Jaké pravděpodobnostní rozdělení kdy použít?

18. března 2016

Toť netriviální otázka. Asi bych mluvil o třech případech. Někdy víme o problému tolik, že pravděpodobnostní rozdělení určíme naprosto jednoznačně. Jindy máme pouze indicie, které napovídají, co to asi bude, někdy máme syrová data a nevíme o nich vůbec nic.

Pravděpodobnostní rozdělení je určeno jednoznačně

Jsou možné jen dva výsledky = Alternativní

Příkladem může být otázka, jestli v příští hodině nějaké auto projede nebo neprojede. Často ale může být cenné zvážit, jestli není lepší položit otázku jinak - např.: "Kolik aut v příští hodině projede?" Pak se ovšem od alternativního rozdělení přesouváme k nějakému jinému.

Sčítáme výsledky opakujících se alternativních pokusů se stejnou pravděpodobností úspěchu = Binomické

Zde je dobré si zformulovat, co je tím jedním alternativním pokusem a jestli je pravděpodobnost opravdu stále stejná.

Příklad 1: V pytlíku je deset kuliček, z toho pět modrých. Vytáhnu bez vracení čtyři kuličky. Jaká je pravděpodobnost, že vytáhnu 0, 1, 2, 3, 4 modré?

V prvním tahu vytáhnu modrou s PSTí 50%. V druhém tahu ... POZOR! Pravděpodobnost se změnila! Tedy nejde o binomické rozdělení!

Příklad 2: V rodině je pět dětí. S jakou pravděpodobností mají 0, 1, 2, 3, 4, 5 děvčat?

Alternativním pokusem je narození dítěte. Úspěchem (v pravděpodobnostním smyslu) je narození děvčete. PST narození děvčete je asi pro každý porod stejná, tedy půjde o binomické rozdělení.

Zjišťujeme, kolik se v daném čase, úseku, objemu vyskytne jednotek, přičemž jejich výskyt je navzájem nezávislý = Poissonovo

Zde musíme uvážit, jestli můžeme opravdu výskyt jednotek považovat za (víceméně) nezávislý.

Příklad 1: Kolik aut projede během minuty měřícím bodem za přechodem se světly?

Tady nemůže být o nezávislosti ani řeč. Pokud zjistím nějaké auto, pak je přechod průjezdný a pojedou i auta další. Pokud naopak auta nejedou, pak je přechod neprůjezdný a nepojedou nejspíš ani v dalších vteřinách. Nepůjde tedy o Poissonovo rozdělení.

Příklad 2: Kolik aut projede během minuty měřícím bodem na silnici druhé třídy mezi vesnicemi?

Pokud není v blízkosti semafor, železniční přejezd nebo něco podobného a doprava nevytváří kongesce, pak nejspíš půjde o Poissonovo rozdělení.

Pravděpodobnost všech výsledků je stejná = Rovnoměrné (diskrétní nebo spojité)

Příkladem může být házení kostkou.

Zajímá nás doba, za jak dlouho zanikne nějaká jednotka a pravděpodobnost, že jednotka zanikne (pokud ještě nezanikla) v příští sekundě (minutě, hodině) je stále stejná = Exponenciální

Krásným příkladem, který má popisovanou vlastnost, je rozpad atomů nebo částic. U rozbití přístrojů nesmí být znatelný vliv stárnutí nebo opotřebení materiálu. Příkladem může být praskání wolframové žárovky. Protipříkladem může být prodření kalhot.

Náhodná veličina vzniká jako součet nebo průměr mnoha náhodných veličin se stejným rozdělením = Přibližně Gaussovo

Toto je tvrzením Centrální limitní věty. Shoda je tím lepší, čím více náhodných veličin sčítám nebo průměruji. Obvykle se za dostatečný počet považuje cca 30.

Ke Gaussovu rozdělení se blíží jak spojitá, tak i diskrétní rozdělení. O těch diskrétních říkáme, že se Gaussovu blíží "v distribuci".

Někdy pravděpodobnostní rozdělení zjistíme přímo ze zadání

Příklad: V platu je 30 vajec. Deset z nich jsou pukavci. Vytáhnu pět z nich. Jaká je pravděpodobnost, že vytáhnu 0, 1, 2, 3, 4, 5 pukavců?

Výsledkem náhodného pokusu je vytažená pětice. Každá pětice má stejnou šanci, že ji vytáhnu. Tedy mohu použít klasickou definici pravděpodobnosti a

PST spočíst jako poměr počtu vyhovujících elementárních jevů ku počtu všech elementárních jevů.

Všech možných pětic z 30 je $\binom{30}{5}$. Pokud vytáhnu k pukavců, tak možných k-tic pukavců je $\binom{10}{k}$ a (5-k)-tic dobrých je $\binom{20}{5-k}$. Tedy dostávám vzorec pro pravděpodobnostní funkci:

$$f(k) = \frac{\binom{10}{k} \cdot \binom{20}{5-k}}{\binom{30}{5}}.$$

Indicie

V tomto případě některé rozdělení tipneme. Pak by měl následovat test, zda data tomuto rozdělení opravdu odpovídají.

Máme diskrétní rozdělení, přičemž jsou možné libovolně velké hodnoty ... možná Poissonovo?

Máme diskrétní rozdělení, přičemž existuje maximální možná hodnota ... možná Binomické nebo Rovnoměrné?

Náhodná veličina je výslednicí mnoha malých vlivů ... asi přibližně Gaussovo?

Zatímco Centrální limitní věta hovoří o sčítání nebo průměrování náhodných veličin, gaussovskost se obvykle zachová, i když je působení jednotlivých vlivů mnohem složitější.

Něco zaniká, umírá, rozpadá se, rozbíjí se ... možná Exponenciální?

Tady v reálu často narazíme i na rozdělení mnohem složitější. Přesně exponenciální bývá vzácností.

Syrová data bez dalších znalostí

Zde začneme obvykle tím, že si nakreslíme histogram. Zkušeným okem ho zhodnotíme a tipneme druh rozdělení. Postupně testujeme a dolaďujeme, až jsme se shodou spokojeni.

Zcela běžně se stává, že různí autoři si zvolí rozdělení různá a mají i trochu odlišné předpovědi. Rozhodčím pak bývá praxe.

Nahrazování

Někdy se rozdělení sice správné, ale výpočetně nešikovné, nahradí rozdělením podobným, se kterým se počítá lépe.

Typicky kombinační číslo $\binom{10000}{5000}$ v binomickém rozdělení je zcela nezvládnutelné.

Binomické nahrazujeme Poissonovým (se stejnou střední hodnotou), pokud je p malé a n velké.

Binomické nahrazujeme Normálním (se stejnou střední hodnotou a rozptylem), pokud je n velké a p není extrémně blízké nule nebo jedničce.

Poissonovo nahrazujeme Normálním $N(\lambda, \lambda)$, pokud λ není příliš malé.