Seriál – Pravděpodobnost a matematická statistika 1

Teorie pravděpodobnosti je nesmírně krásnou vědou. Zatímco většina odvětví matematiky se věnuje pojmům a veličinám, jež jsou „přesně“ dány, umožňuje teorie pravděpodobnosti hovořit i o nejistotě a neurčitosti. A protože v reálném světě je více věcí nejistých nežli jistých, má široké uplatnění v praxi.

Seriál bude mít tři části. V první části zavedeme základní pojmy a probereme to, co někteří z vás ještě nestihli probrat ve škole. Ve druhé si uděláme výlet do světa náhodných veličin a podmíněných nezávislostí. Třetí část se bude věnovat praxi matematické statistiky: normálnímu rozdělení, zákonu velkých čísel a centrální limitní větě.

Anžto jsou znalosti středoškolských studentů v této oblasti značně rozdílné, předem se omlouvám jak těm, kteří se při čtení prvního dílu seriálu budou nudit, tak těm, kteří se v něm zcela ztratí. Pokud nebudete rozumět nějaké definici či větě, podívejte se na osvětlující příklad za ní. Pokud vám scházejí středoškolské znalosti z kombinatoriky a pojmy jako permutace, kombinace či variace vám nic neříkají, doporučuji si je doplnit. Ať už z nějaké lepší středoškolské učebnice nebo z internetu. Můžete zkusit kupříkladu

`http://www.czech-ware.net/mathes/zobuceb.aspx?zob=kombinatorika`

nebo stručný, ale všezahrnující textík

`http://mks.mff.cuni.cz/knihovna/kombinatorika.pdf`.

Klasická pravděpodobnost – verze pro mateřské školy

Mějme množinu všech možných výsledků nějakého náhodného pokusu. Tuto množinu budeme nazývat *množina elementárních jevů* a značit Ω. O jejích podmnožinách $A \subset \Omega$ budeme hovořit jako o jevech, o jejích prvcích $\omega \in \Omega$ jako o elementárních jevech.

Příklad. Uvažujme výsledek hodu šestistěnnou kostkou[1]. Potom množina elementárních jevů jest

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Jevem je kupříkladu to, že padne sudé číslo, $A = \{2, 4, 6\}$. Elementární jevy jsou $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, $\{6\}$.

Poměrně jednoduchý případ nastává tehdy, je-li zaručeno, že Ω je konečná a že všechny elementární jevy jsou „stejně pravděpodobné“. Potom *pravděpodobností jevu* A rozumíme číslo

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

kde $|X|$ značí počet prvků množiny X.

Jev A nazveme *jistý*, pokud $\mathrm{P}(A) = 1$, a *nemožný*, pakliže $\mathrm{P}(A) = 0$. Jevem doplňkovým k jevu A míníme jev $A^c = \Omega \setminus A$, zjevně $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$.

[1] Pokud nebude řečeno jinak, všechny další v textu zmíněné kostky jsou šestistěnné.

Příklad. Tedy v minulém příkladu pravděpodobnost jevu A, že padne sudé číslo, je rovna

$$\mathrm{P}(A) = \frac{|A|}{|\Omega|} = \frac{3}{6} = 0{,}5.$$

Příklad. Na pouti se objevil stánek, kde si návštěvníci mohli vsadit na hod třemi kostkami[2]. Stánkaři zaplatili 10 Kč a pokud hodili v součtu alespoň 17 byla jim vyplacena stokoruna. Jinak vklad propadal. Jaká je šance hráče na výhru?

Uvažujme jako náhodný pokus hod třemi kostkami a zajímejme se, zda je jejich součet alespoň 17. Množinu elementárních jevů Ω tvoří $6 \cdot 6 \cdot 6 = 216$ trojic možných hodů. Z nich čtyři, konkrétně $(6,6,6), (5,6,6), (6,5,6), (6,6,5)$, mají součet alespoň 17. Tudíž

$$\mathrm{P}(\text{součet je alespoň 17}) = \frac{4}{216} \doteq 0{,}019.$$

Jako nepříliš těžké cvičení se můžete pokusit přijít na to, jaká byla „průměrná“ výhra hráče, a kolik tedy průměrně na každé sázce vydělával stánkař.

Příklad. Mějme 6 párů (tzn. 12 kusů) střevíců. Náhodně z nich vyberme 5 střevíců. Jaká je pravděpodobnost, že mezi nimi bude alespoň jeden pár?

Množinou elementárních jevů Ω jsou všechny (neuspořádané) pětice vybrané z množiny všech bot $\{a1, a2, b1, b2, c1, c2, d1, d2, e1, e2, f1, f2\}$, kde stejné písmeno značí příslušnost k témuž páru. Takovýchto pětic je $\binom{12}{5} = 792$.

Nás zajímá pravděpodobnost jevu A, že ve výběru bude alespoň jeden pár. Tu je ale značně nesnadné spočítat. Co je však mnohem snazší, je spočítat pravděpodobnost doplňkového jevu A^c, že tam žádný pár nebude,

$$\mathrm{P}(A^c) = \frac{2^5 \cdot \binom{6}{5}}{\binom{12}{5}} = \frac{192}{792} \doteq 0{,}24,$$

neboť vybraných pět střevíců musí patřit do pěti různých párů (a pět párů z šesti lze vybrat $\binom{6}{5}$ způsoby) a u každého páru máme na výběr ze dvou střevíců.

Odtud už snadno dopočteme pravděpodobnost samotného jevu

$$\mathrm{P}(A) = 1 - P(A^c) = \frac{600}{792} \doteq 0{,}76.$$

Cvičení. "Když se náhodně vyberou dvě mé děti, je stejná pravděpodobnost, že mají stejné pohlaví, jako pravděpodobnost, že mají různé pohlaví," pravil sultán.

„Jaká je pravděpodobnost, že to budou dvě dívky?“ otázal se kalif.

„Stejná jako že náhodně vybrané dítě bude chlapec,“ odvětil sultán.

Pokud jste dobře pochopili, co to je pravděpodobnost, a umíte řešit kvadratické rovnice, nemělo by pro vás být těžké určit, kolik má sultán dětí.

Podmínění náhodným jevem, nezávislost jevů

[2] Příhoda není tak úplně imaginární. Podobný stánek jsem skutečně viděl na Mohelnickém dostavníku.

Nechť máme dva jevy A a B, $\mathrm{P}(B) > 0$. *Podmíněnou pravděpodobnost jevu A* za podmínky, že nastal jev B, definujeme jako

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

Dva jevy A a B nazveme *nezávislé*, jestliže

$$\mathrm{P}(A \cap B) = \mathrm{P}(A) \cdot \mathrm{P}(B).$$

Pokud tyto jevy nejsou nemožné, snadno z nezávislosti odvodíme

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = P(A),$$
$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)} = P(B).$$

Vidíme tedy, že u nezávislých jevů to, že nastal jev A, neovlivní pravděpodobnost, že nastane jev B, a naopak.

Příklad. Uvažujme hod kostkou: Nechť A je jev, že padne sudé číslo. Nechť B je jev, že padne jednička nebo dvojka. Potom si snadno můžete ověřit, že jevy A a B jsou nezávislé.

Rozklad množiny elementárních jevů

Jevy $B_1, B_2, \ldots, B_m$ nazveme rozkladem množiny elementárních jevů, jestliže průnik libovolných dvou z nich je prázdná množina a sjednocení všech je Ω.

Jinak řečeno, každý elementární jev $\omega \in \Omega$ náleží do právě jednoho z těchto jevů.

Příklad. U hodu kostkou můžeme uvažovat rozklad na jevy, že padne sudé/liché číslo. Nebo rozklad na jevy $B_1 = \{1\}$, $B_2 = \{3, 4, 6\}$, $B_3 = \{2, 5\}$.

Věta. *Nechť $B_1, B_2, \ldots, B_m$ je rozklad množiny elementárních jevů a žádný z těchto jevů není nemožný. Potom pro libovolný jev A platí*

$$\mathrm{P}(A) = \sum_{i=1}^{m} \mathrm{P}(B_i) \cdot \mathrm{P}(A|B_i).$$

Důkaz. Důkaz je snadný:

$$\sum_{i=1}^{m} \mathrm{P}(B_i) \cdot \mathrm{P}(A|B_i) = \sum_{i=1}^{m} \mathrm{P}(A \cap B_i) = \mathrm{P}(A).$$

Příklad. K lékaři chodí nemocní i zdraví pacienti. Ví se, že 60% z pacientů, kteří přijdou do ordinace, je nemocných. Pan doktor není tvor neomylný, a tak se u 20% pacientů (ať už zdravých či nemocných) splete. Kolik procent pacientů označí pan doktor jako nemocné?

Tupě aplikujme předchozí větu s rozkladem pacient je/není nemocný:

$$\mathrm{P}(\text{pacient označen za nemocného}) =$$

$$\mathrm{P}(\text{pacient označen za nemocného}|\text{pacient je nemocný}) \cdot \mathrm{P}(\text{pacient je nemocný}) +$$

$$\mathrm{P}(\text{pacient označen za nemocného}|\text{pacient je zdravý}) \cdot \mathrm{P}(\text{pacient je zdravý}) =$$

$$= 0{,}8 \cdot 0{,}6 + 0{,}2 \cdot 0{,}4 = 0{,}56$$

Povšimněme si, že nás v tomto případě vůbec nemuselo zajímat, jak vypadá množina Ω.

Následující věta, jež nese jméno po ctihodném otci Bayesovi, slouží k tomu, abychom mohli „prohodit“ jevy v podmiňování:

Věta (Bayesova). *Nechť B_1, B_2, ..., B_m je rozklad množiny elementárních jevů a žádný z těchto jevů není nemožný. Potom pro libovolný jev A, $\mathrm{P}(A) > 0$ platí*

$$\mathrm{P}(B_k|A) = \frac{P(B_k) \cdot \mathrm{P}(A|B_k)}{\sum_{i=1}^{m} \mathrm{P}(B_i) \cdot \mathrm{P}(A|B_i)}\,, \qquad k = 1, \ldots, m.$$

Důkaz. Důkaz je opět velmi snadný:

$$\mathrm{P}(B_k|A) = \frac{\mathrm{P}(A \cap B_k)}{\mathrm{P}(A)} = \frac{\mathrm{P}(B_k) \cdot \mathrm{P}(A|B_k)}{\sum_{i=1}^{m} \mathrm{P}(B_i) \cdot \mathrm{P}(A|B_i)}.$$

Příklad. Předchozí věta je značně oblíbená v lékařské diagnostice. Uvažujme testování, zda pacient má AIDS. Máme k dispozici vcelku spolehlivý test, který dá v 95 procentech případů správnou odpověď a v 5 procentech mylnou. Víme, že ve sledované populaci má AIDS 0,5% lidí. Test dopadl pozitivně[3]. Jaká je pravděpodobnost, že pacient nemoc skutečně má?

Označme N jev, že pacient je nemocný, a Z, že je zdravý. Označme $\oplus$ jev, že test dopadl pozitivně, a tupě dosaďme do předchozí věty:

$$\mathrm{P}(N|\oplus) = \frac{\mathrm{P}(N) \cdot \mathrm{P}(\oplus|N)}{\mathrm{P}(N) \cdot \mathrm{P}(\oplus|N) + \mathrm{P}(Z) \cdot \mathrm{P}(\oplus|Z)} = \frac{0{,}005 \cdot 0{,}95}{0{,}005 \cdot 0{,}95 + 0{,}995 \cdot 0{,}05} \doteq 0{,}087.$$

Odpověď tedy zní: velice nízká, ani ne desetina. Poněkud překvapivý výsledek u tak spolehlivého testu, nezdá se vám? Zkuste přijít na to, proč tomu tak je.

Cvičení. Na procvičení doporučuji následující příklad: Američan, Rus a Čech se hádají, kdo vydrží déle pod vodou. Američan tvrdí, že dvě minuty (pravděpodobnost přežití 0,8). Rus tvrdí, že pět minut (pravděpodobnost přežití 0,5). Čech tvrdí, že deset minut (pravděpodobnost přežití 0,1). Všichni se potopili a jeden z nich to nepřežil. Jaká je pravděpodobnost, že to byl Čech?

Námět k zamyšlení

Jaká je hlavní nevýhoda námi definované pravděpodobnosti? Tušíte správně – v tom, že všechny elementární jevy nemusí být nutně stejně pravděpodobné. Představte si třeba „falešnou kostku“, na které padá šestka o něco častěji než jednička.

Pokud se tedy chceme bavit i o „falešných kostkách“, musíme opustit mateřskou školku. Ale o tom zase až příště ...

[3] To znamená, že svědčí pro to, že pacient nemoc má.

Seriál – Pravděpodobnost a matematická statistika 2

Motto: Bohužel vaše planeta je jednou z těch, které byly určeny k demolici. Tento proces započne za necelé dvě vaše pozemské minuty. **Nepodléhejte panice!** Děkuji vám.
(z poselství Vogona Jetze, Stopařův průvodce po galaxii, Dougles Adams)

V druhé část seriálu rozšíříme naše znalosti z části první. Místo náhodných jevů budeme pracovat s náhodnými veličinami a krom obyčejné nezávislosti se budeme zabývat i nezávislostí podmíněnou.

Zkrátka a dobře, všechno bude mnohem zajímavější a dobrodružnější nežli minule.

Podmíněná nezávislost

Minule jsme zavedli pojem „nezávislost jevů“. Rozšíříme jej nyní na podmíněnou nezávislost, tedy nezávislost dvou jevů za podmínky, že nastal jev třetí:

Definice. Jevy A a B nazveme *nezávislé za podmínky, že nastal jev* C ($\mathrm{P}(C) > 0$), pakliže platí

$$\mathrm{P}(A \cap B|C) = \mathrm{P}(A|C)\,\mathrm{P}(B|C).$$

Příklad. Nechť A je jev, že se v náhodně vybraném okrese vyskytuje značně nadprůměrné množství čápů, a B jev, že se tu rodí značně nadprůměrné množství dětí, přičemž význam spojení „značně nadprůměrné“ ponechávám čtenářově fantazii.

Tabulka pravděpodobností A a B by mohla vypadat například takto[4]:

$$P(A \cap B) = \tfrac{7}{64} \qquad P(A \cap B^c) = \tfrac{13}{64}$$
$$P(A^c \cap B) = \tfrac{13}{64} \qquad P(A^c \cap B^c) = \tfrac{31}{64}$$

Odtud snadno spočítáme $\frac{7}{64} = \mathrm{P}(A \cap B) \neq \mathrm{P}(A) \cdot \mathrm{P}(B) = \frac{6{,}25}{64}$. Jevy A a B tedy nejsou nezávislé, ke všemu si můžete spočítat, že pravděpodobnost výskytu enormního množství novorozenců je vyšší, pokud nastal výskyt enormního množství čápů, než když tomu tak nebylo.

Teď se možná smějete a myslíte si, že takto by to ve skutečnosti být nemohlo. A to se mýlíte! Skutečně existuje výzkum, který potvrzuje statisticky významnou souvislost mezi počtem čápů a novorozenců. Znamená to tedy, že děti nosí čáp?

Nikoli, znamená to jen, že příčinnou souvislost dvou jevů nelze zkoumat bez ohledu na jevy ostatní. Označme si náš hypotetický vysvětlující jev třeba C. Může odrážet vliv životního prostředí, zeměpisné šířky, ... Jeho pravděpodobnost nechť je $\mathrm{P}(C) = \frac{1}{4}$ a pravděpodobnosti A a B podmíněno C jsou

$$P(A \cap B|C) = \tfrac{1}{4} \qquad P(A \cap B^c|C) = \tfrac{1}{4} \qquad P(A \cap B|C^c) = \tfrac{1}{16} \qquad P(A \cap B^c|C^c) = \tfrac{3}{16}$$
$$P(A^c \cap B|C) = \tfrac{1}{4} \qquad P(A^c \cap B^c|C) = \tfrac{1}{4} \qquad P(A^c \cap B|C^c) = \tfrac{3}{16} \qquad P(A^c \cap B^c|C^c) = \tfrac{9}{16}$$

Jako snadné cvičení si můžete ověřit, že A a B jsou podmíněně nezávislé za podmínky C, a obě tabulky jsou navzájem konzistentní. A jako první statistikovo přikázání si zapište: „Nebudu z popření nezávislosti dvou jevů vyvozovat přímou příčinnou souvislost mezi nimi!“. Kéž by toto přikázání ctili všichni praktičtí výzkumníci.

[4] Připomeňme, že A^c značí doplňkový jev k A, tedy $\Omega \setminus A$.

Klasická pravděpodobnost – rozšíření na různě pravděpodobné elementární jevy

Mějme opět množinu elementárních jevů $\Omega = \{\omega_1, \ldots, \omega_n\}$. Každému elementárnímu jevu ω_i přiřadíme pravděpodobnost $p_i \in \langle 0, 1\rangle$, že tento jev nastane, přičemž součet těchto pravděpodobností je nutně jedna:

$$\sum_{i=1}^{n} p_i = p_1 + p_2 + \cdots + p_n = 1 \,.$$

Potom *pravděpodobnost jevu* A je rovna součtu všech pravděpodobností elementárních jevů náležících do A:

$$\mathrm{P}(A) = \sum_{i:\ \omega_i \in A} p_i.$$

Příklad. Uvažujme falešnou kostku, na níž padá šestka s pravděpodobností $\frac{1}{2}$, zatímco ostatní čísla s pravděpodobností $\frac{1}{10}$. Množina elementárních jevů $\Omega = \{1, 2, 3, 4, 5, 6\}$. Jev, že padne sudé číslo, $A = \{2, 4, 6\}$:

$$\mathrm{P}(A) = \frac{1}{10} + \frac{1}{10} + \frac{1}{2} = 0{,}7.$$

Poznámka. Povšimněte si toho, že pokud jsou všechny elementární jevy stejně pravděpodobné $p_1 = p_2 = \cdots = p_n = \frac{1}{|\Omega|}$, dostáváme teorii pravděpodobnosti pro mateřskou školku.

Dále si povšimněte toho, že jsme nijak nevyužili faktu, že Ω je konečná, a pokud tedy umíme pracovat s nekonečnými součty, můžeme (a také budeme) mít Ω nekonečnou spočetnou[5].

Cvičení. Ověřte, že při takto modifikované definici pravděpodobnosti jevu mají smysl všechny definice z minula a platí všechny věty.

Výše uvedenou definici typicky využijeme, když se snažíme popsat nějakou náhodnou věc z praxe, jež je číselné povahy – například počet narozených dětí, počet uzdravených pacientů, ... Zde je použití stejně pravděpodobných elementárních jevů přinejmenším krkolomné.

Nechť tedy Ω je spočetnou podmnožinou reálných čísel[6] a P pravděpodobnost na Ω. Potom budeme mluvit o náhodné veličině X s rozdělením P a $\mathrm{P}(X = k)$ značí pravděpodobnost elementárního jevu k, $\mathrm{P}(k)$; resp. $\mathrm{P}(X \in A)$ značí pravděpodobnost jevu A, $\mathrm{P}(A)$.

Náhodné veličiny budeme obvykle značit velkými písmeny z konce abecedy.

Příklad. U hodu šestistěnnou kostkou můžeme označit číslo, jež padne, jako náhodnou veličinu X. Poté můžeme hovořit o pravděpodobnosti, že X je tři, $\mathrm{P}(X = 3) = \frac{1}{6}$; resp. že X je sudé, $\mathrm{P}(X \in \{2, 4, 6\}) = \frac{1}{2}$.

Definice. *Střední hodnotou* náhodné veličiny X míníme

$$\sum_{x \in \Omega} x \cdot P(X = x)$$

a značíme $\mathrm{E}\,X$. Jedná se tedy o vážený průměr možných hodnot X, kde váhy tvoří pravděpodobnosti těchto hodnot.

---

[5] Pokud této poznámce ani zbla nerozumíte, poněvadž netušíte, co znamená slovíčko „spočetný“, tak si s ní nelamte hlavu. Značně volně řečeno, spočetná je taková množina, která má „stejně nebo méně“ prvků nežli množina celých čísel.

[6] tj. například celá čísla, přirozená čísla, konečně mnoho čísel, ...

V angličtině se místo o střední hodnotě mluví o očekávané hodnotě. A právě takový je její význam – značí, kolik asi tak můžeme očekávat, že bude průměr z mnoha realizací X.

Příklad. Nechť X je náhodná veličina udávající hod spravedlivou kostkou. Potom

$$\mathrm{E}\,X = 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + 3\cdot\frac{1}{6} + 4\cdot\frac{1}{6} + 5\cdot\frac{1}{6} + 6\cdot\frac{1}{6} = 3{,}5.$$

Nechť Y je náhodná veličina udávající hod falešnou kostkou z minulého příkladu. Potom

$$\mathrm{E}\,Y = 1\cdot\frac{1}{10} + 2\cdot\frac{1}{10} + 3\cdot\frac{1}{10} + 4\cdot\frac{1}{10} + 5\cdot\frac{1}{10} + 6\cdot\frac{1}{2} = 4{,}5.$$

Definice. *Rozptylem* náhodné veličiny X míníme

$$\mathrm{E}\,X^2 - (\mathrm{E}\,X)^2 = \mathrm{E}(X - \mathrm{E}\,X)^2 = \sum_{x\in\Omega}(x - \mathrm{E}\,X)^2\cdot P(X = x)$$

a značíme $\mathrm{Var}\,X$.

Rozptyl je mírou toho, jak se X drží kolem své střední hodnoty, resp. jaká je v něm ukryta neurčitost. Čím větší rozptyl, tím větší neurčitost. Jako snadné cvičení můžete ověřit, že $\mathrm{Var}\,X = 0$ právě tehdy, když $X = EX$ s pravděpodobností jedna, a že rozptyl je vždy nezáporný.

Poznámka. Pozorný čtenář si zajisté povšiml, že předchozí definice není zcela korektní, neboť není zcela jasné, co vlastně míníme výrazem $\mathrm{E}\,X^2$, resp. $\mathrm{E}(X - \mathrm{E}\,X)^2$, neboť umíme zatím spočítat střední hodnotu náhodné veličiny X, ale nikoli funkce této veličiny $f(X)$.

Pozornému čtenáři je nutno zcela dát zcela za pravdu a přiznat, že autor zde zneužívá toho, že intuitivně je jasné, co se tím „myslí“. Poctivě by bylo potřeba dokázat, že pokud X je náhodná veličina, potom $X' = f(X)$ je opět náhodná veličina (na nějakém jiné množině Ω' a s jiným rozdělením P'), a tudíž můžeme spočítat její střední hodnotu $\mathrm{E}'\,X'$.

Pro praktický výpočet je možno využít toho, že

$$\begin{aligned}\mathrm{E}\,f(X) &= \sum_{x\in\Omega} f(x)\cdot P(X = x),\\ \mathrm{Var}\,f(X) = \mathrm{E}\,f(X)^2 - (\mathrm{E}\,f(X))^2 &= \sum_{x\in\Omega}(f(x) - \mathrm{E}\,f(X))^2\cdot P(X = x).\end{aligned}$$

Důkaz je ponechán pozornému čtenáři.

Příklad. Kdyby nás tedy zajímalo, jaká se střední hodnota třetí mocniny čísla, jež nám padne na spravedlivé kostce (označme jej X), zvětšeného o jedničku, dostali bychom

$$\mathrm{E}(X+1)^3 = 8\cdot\frac{1}{6} + 27\cdot\frac{1}{6} + 64\cdot\frac{1}{6} + 125\cdot\frac{1}{6} + 216\cdot\frac{1}{6} + 343\cdot\frac{1}{6} = 130{,}5.$$

Příklad. Nechť X je náhodná veličina udávající hod spravedlivou kostkou, potom

$$\begin{aligned}\mathrm{E}\,X^2 &= 1^2\cdot\frac{1}{6} + 2^2\cdot\frac{1}{6} + 3^2\cdot\frac{1}{6} + 4^2\cdot\frac{1}{6} + 5^2\cdot\frac{1}{6} + 6^2\cdot\frac{1}{6} \doteq 15{,}2,\\ \mathrm{Var}\,X &= \mathrm{E}\,X^2 - (\mathrm{E}\,X)^2 \doteq 2{,}9.\end{aligned}$$

Věta. *Pro náhodnou veličinu X a libovolné nenáhodné konstanty a a b platí:*

$$\mathrm{E}(aX+b) = a(\mathrm{E}\,X) + b,$$
$$\mathrm{Var}(aX+b) = a^2\,\mathrm{Var}\,X.$$

Důkaz.

$$\mathrm{E}(aX+b) = \sum_{x\in\Omega}(ax+b)\cdot\mathrm{P}(X=x) = b\cdot\sum_{x\in\Omega}\mathrm{P}(X=x) + a\cdot\sum_{x\in\Omega}x\cdot\mathrm{P}(X=x) = b + a\cdot\mathrm{E}\,X.$$

Obdobně

$$\mathrm{Var}(aX+b) = \sum_{x\in\Omega}\left((ax+b)-(a\,\mathrm{E}\,X+b)\right)^2\cdot\mathrm{P}(X=x) =$$

$$= a^2\sum_{x\in\Omega}((x-\mathrm{E}\,X)^2\cdot\mathrm{P}(X=x) = a^2\,\mathrm{Var}\,X.$$

Poznámka. Pozorného čtenáře zajisté napadlo, že sice umíme už pracovat s jednou náhodnou veličinou, ale nikoli se dvěma. Nedokážeme tedy říci, jaká je například pravděpodobnost $\mathrm{P}(X=x, Y=y)$ nebo střední hodnota $\mathrm{E}\,XY$.

To je pochopitelně chyba a nelze ji jednoduše napravit. Abychom mohli mluvit o vzájemné interakci více náhodných veličin X_1, X_2, …, X_k, musí být nadefinovány na téže množině elementárních jevů Ω, tedy musíme pracovat s jedinou náhodnou veličinou – uspořádanou k-ticí $(X_1, X_2, \ldots, X_k)$ a to si žádá opět trochu obecnější definici pravděpodobnosti. Pro názornost tu provedu konstrukci pro případ dvou náhodných veličin X a Y, přičemž pro větší počet se užije obdobného triku.

Mějme dvě náhodné veličiny: X nabývající hodnot z Ω_X a Y nabývající hodnot z Ω_Y. Potom za množinu elementárních jevů Ω budeme brát množinu uspořádaných dvojic $\{(x,y) : x\in\Omega_X, y\in\Omega_Y\}$ a pod $\mathrm{P}(X=x, Y=y)$ budeme rozumět $\mathrm{P}(x,y)$, kde P je nějaká pravděpodobnost na Ω. V tomto případě $\mathrm{P}(X=x)$ značí

$$\mathrm{P}(X=x, Y=\text{„cokoli“}) = \mathrm{P}(X=x, Y\in\Omega_Y) = \sum_{y\in\Omega_y}\mathrm{P}(X=x, Y=y).$$

Povšimněme si, že pokud se budeme zabývat pouze X, Y či libovolnou nenáhodnou funkcí $f(X,Y)$, dostaneme náhodnou veličinu podle dřívější definice.

Věta. *Nechť X a Y jsou náhodné veličiny. Potom*

$$\mathrm{E}(X+Y) = \mathrm{E}\,X + \mathrm{E}\,Y.$$

Důkaz. Požadované tvrzení dostaneme snadnou úpravou:

$$\mathrm{E}(X+Y) = \sum_{x\in\Omega_X, y\in\Omega_Y}(x+y)\,\mathrm{P}(X=x, Y=y) =$$

$$= \sum_{x\in\Omega_X}\sum_{y\in\Omega_Y}x\,\mathrm{P}(X=x, Y=y) + \sum_{y\in\Omega_Y}\sum_{x\in\Omega_X}y\,\mathrm{P}(X=x, Y=y) =$$

$$= \sum_{x \in \Omega_X} x \,\mathrm{P}(X = x) + \sum_{y \in \Omega_Y} y \,\mathrm{P}(Y = y) = \mathrm{E}\, X + \mathrm{E}\, Y.$$

Definice. Řekneme, že dvě náhodné veličiny X a Y jsou *nezávislé*, jestliže pro všechna x, y platí

$$\mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x) \cdot \mathrm{P}(Y = y),$$

neboli jevy $\{X = x\}$ a $\{Y = y\}$ jsou nezávislé.

Obdobně, jestliže X_1, X_2, ... , X_n jsou náhodné veličiny nabývající hodnot z Ω_1, Ω_2, ... , Ω_n, potom je nazveme *navzájem nezávislé*, jestliže pro libovolné $x_1 \in \Omega_1$, $x_2 \in \Omega_2$, ... , $x_n \in \Omega_n$ platí

$$\mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \mathrm{P}(X_1 = x_1) \cdot \mathrm{P}(X_2 = x_2) \cdot \ldots \cdot \mathrm{P}(X_n = x_n).$$

Příklad. Například pokud hodíme n kostkami, tak čísla, jež padnou na jednotlivých kostkách, jsou navzájem nezávislé náhodné veličiny (a to i v případě falešných kostek).

Cvičení. Ověřte, že pokud máme n navzájem nezávislých veličin, pak libovolných k ($k < n$) z nich vybraných jsou opět navzájem nezávislé náhodné veličiny.

Věta. *Nechť X a Y jsou nezávislé náhodné veličiny, potom*

i) $\mathrm{E}(XY) = \mathrm{E}\, X \,\mathrm{E}\, Y$,
ii) $\mathrm{Var}(X + Y) = \mathrm{Var}\, X + \mathrm{Var}\, Y$.

Cvičení. Dokažte tuto větu: použijte nezávislost a v druhé části definici rozptylu $\mathrm{Var}(X + Y) = \mathrm{E}(X + Y)^2 - (\mathrm{E}(X + Y))^2$. Pokud vám to přišlo moc jednoduché, formulujte obdobné tvrzení pro více náhodných veličin.

Příklady nejběžnějších rozdělení

Aby nebylo nutné pokaždé znovu odvozovat vlastnosti těch nejběžnějších náhodných veličin, rozlišujeme několik jejich „typů“.

Alternativní rozdělení

Definice. Řekneme, že náhodná veličina X má *alternativní rozdělení* s parametrem p, jestliže X nabývá pouze hodnot 0 a 1 a $\mathrm{P}(X = 1) = p$. Značíme $X \sim Alt(p)$. Pro $X = 1$ budeme mluvit o „úspěchu“, pro $X = 0$ o „neúspěchu“.

Příklad. Typickým příkladem náhodné veličiny s alternativním rozdělením je indikátor nějakého jevu A, tj. náhodná veličina, jež se rovná jedna, pokud jev A nastal (např. padla šestka, padlo sudé číslo, ...), a nula, pokud jev A nenastal.

Střední hodnotu $X \sim Alt(p)$ spočteme snadno:

$$\mathrm{E}\, X = 1 \cdot p + 0 \cdot (1 - p) = p,$$

rozptyl taktéž:

$$\mathrm{Var}\, X = \mathrm{E}\, X^2 - (\mathrm{E}\, X)^2 = 1^2 \cdot p + 0^2 \cdot (1 - p) - p^2 = p(1 - p).$$

Binomické rozdělení

Definice. Součet n navzájem nezávislých náhodných veličin s rozdělením $Alt(p)$ nazveme náhodnou veličinou X s *binomickým rozdělením* a značíme $X \sim Binom(n,p)$. Jinak řečeno, X značí počet úspěchů v n nezávislých veličinách s rozdělením $Alt(p)$.

Příklad. Pokud hodíme n spravedlivými kostkami, potom počet šestek, jež nám padne, má binomické rozdělení, konkrétně $Binom(n, \frac{1}{6})$.

Z definice je jasné, že $X \sim Binom(n,p)$ nabývá hodnot 0,1, ... , n. Pravděpodobnost, že nabude hodnoty k, tedy že alternativní náhodné veličiny, jejichž je součtem, obsahují právě k úspěchů, je rovna počtu možných výběrů k prvků z n, t.j. $\binom{n}{k}$, krát pravděpodobnost každé z posloupností k úspěchů a $(n-k)$ neúspěchů, t.j. $p^k(1-p)^{(n-k)}$. Dohromady tedy

$$\mathrm{P}(X=k) = \binom{n}{k} p^k (1-p)^{(n-k)}.$$

Střední hodnotu a rozptyl spočítáme podle věty pro součet náhodných veličin, u rozptylu navíc využijeme nezávislosti alternativně rozdělených veličin Y_1, Y_2, ... , Y_n, jejichž je X součtem,

$$\mathrm{E}\,X = \mathrm{E}\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \mathrm{E}\,Y_i = np,$$
$$\mathrm{Var}\,X = \mathrm{Var}\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \mathrm{Var}\,Y_i = np(1-p).$$

Příklad. Známý hráč Samuel Pepys kdysi přišel za jistým Johnem Smithem s následující úlohou: Co si myslíte, že je snazší – hodit 6 kostkami alespoň jednu šestku, hodit 12 kostkami alespoň dvě šestky nebo hodit 18 kostkami alespoň tři šestky?

Ti, co mají dobrou intuici, by měli řešení této úlohy uhádnout i bez počítání. Pro ty ostatní, označme počet šestek při hodu 6, 12 a 18 kostkami po řadě X, Y a Z. Zřejmě $X \sim Binom(6, \frac{1}{6})$, $Y \sim Binom(12, \frac{1}{6})$, $Z \sim Binom(18, \frac{1}{6})$. Tudíž

$$\mathrm{P}(X \geq 1) = 1 - \mathrm{P}(X=0) = 1 - \binom{6}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^6 \doteq 0{,}67,$$
$$\mathrm{P}(Y \geq 2) = 1 - \mathrm{P}(Y=0) - \mathrm{P}(Y=1) = 1 - \binom{12}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^{12} - \binom{12}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^{11} \doteq 0{,}62,$$
$$\mathrm{P}(Z \geq 3) = 1 - \mathrm{P}(Z=0) - \mathrm{P}(Z=1) - \mathrm{P}(Z=2) \doteq 0{,}60.$$

Geometrické rozdělení

Definice. Mějme posloupnost nezávislých náhodných veličin s rozdělením $Alt(p)$: $Y_1, Y_2, \ldots$ Náhodnou veličinou s *geometrickým rozdělením* s parametrem p rozumíme počet neúspěchů v předchozí posloupnosti před prvním úspěchem. Značíme $Geom(p)$.

Příklad. Počet hodů kostkou, nežli[7] nám padne šestka, má rozdělení $Geom(\frac{1}{6})$.

Pravděpodobnost, že $X \sim Geom(p)$ nabude hodnoty k, snadno spočteme jako pravděpodobnost, že nastane k neúspěchů a po nich jeden úspěch:

$$\mathrm{P}(X=k) = (1-p)^k p.$$

[7] Tzn. hod, ve kterém nám padne šestka, se už nepočítá.

Vzorec pro střední hodnotu a rozptyl zde uvedeme bez důkazu:

$$\mathrm{E}\,X = \frac{1-p}{p},$$
$$\mathrm{Var}\,X = \frac{1-p}{p^2}.$$

Příklad. V každé žvýkačce Pedro je ukrytý některý z n obrázků Harry Pottera. Předpokládejme, že každý obrázek se ve žvýkačkách vyskytuje se stejnou pravděpodobností. Jaká je střední hodnota počtu žvýkaček, které musíme zakoupit, abychom získali všechny Harryho obrázky?

Označme X náhodnou veličinu udávající počet žvýkaček, jež musíme zakoupit. Označme Y_i počet žvýkaček, zakoupených po dosažení i obrázků do doby, nežli[8] získáme $(i+1)$ obrázků.

Zjevně

$$X = 1 + Y_1 + 1 + Y_2 + 1 + Y_3 + 1 + \ldots Y_{n-1} + 1$$

a veličiny Y_i mají rozdělení $Geom(\frac{n-i}{n})$. Tudíž

$$\mathrm{E}\,X = n + \sum_{i=1}^{n-1} \mathrm{E}\,Y_i = n + \sum_{i=1}^{n-1} \frac{i}{n-i} = \sum_{i=1}^{n} \frac{n}{i}.$$

Tedy například pro $n = 20$ vyjde $\mathrm{E}\,X = 72$.

Seriál – Pravděpodobnost a matematická statistika 3

Motto: Po aplikaci preparátu B se 33,3% kuřat uzdravilo, 33,3% kuřat uhynulo a o zbývajících 33,3% kuřat nejsme schopni poskytnout uspokojující informaci. Dosud se nám nepodařilo to třetí kuře chytit.

V závěrečné části seriálu opustíme šedou teorii a vrhneme se do mnohabarevné praxe. Předem varuji, že se zde setkáte s větší mírou neurčitosti a menší mírou preciznosti, než na jakou jste asi zvyklí a než jaká je po vás v korespondenčním semináři vyžadována. Tuto daň za dosažení užitečných a zajímavých výsledků je nutné zaplatit, neboť teorie, jež se za těmito výsledky skrývá, zdaleka přesahuje rámec středoškolské matematiky.

Poissonovo rozdělení

K běžným rozdělením z minula si ještě přidáme tzv. *Poissonovo* rozdělení:

Definice. Řekneme, že náhodná veličina X má Poissonovo rozdělení s parametrem $0 < \lambda < \infty$ (značíme $X \sim Poiss(\lambda)$), pokud nabývá pouze nezáporných celých čísel a to s pravděpodobností

$$\mathrm{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k = 0, 1, \ldots,$$

[8] Ten s $i+1$. obrázkem se už nepočítá.

kde $e \doteq 2{,}718$ je základ přirozeného logaritmu.

Poissonovo rozdělení se používá k vyjádření počtu náhodných událostí, které nastávají s nějakou intenzitou (tu vyjadřuje parametr λ), tedy například pro počet telefonních hovorů na ústředně v daném časovém intervalu, nebo pro počet lidí, které zvládne obsloužit pokladní v supermarketu v danou hodinu.

Bez důkazu si uveďme střední hodnotu a rozptyl pro $X \sim Poiss(\lambda)$:

$$\mathrm{E}\, X = \mathrm{Var}\, X = \lambda.$$

Příklad. Pokladní v supermarketu obslouží za pět minut v průměru jednoho zákazníka. Pokud budeme předpokládat, že počet obsloužených zákazníků v pěti minutách má Poissonovo rozdělení s parametrem $\lambda = 1$, spočtěte pravděpodobnost, že pokladní se překoná a obslouží v příštích pěti minutách alespoň tři zákazníky.

Řešení je jednoduché, spočteme pravděpodobnost doplňkového jevu a odečteme ji od jedničky:

$$\mathrm{P}(\text{obslouží alespoň tři}) =$$

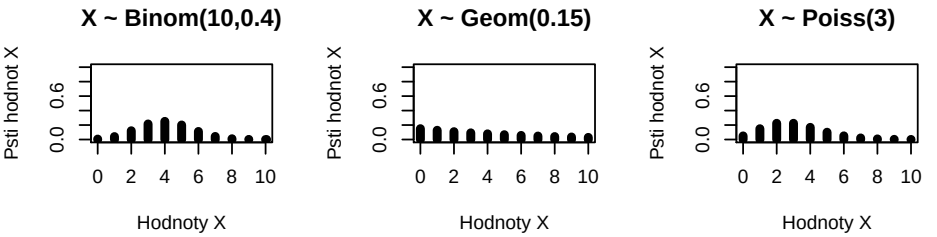
$$= 1 - \mathrm{P}(\text{obslouží dva}) - \mathrm{P}(\text{obslouží jednoho}) - \mathrm{P}(\text{neobslouží ani jednoho}) =$$

$$= 1 - e^{-1} \cdot \frac{1}{2} - e^{-1} \cdot \frac{1}{1} - e^{-1} \cdot \frac{1}{1} \doteq 0{,}08.$$

Pravděpodobnost, že pokladní trhne rekord, je tedy přibližně 8%.

Grafické znázornění

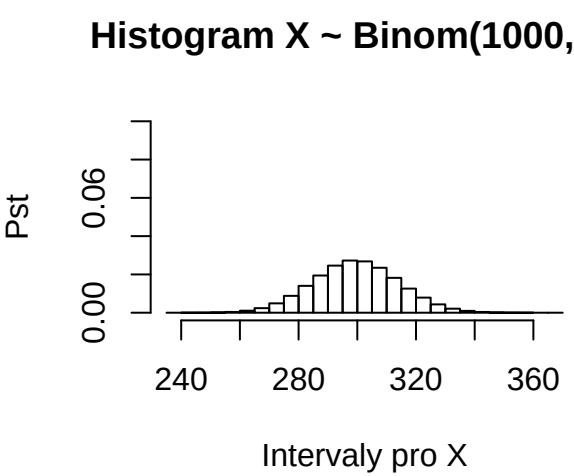
Možná už od minula máte pocit, že popis náhodné veličiny jakožto seznamu (či tabulky) pravděpodobností jednotlivých hodnot je sice přesný, ale zato nepříliš názorný. Nejjednodušším způsobem, jak náhodnou veličinu X graficky znázornit, je vzít jako vodorovnou osu seznam možných hodnot X (či alespoň těch nejvíce pravděpodobných) a ke každé z nich nakreslit sloupec, jehož výška je rovna pravděpodobnosti toho, že X nabude této hodnoty.



Ukázku na třech různých náhodných veličinách můžete vidět na obrázku výše.

Druhým možným způsobem, jenž oceníte především u veličin s velkým počtem možných hodnot, je použití tzv. *histogramu*. V tomto případě si vodorovnou osu rozdělíte na intervaly stejné délky d a ke každému intervalu I poté nakreslíte sloupec výšky $\mathrm{P}(X \in I)/d$.

Histogram je pro $d = 1$ určitou aproximací (odhadem) předchozího obrázku.



Ukázku pro binomické rozdělení s vysokým parametrem n můžete vidět výše.

Posloupnost i.i.d.

Představme si, že máme nějakou náhodnou veličinu X, která nám udává výsledek nějakého pokusu, např. hodu kostkou. Pokud tento pokus budeme několikrát opakovat, budou jeho výsledky $X_1, X_2, \ldots$ nezávislé náhodné veličiny, jež budou mít všechny stejné rozdělení jako X.

Definice. Posloupností *nezávislých stejně rozdělených veličin*[9] neboli posloupností *i.i.d.* (independent, identically distributed) míníme posloupnost náhodných veličin, jež jsou navzájem nezávislé a všechny mají stejné rozdělení X.

Příklad. Posloupností i.i.d. jsou třeba již výše zmiňované hody kostkou. Naopak posloupností i.i.d. není množství srážek v jednotlivé dny (když pršelo dneska, bude asi pršet i zítra, tzn. je porušena nezávislost) nebo vzrůst rostlin, z nichž část jsme hnojili a část ne (porušen předpoklad stejného rozdělení).

V praxi často nemusí být snadné rozhodnout, zda se o posloupnost i.i.d. jedná či nikoli.

Poznámka. V textu o výsledcích experimentů mohou vznikat nejasnosti ohledně toho, zda pod značením X myslíme náhodnou veličinu (t.j. něco neurčitého) nebo její naměřenou hodnotu (tj. reálné číslo). Domluvme se tedy, že náhodné veličiny budeme značit velkými písmeny, zatímco jejich realizace obvykle písmeny malými.

Otázkou, která nás pochopitelně bude zajímat, jest, kterak z pozorované posloupnosti i.i.d. x_1, $x_2, \ldots, x_n$ získat informaci o (neznámém) rozdělení, z něhož pocházejí. Například si představte, že jsme 100 krát hodili falešnou kostkou a zajímá nás, jak je tato kostka falešná (t.j. s jakou pravděpodobností padne které číslo).

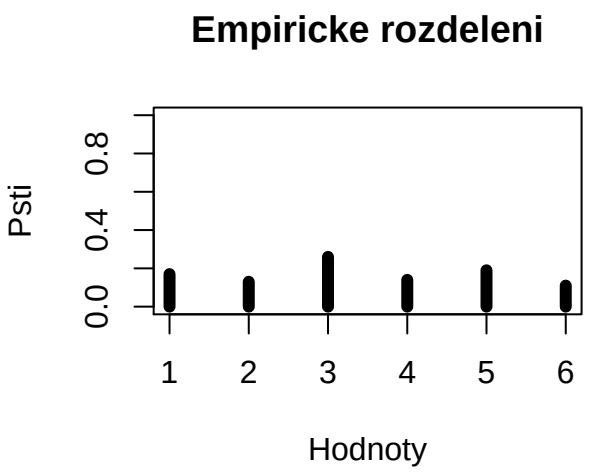
[9] Někdy se také mluví o tzv. náhodném výběru.

Definice. Empirickým rozdělením[10] k posloupnosti i.i.d. $X_1, \ldots, X_n$ budeme rozumět náhodnou veličinu $\hat{X}$ takovou, že

$$\mathrm{P}(\hat{X} = x) = \frac{pocet(x)}{n},$$

kde výrazem $pocet(x)$ myslíme počet výskytů x v posloupnosti $x_1, x_2, \ldots, x_n$.

Příklad. Při 100 hodech kostkou padlo 17 jedniček, 13 dvojek, 26 trojek, 14 čtyřek, 19 pětek a 11 šestek. Empirické rozdělení pravděpodobnosti jednotlivých hodnot je vidět na obrázku.



Průměr

Definice. *Průměrem* posloupnosti i.i.d. $X_1, X_2, \ldots, X_n$ budeme rozumět

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Poznámka. Z toho, co víme z minula, můžeme odvodit, že pokud jsou $X_1, \ldots, X_n$ realizace X, potom

$$\mathrm{E}\,\overline{X}_n = \mathrm{E}\,\hat{X} = \mathrm{E}\,X,$$
$$\mathrm{Var}\,\overline{X}_n = \frac{1}{n}\,\mathrm{Var}\,X.$$

Průměr je tedy jakýmsi odhadem střední hodnoty.

[10] Slovo „empirie“ znamená praxe, zkušenost. Jedná se tedy o odhad rozdělení X na základě zkušenosti z pokusu.

Obdobně můžeme definovat i odhad rozptylu

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

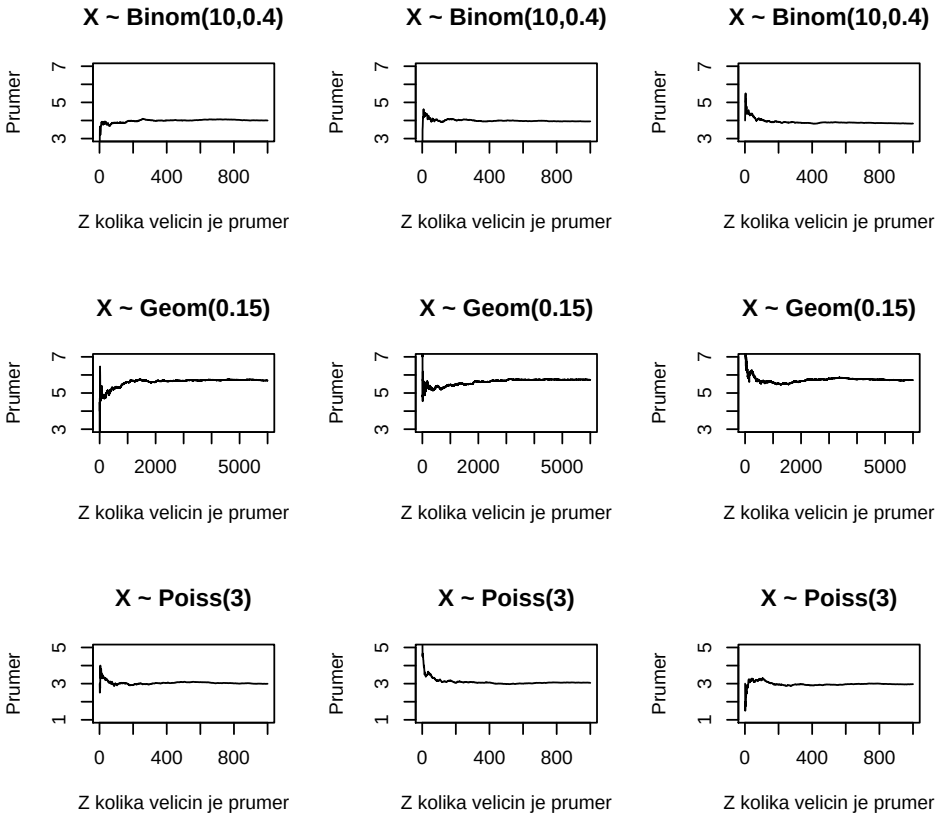
Opět bude platit

$$\mathrm{E}\, S^2 = \mathrm{Var}\, X.$$

Tento odhad zde však nepoužijeme.

Zákon velkých čísel

Zmínili jsme, že průměr je odhadem střední hodnoty. Pojďme se přesvědčit, zda tomu tak skutečně je. Pro náhodné veličiny, jejichž rozdělení jsou znázorněna na prvním obrázku, jsme na počítači nagenerovali velké množství jejich realizací. Z těchto jsme pak vždy z prvních i spočítali průměr $\overline{x}_i = \frac{1}{i}\sum_{j=1}^{i} x_i$ a ten nakreslili do grafu. Pro každou z náhodných veličin jsme tento postup třikrát zopakovali. Výsledky vidíte na obrázcích.



Vidíme, že ačkoli pro malý počet pozorování se průměr od střední hodnoty značně liší, u velkého počtu pozorování je téměř konstantní a jedná se tedy o její dobrý odhad. V statistickém žargonu se tomuto faktu říká zákon velkých čísel (ZVČ):

Věta (ZVČ). *Mějme náhodnou veličinu X s konečnou střední hodnotou $\mathrm{E}\,X$ a i.i.d. posloupnost jejích realizací $X_1, X_2, \ldots$ Potom pro n dostatečně vysoké je $\overline{X}_n$ přibližně $\mathrm{E}\,X$, značíme*

$$\overline{X}_n \approx \mathrm{E}\,X.$$

Správná otázka pochopitelně je, jak to n má být vysoké či jak blízko se k $\mathrm{E}\,X$ pro dané n dostaneme. Tuto otázku řešit nebudeme či lépe řečeno řešit ji obecně nedokážeme. Všimněte si však, že na předchozích obrázcích máme různé měřítko na vodorovné ose, a že se tedy rychlost přibližování průměru k střední hodnotě může pro různá rozdělení lišit.

Příklad. Student řeší test složený z n stejně obtížných otázek. Každou otázku zodpoví správně s pravděpodobností p, kterou neznáme. Jak tuto pravděpodobnost odhadneme? Jednoduše – dáme mu napsat test a za odhad p zvolíme počet správně zodpovězených otázek podělený n. Pokud bude n dostatečně vysoké, získáme dostatečně přesný odhad p.

Přišlo vám to jednoduché, ba až zřejmé? Dobrá, ukažme si něco obtížnějšího.

Centrální limitní věta

Už víme, že pro velká n se průměr blíží střední hodnotě. To však rozhodně neznamená, že by jí byl roven! Co nás tedy může zajímat, je rozdělení odchylek průměru od střední hodnoty či lépe řečeno výraz

$$Y_n = \sqrt{n}\,\frac{\overline{X}_n - \mathrm{E}\,X}{\sqrt{\mathrm{Var}\,X}}.$$

Poznámka. Uvědomme si, že $\mathrm{E}\,X$ a $\mathrm{Var}\,X$ jsou (nenáhodné) konstanty. Tudíž s použitím vět z minula snadno spočítáme střední hodnotu Y_n

$$\mathrm{E}\,Y_n = \sqrt{\frac{n}{\mathrm{Var}\,X}}(\mathrm{E}\,\overline{X}_n - \mathrm{E}\,X) = 0,$$

obdobně rozptyl Y_n je roven

$$\mathrm{Var}\,Y_n = \frac{n}{\mathrm{Var}\,X}\,\mathrm{Var}\,\overline{X}_n = 1.$$

Pro libovolné rozdělení X bude mít tudíž Y_n stejnou střední hodnotu i rozptyl. Za chvíli uvidíme, že pro vysoké n dokonce můžeme říci, že rozdělení Y_n je vždy přibližně stejné, ať bylo původní rozdělení X jakékoli.

Zkusme si pro naše tři staré známé náhodné veličiny počítačem nasimulovat (empirické) rozdělení výše uvedeného výrazu (Y_n). Postupně budeme zkoušet $n = 1$, $n = 10$ a $n = 1\,000$. Pro každý případ nagenerujeme dostatečně mnoho (10 000) realizací posloupnosti $X_1, X_2, \ldots, X_n$, neboli nagenerujeme

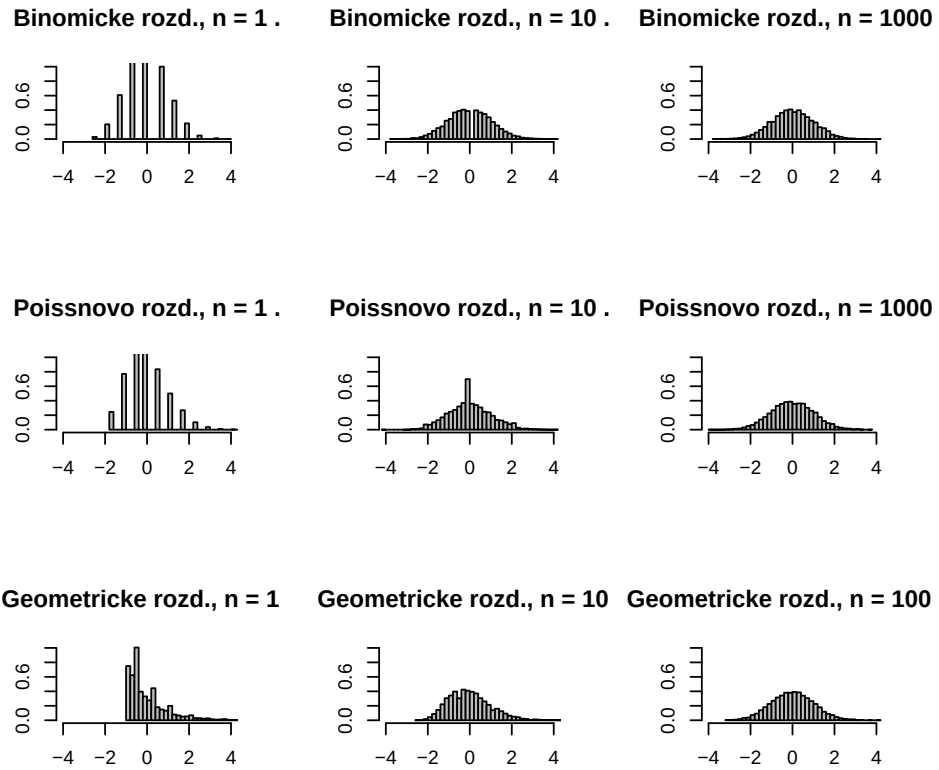
$$X_{i,1}, X_{i,2}, \ldots, X_{i,n} \qquad i = 1, \ldots, 10\,000,$$

a z každé této i-té posloupnosti spočítáme průměr

$$\overline{X}_{i,n} = \frac{1}{n}\sum_{j=1}^{n} X_{i,j}.$$

Tak dostaneme empirické rozdělení průměru $\widehat{\overline{X}}_n$ a z něj i empirické rozdělení výrazu, jenž nás zajímá.

Výsledky (histogramy) můžete vidět na obrázcích:

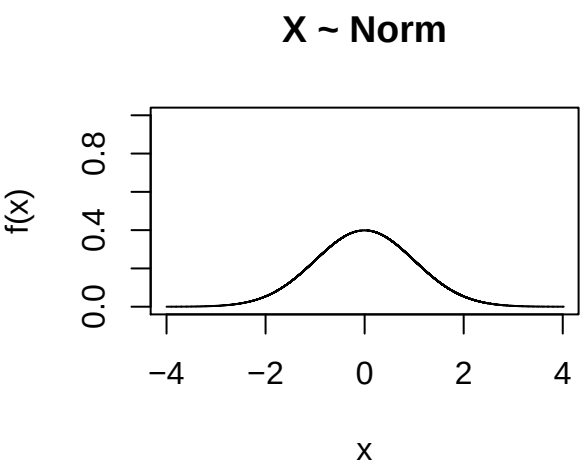


Pokud si výše uvedené obrázky dobře prohlédnete, zjistíte zajímavý fakt, že totiž ačkoli se pro nízké n se jsou obrázky pro různá rozdělení pochopitelně různé, pro n vysoké (zde $n = 1\,000$) jsou si obrázky nápadně podobné.

Kdybychom dále zvyšovali n a počet nagenerovaných hodnot a naopak snižovali velikost intervalu v histogramu, začal by se nám (pro libovolné počáteční rozdělení) obrázek histogramu blížit ke grafu funkce

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}},$$

kde $e \doteq 2{,}718$ je základ základ přirozeného logaritmu. Obrázek (graf) této funkce vidíte dále.



Obecně se tomuto faktu říká v statistické hantýrce centrální limitní věta (CLV):

Věta (CLV). *Mějme náhodnou veličinu X s konečným nenulovým rozptylem a i.i.d. posloupnost jejích realizací $X_1, X_2, \ldots$ Potom pro n dostatečně vysoké je rozdělení náhodné veličiny*

$$\sqrt{n}\,\frac{\overline{X}_n - \mathrm{E}\,X}{\sqrt{\mathrm{Var}\,X}}$$

vždy přibližně[11] *stejné. Toto rozdělení budeme nazývat (normované) normální rozdělení. Píšeme rozdělení* $\sqrt{n}\,\frac{\overline{X}_n - \mathrm{E}\,X}{\sqrt{\mathrm{Var}\,X}} \approx Norm$.

Poznámka. Názorněji je možná význam věty vidět, když ji napíšeme ve tvaru

$$\text{rozdělení } \overline{X}_n \approx \mathrm{E}\,X + \sqrt{\frac{\mathrm{Var}\,X}{n}}Z,$$

kde $Z \sim Norm$. Neboli (přibližné) rozdělení průměru posloupnosti i.i.d. pro dostatečně vysoké n závisí pouze na střední hodnotě a rozptylu X.

Poznámka. Často místo průměru potřebujeme pracovat se sumou, v tomto případě:

$$\sum_{i=1}^{n} X_i = n\overline{X}_n \approx n\,\mathrm{E}\,X + \sqrt{n\,\mathrm{Var}\,X}\;Z.$$

[11] Co toto „přibližně“ znamená by vydalo na samostatný seriál. Jedno z možných zjednodušení říká, že pro libovolně rozdělené X je pravděpodobnost, že Y_n padne do daného intervalu, přibližně tatáž. Neboli že histogramy se příliš neliší.

Náhodnou veličinu $Z \sim Norm$ s normálním rozdělením je za pomoci naší teorie obtížné popsat. Může totiž nabýt libovolné reálné hodnoty, tedy nespočetně mnoha hodnot, každé s nekonečně malou pravděpodobností.

Nelze tedy mluvit o pravděpodobnosti, že Z je rovno z, ale pouze o tom, že $Z \leq z$ nebo že $z_1 \leq Z \leq z_2$. Tyto pravděpodobnosti nelze jednoduše spočítat, lze je však najít v statistických tabulkách nebo vyčíslit za pomoci počítače:

z	$\mathrm{P}(Z \leq z)$
-4	0,0000317
-3	0,0013499
-2	0,0227501
-1	0,1586553
0	0,5000000
1	0,8413447
2	0,9772499
3	0,9986501
4	0,9999683

Často potřebujeme také řešit opačnou úlohu. Nalézt pro dané α číslo z takové, že $\mathrm{P}(Z \leq z) = \alpha$:

α	$z : \mathrm{P}(Z \leq z) = \alpha$
0,5	0
0,75	0,674490
0,9	1,281552
0,95	1,644854
0,975	1,959964
0,99	2,326348
0,995	2,575829

Poznámka. Dlužno podotknout, že normované normální rozdělení je symetrické kolem nuly, neboli platí

$$\mathrm{P}(Z \leq -z) = \mathrm{P}(Z \geq z) = 1 - \mathrm{P}(Z \leq z).$$

Příklad. Stroj na výrobu mikroprocesorů pracuje spolehlivě na 99%. Stroj vyrobil deset tisíc výrobků. Odhadněte počet zmetků. Jaká je pravděpodobnost, že zmetků bude alespoň 120?

Indikátor toho, že je výrobek zmetek, je alternativně rozdělená náhodná veličina (rozdělení $X \sim Alt(0{,}01)$, $\mathrm{E}\,X = 0{,}01$, $\mathrm{Var}\,X = 0{,}0099$). Odhad počtu zmetků v deseti tisících realizacích X je tedy vlastně odhad $10000 \cdot \overline{X}_{10000}$. Veličinu $\overline{X}_{10000}$ odhadneme ze (ZVČ) jakožto 0,01, tedy počet zmetků bude přibližně $10000 \cdot 0{,}01 = 100$.

Pravděpodobnost, že počet zmetků bude alespoň 120, je stejná jako pravděpodobnost, že $\overline{X}_{10000} - \mathrm{E}\,X \geq 0{,}002$. Tato pravděpodobnost bude z (CLV) přibližně rovna

$$\mathrm{P}\left(\sqrt{\frac{\mathrm{Var}\,X}{n}}Z \geq 0{,}002\right) = 1 - \mathrm{P}\left(Z < \sqrt{\frac{n}{\mathrm{Var}\,X}}0{,}002\right) \doteq 1 - \mathrm{P}(Z \leq 2) \doteq 0{,}02275,$$

kde $Z \sim Norm$.

Průměrný počet zmetků tedy bude přibližně 100, pravděpodobnost, že by jich bylo více než 120 je však pouze okolo[12] 2,3%.

Příklad. Obchodník jménem Kolík Ať sepicnu se rozhodl věnovat každému zákazníkovi v jubilejním roce 2000 malého plyšového medvídka. Počet zákazníků Kolíkova krámku má v kterýkoli den v roce Poissonovo rozdělení s parametrem 100. Všichni medvídci musí být objednáni naráz. Kolik jich má Kolík objednat?

Jak správně tušíte není zatím úloha příliš řešitelná, protože když Ať sepicnu objedná z medvídků, je vždy možné (ať už je z vysoké sebevíc), že třeba i v jediný den přijde $z+1$ zákazníků. Je tedy potřeba určit nějaké riziko, které jsme ochotni podstoupit: Řekněme, že jsme ochotni tolerovat pravděpodobnost nedostatku medvídků nanejvýš na úrovni $\alpha = 5\%$.

Snažíme se tedy najít z takové, že

$$\mathrm{P}(\sum_{i=1}^{366} X_i > z) = 0{,}05,$$

neboli

$$\mathrm{P}(\sum_{i=1}^{366} X_i \leq z) = 0{,}95,$$

kde X_1, X_2, ... , X_{366} jsou i.i.d. s rozdělením $Poiss(100)$.

Za použití (CLV) dostáváme

$$\sum_{i=1}^{366} X_i \approx 366\,\mathrm{E}\,X + \sqrt{366\,\mathrm{Var}\,X}\;Z,$$

kde $\mathrm{E}\,X = \mathrm{Var}\,X = 100$, $Z \sim Norm$.

Tedy

$$\mathrm{P}(\sum_{i=1}^{366} X_i \leq z) \doteq \mathrm{P}(366 \cdot 100 + \sqrt{366 \cdot 100}\;Z \leq z) = \mathrm{P}(Z \leq \frac{z - 36600}{\sqrt{36600}}) = 0{,}95.$$

Ovšem z tabulky můžeme vyčíst, že $\mathrm{P}(Z \leq x) = 0{,}95$ právě pro $x = 1{,}644854$. Dosazením

$$\frac{z - 36600}{\sqrt{36600}} = 1.644854$$

a odtud již snadným výpočtem

$$z = 36914{,}7.$$

Kolík Ať sepicnu tedy objedná 36 915 medvídků[13].

Závěr

[12] S pomocí počítače (přesným výpočtem, bez použití CLV) je možné ověřit, že ve skutečnosti je tato pravděpodobnost asi o půl promile nižší, jedná se tedy o vcelku dobrý odhad.

[13] Přesný výpočet na počítači potvrdí, že tentokráte jsme se nespletli ani o jednoho medvídka.

Na závěr tohoto dílu, jakožto i celého seriálu, bych uvedl několik knížek pro ty, kteří by si rádi z tohoto oboru ještě něco přečetli. Řazení je od jednodušších k obtížnějším:

[1] Disman M.: *Jak se vyrábí sociologická znalost*, pohled na pravděpodobnost a statistiku očima sociologa, vcelku zábavné

[2] Anděl J.: *Matematika náhody*, nejkrásnější sbírka úloh z diskrétní pravděpodobnosti, jakou znám

[3] Zvára, K., Štěpán, J.: *Pravděpodobnost a matematická statistika*, pozdější kapitoly předpokládají znalost integrálního počtu

[4] Anděl J.: *Matematická statistika*, statistikova bible, určeno pro náročnější čtenáře