ukázány troje dveře, jedny skrývají auto, zbylé dvoje kozu. Hráč zvolí jedny dveře. Moderátor, který ví, kde je auto umístěno, otevře některé z neoznačených dveří a odkryje kozu. Poté nabídne hráči, jestli chce svou volbu změnit. Je výhodné provést změnu?

Řešení: Existují různé varianty tohoto problému, proto zdůrazněme, že předpokládáme, že pokud má moderátor více možností, které dveře otevřít, rozhodne se náhodně. Pokud by hráč změnil svou volbu, tak vyhraje auto právě tehdy, když na začátku vybral dveře s kozou, což má pravděpodobnost 2/3, proto je výhodné svou volbu změnit. Přesto se najde hodně lidí, kteří argumentují tím, že po odkrytí jedněch dveří zbývají dvě možnosti a pravděpodobnost výhry je tudíž 1/2.

Bez újmy na obecnosti uvažujme, že soutěžící zvolil první dveře. Rozlišíme-li obě kozy, tak existuje 6 stejně pravděpodobných rozmístění objektů za dveře (AK_1K_2, AK_2K_1, K_1AK_2, K_2AK_1, K_1K_2A a K_2K_1A). Pokud má moderátor dvě možné volby, rozhoduje se náhodně. Můžeme uvažovat, že v tom případě zvolí K_1. Potom v případech AK_1K_2, K_1K_2A a K_2K_1A moderátor otevře druhé dveře, zatímco v případech AK_2K_1, K_1AK_2 a K_2AK_1 otevře třetí dveře. Při strategii změny rozhodnutí je podmíněná pravděpodobnost výhry auta za podmínky, že moderátor otevře druhé dveře, rovna 2/3 (nepříznivý je jen případ AK_1K_2). Podobně podmíněná pravděpodobnost výhry auta za podmínky, že moderátor otevře třetí dveře, je také 2/3 (nepříznivý je jen případ AK_2K_1).

Srozumitelné znázornění je pomocí stromu s možnými průběhy. První větvení (vlevo) znázorňuje volbu soutěžícího, druhé větvení (vpravo) ukazuje, které dveře moderátor odkryl.



Zkusme vše zapsat ještě podrobněji pomocí podmíněných pravděpodobností. Označme A_i jev, že auto se skrývá za dveřmi s pořadovým číslem i; B_i jev, že hráč zvolí dveře s pořadovým číslem i; a C_i jev, že moderátor otevře dveře s číslem i. Potom

$$
\begin{aligned}
P(A_1 \mid B_1 \cap C_3) &= \frac{P(A_1 \cap B_1 \cap C_3)}{P(B_1 \cap C_3)} \\
&= \frac{P(A_1 \cap B_1 \cap C_3)}{P(A_1 \cap B_1 \cap C_3) + P(A_2 \cap B_1 \cap C_3) + P(A_3 \cap B_1 \cap C_3)} \\
&= \frac{P(C_3 \mid A_1 \cap B_1)P(A_1 \cap B_1)}{P(C_3 \mid A_1 \cap B_1)P(A_1 \cap B_1) + P(C_3 \mid A_2 \cap B_1)P(A_2 \cap B_1) + 0} \\
&= \frac{\frac{1}{2} \cdot \frac{1}{3} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} \cdot \frac{1}{3}} = \frac{1}{3}
\end{aligned}
$$

a

$$
\begin{aligned}
P(A_2 \mid B_1 \cap C_3) &= \frac{P(C_3 \mid A_2 \cap B_1)P(A_2 \cap B_1)}{P(C_3 \mid A_1 \cap B_1)P(A_1 \cap B_1) + P(C_3 \mid A_2 \cap B_1)P(A_2 \cap B_1) + 0} \\
&= \frac{1 \cdot \frac{1}{3} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} \cdot \frac{1}{3}} = \frac{2}{3}.
\end{aligned}
$$

Na podobném logickém principu funguje úloha o třech vězních nebo Bertrandův přihrádkový paradox (viz prezentaci s úlohami).