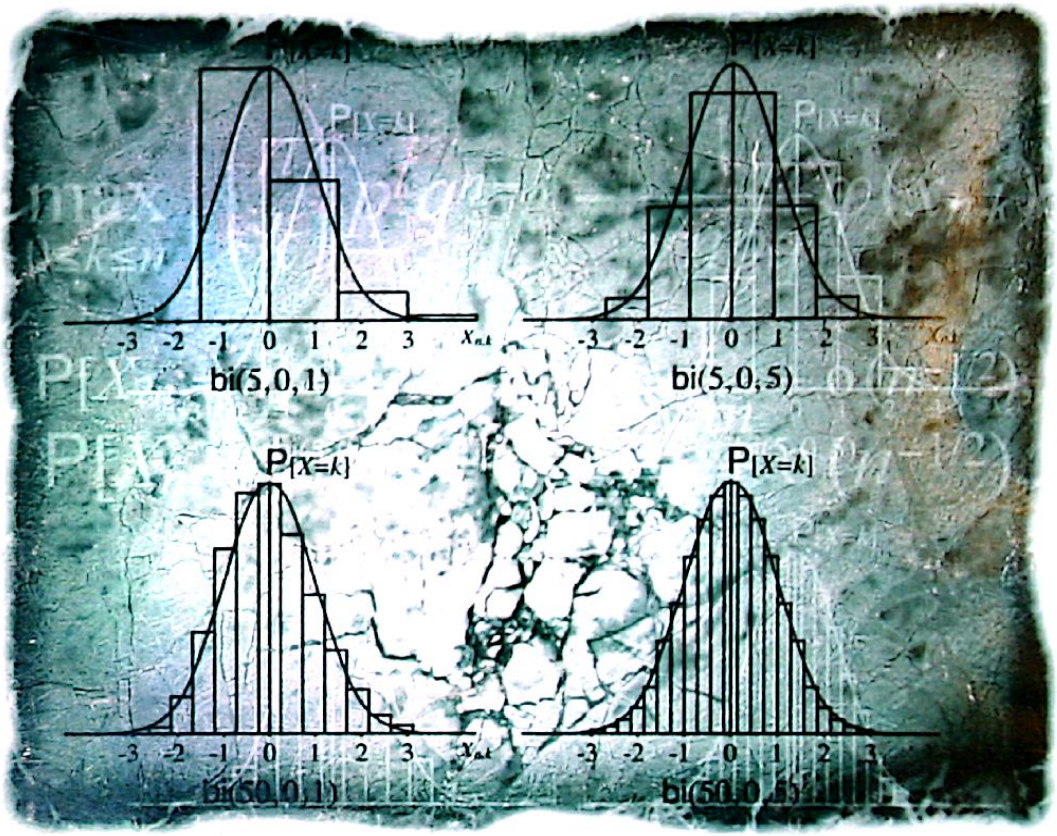
PRAVDĚPODOBNOST A MATEMATICKÁ STATISTIKA



matfyzpress
VYDAVATELSTVÍ
MATEMATICKO-FYZIKÁLNÍ FAKULTY
UNIVERZITY KARLOVY

ISBN 80-85863-24-3

Obsah

Úvodem

Text je určen především studentům učitelství všeobecně vzdělávacích předmětů, kteří na přírodovědných fakultách staroslavné Univerzity Karlovy studují kombinace s matematikou.

Skripta lze použít pro dvousemestrální dvouhodinovou přednášku, kterou lze uspořádat několika způsoby. Pokud je kladen větší důraz na teorii pravděpodobnosti, lze probrat celou první část skript o pravděpodobnosti a výklad statistiky ukončit kapitolou o základech statistické indukce. Chceme-li věnovat více pozornosti statistickým metodám, je možno vynechat výklad některých klasických modelů či asymptotických vlastností. Na teorii lineárních modelů může navázat buď kapitola o speciálních lineárních modelech nebo kapitola o použití multinomického rozdělení v testech dobré shody.

Autoři jsou vděčni Prof. Ing. Václavu Čermákovi, DrSc. a RNDr. Jarmile Zocové za řadu námětů na vylepšení textu a za upozornění na chyby a nedopatření. Stejné poděkování směřuje také k Mgr. Petru Ševčíkovi.

V Praze dne 4. října 1997.

1. Definice pravděpodobnosti

1.1 Klasická pravděpodobnost

O *náhodném pokusu* hovoříme tehdy, když konáme pokus, jehož výsledek není jednoznačně určen podmínkami, za nichž je prováděn. Nechť A je nějaké ověřitelné tvrzení o výsledku náhodného pokusu. Prohlášení *nastal jev* A znamená, že je pravdivé tvrzení A o výsledku náhodného pokusu. Zajímají nás při tom jen takové pokusy, v nichž sledovaný jev vykazuje v opakovaných pokusech jakousi stabilitu: relativní četnost $f_n(A) = n_A/n$ výskytu jevu A v posloupnosti n „nezávislých" pokusů má tendenci při velkých hodnotách n se příliš neměnit, má tendenci držet se nějaké konstanty. Tuto konstantu budeme nazývat *pravděpodobnost* a cílem první části skript je pro pojmy náhodný jev a pravděpodobnost najít vhodný matematický model.

Mějme dva náhodné jevy A a B. Jistě má smysl zjišťovat, zda nastaly oba *současně*, což vede k potřebě definovat **průnik** náhodných jevů $A \cap B$ jako nový náhodný jev. Podobné platí o otázce, zda nastal *aspoň jeden* z jevů A a B, takže **sjednocení** náhodných jevů $A \cup B$ musí být také náhodným jevem. **Jev opačný** A^c nastává, právě když *nenastal* jev A. Odtud plyne, že k systému náhodných jevů musí patřit také **jev jistý**, vzniklý sjednocením $A \cup A^c$. Tento „maximální" jev označíme symbolem Ω. Jev opačný k jevu jistému nazveme **jevem nemožným** a označíme jej symbolem $\emptyset$. Pokud je průnikem náhodných jevů A a B jev nemožný, nazývají se jevy A a B **neslučitelné**. Příkladem dvojice neslučitelných jevů jsou jevy A a A^c. Kromě toho má smysl jevy porovnávat. Pokud má jev A za následek jev B, budeme říkat, že náhodný jev A je **podjevem** náhodného jevu B. Každý z náhodných jevů musí být nepochybně podjevem jistého jevu Ω.

Zřejmě bude vhodné použít množinový jazyk.

Definice 1.1. Systém množin $\mathcal{A}$ se nazývá **algebra**, jestliže platí:

$$A, B \in \mathcal{A} \Rightarrow A \cup B \in \mathcal{A}, \tag{1.1}$$

$$A, B \in \mathcal{A} \Rightarrow A \cap B \in \mathcal{A}, \tag{1.2}$$

$$A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}, \tag{1.3}$$

$$\Omega \in \mathcal{A}, \quad \emptyset \in \mathcal{A}. \tag{1.4}$$

Jako náhodné jevy budeme tedy v našem matematickém modelu chápat množiny z nějaké algebry podmnožin množiny Ω. Obecný prvek ω množiny Ω se v této souvislosti nazývá **elementární jev**. Množinu Ω budeme v konkrétní situaci volit tak, aby elementární jevy ω byly těmi nejjemnějšími výsledky náhodného pokusu, které je ještě třeba rozlišovat.

Nyní se budeme zabývat pravděpodobností, kterou chceme připisovat každému náhodnému jevu. Tato reálná funkce definovaná na algebře $\mathcal{A}$ by měla mít podobné vlastnosti jako relativní četnost, kterou má modelovat. K takovým vlastnostem patří

$$A \in \mathcal{A} \Rightarrow f_n(A) \geq 0, \tag{1.5}$$

$$A, B \in \mathcal{A}, \quad A \cap B = \emptyset \Rightarrow f_n(A \cup B) = f_n(A) + f_n(B), \tag{1.6}$$

$$f_n(\Omega) = 1, \quad f_n(\emptyset) = 0. \tag{1.7}$$

Definice 1.2. Reálnou funkci $\mathsf{P}(A)$ definovanou na algebře $\mathcal{A}$ podmnožin množiny Ω budeme nazývat **pravděpodobnost**, jestliže platí

$$A \in \mathcal{A} \Rightarrow \mathsf{P}(A) \geq 0, \tag{1.8}$$

$$A, B \in \mathcal{A}, \quad A \cap B = \emptyset \Rightarrow \mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B), \tag{1.9}$$

$$\mathsf{P}(\Omega) = 1, \quad \mathsf{P}(\emptyset) = 0. \tag{1.10}$$

Základní vlastnosti pravděpodobnosti prohlubuje následující tvrzení:

Věta 1.1. Je-li P pravděpodobnost definovaná na algebře $\mathcal{A}$, $A, B \in \mathcal{A}$ resp. $A_i \in \mathcal{A}, 1 \leq i \leq n$, pak platí

$$A \subset B \Rightarrow \mathsf{P}(A) \leq \mathsf{P}(B), \tag{1.11}$$

$$A \subset B \Rightarrow \mathsf{P}(B - A) = \mathsf{P}(B) - \mathsf{P}(A), \tag{1.12}$$

$$\mathsf{P}(A^c) = 1 - \mathsf{P}(A), \tag{1.13}$$

$$\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B), \tag{1.14}$$

$$\mathsf{P}\left(\cup_{i=1}^{n} A_i\right) = \sum_{1 \leq i \leq n} \mathsf{P}(A_i) - \sum_{1 \leq i < j \leq n} \mathsf{P}(A_i \cap A_j) + \sum_{1 \leq i < j < k \leq n} \mathsf{P}(A_i \cap A_j \cap A_k) + \ldots + (-1)^{n+1} \mathsf{P}\left(\cap_{i=1}^{n} A_i\right). \tag{1.15}$$

D ů k a z: Výše uvedené implikace jistě umíte dokázat sami, pokud nikoliv, použijte se v důkazu věty 1.2. □

Užitečný speciální případ dostaneme, když budeme předpokládat:

1. Jistý jev Ω sestává pouze z konečného počtu elementárních jevů.
2. Není důvod považovat některý z elementárních jevů $\omega \in \Omega$ za více možný („pravděpodobnější“) než ostatní elementární jevy.

Na těchto předpokladech je založena následující definice pravděpodobnosti používaná ve středoškolské matematice.

Definice 1.3. (Klasická definice pravděpodobnosti) Nechť Ω je konečná množina, nechť $\mathcal{A}$ je algebra všech podmnožin množiny Ω. Pravděpodobnost je dána vztahem

$$\mathsf{P}(A) = \frac{|A|}{|\Omega|}.$$

Trojice $(\Omega, \mathcal{A}, \mathsf{P})$ se nazývá **klasický pravděpodobnostní prostor**.

Snadno zjistíme, že klasická definice opravdu zavádí na konečné množině Ω pravděpodobnost. Vlastnost (1.8) je přímým důsledkem toho, že počet prvků $|A|$ množiny A, která je podmnožinou konečné množiny Ω, musí být nezáporný. Pro disjunktní množiny A a B platí dále $|A \cup B| = |A| + |B|$, takže je splněn také požadavek (1.9). Podobně jsou zřejmě splněny požadavky (1.10).

Příklad 1.1. Uvažujme n hodů symetrickou mincí. Strana, na níž je vyobrazen státní znak, se zpravidla nazývá líc mince. Zajímá nás náhodný jev A_k = [padlo právě k líců], $0 \leq k \leq n$. Máme dvě možnosti volby množiny elementárních jevů. Protože sledovaný počet líců musí být v rozmezí od nuly do n, stačilo by zvolit $\Omega = \{0, 1, \ldots, n\}$. Náhodné jevy A_k jsou pak totožné s elementárními jevy. Jinou možností je chápat výsledek pokusu jako uspořádanou n-tici nul a jedniček. Prostor elementárních jevů má pak tvar $\Omega_n = \{0,1\} \times \{0,1\} \times \ldots \times \{0,1\}$. Každý z jevů A_k v tomto případě musíme sestavit vždy z těch elementárních jevů, které obsahují právě k jedniček. Klasická definice pravděpodobnosti je zřejmě použitelná pouze ve druhém případě. ○

Schopnost umět dobře počítat klasické (kombinatorické) pravděpodobnosti je užitečný pomocník při studiu pokročilejších partií pravděpodobnosti a statistiky. Uvědomme si, že řešení následujících příkladů není pouze otázkou znalosti kombinatorických postupů a pravidel. Prvním krokem je vždy vytvoření pravděpodobnostního modelu, tj. stanovení množiny elementárních jevů Ω tak, aby nebylo pochyb o oprávněnosti nahoře uvedeného výroku 2. V jednotlivých příkladech i ve výkladu budeme pro některé kombinatorické pojmy používat označení zavedená v dodatku A1

Příklad 1.2. Předmětem výběrových šetření (viz též oddíl 11.1) je zjištění určité vlastnosti velké populace tzv. statistických jednotek $S = \{1, 2, \ldots, N\}$ na základě vyčerpávajícího průzkumu malého vzorku (výběru) $s = \{i_1, \ldots, i_n\} \subset S$. V zájmu reprezentativnosti vzorku je třeba pořídit výběr tak, aby každá jednotka měla stejnou možnost být zahrnuta do výběru. Jedna z možností, jak toho docílit, spočívá v tom, že každý výběr $s \subset S$,

který má n prvků ($s \in \exp_n S$), bude volen se stejnou pravděpodobností $\binom{N}{n}^{-1}$. Uvažujme tedy klasický pravděpodobnostní prostor, kde množinou elementárních jevů je množina všech podmnožin mohutnosti n uvnitř S. Uvažujme jednotky $1 \le i \ne k \le N$ a vypočtěme pravděpodobnosti náhodných jevů:

V_i = [jednotka i je zahrnuta do výběru] = $\{s \in \exp_n S : i \in s\}$,
$V_{i,k}$ = [jednotky i a k jsou zahrnuty do výběru],
V_{i+k} = [alespoň jedna z jednotek i a k je zahrnuta do výběru],
V_{i-k} = [jednotka i je zahrnuta, jednotka k nikoliv],
$V_{i\nabla k}$ = [právě jedna z jednotek i a k je zahrnuta do výběru].

Zřejmě je

$$\mathsf{P}(V_i) = \binom{N-1}{n-1}\binom{N}{n}^{-1} = \frac{n}{N},$$

protože výběr zahrnující jednotku i je jednoznačně určen podmnožinou $s \in \exp_{n-1}(S - \{i\})$. Podobně vypočteme

$$\mathsf{P}(V_{i,k}) = \mathsf{P}(V_i \cap V_k) = \binom{N-2}{n-2}\binom{N}{n}^{-1} = \frac{n(n-1)}{N(N-1)}.$$

Povšimněme si, že platí

$$\mathsf{P}(V_i \cap V_k) \ne \mathsf{P}(V_i) \cdot \mathsf{P}(V_k),$$

tj. zahrnutí dvou různých jednotek nejsou „nezávislé události".

Při výpočtu zbývajících pravděpodobností použijeme vlastnosti uvedené ve větě 1.1:

$$\mathsf{P}(V_{i+k}) = \mathsf{P}(V_i \cup V_k) = \mathsf{P}(V_i) + \mathsf{P}(V_k) - \mathsf{P}(V_{i,k}) = \frac{n}{N}\left(2 - \frac{n-1}{N-1}\right),$$

$$\mathsf{P}(V_{i-k}) = \mathsf{P}(V_i - V_{i,k}) = \frac{n}{N}\left(1 - \frac{n-1}{N-1}\right),$$

$$\mathsf{P}(V_{i\nabla k}) = \mathsf{P}(V_i \nabla V_k) = \mathsf{P}(V_{i-k}) + \mathsf{P}(V_{k-i}) = \frac{2n}{N}\left(1 - \frac{n-1}{N-1}\right). \qquad \bigcirc$$

Příklad 1.3. K roztržité šatnářce přichází n hostů restaurace odložit si svůj kabát. Šatnářka vydává čísla označující příslušný věšák zcela chaoticky. Náhodný svědek jejího počínání si může položit otázku, jaká je vlastně naděje (pravděpodobnost), že alespoň jeden z hostů dostane při odchodu svůj vlastní kabát.

Nejpodrobnějším zápisem výsledku tohoto „experimentu" je posloupnost $(k_1, \ldots, k_n)$, kde $1 \le k_i \le n$ je číslo přidělené hostu, který přišel jako i-tý, tedy jistá permutace množiny $\{1, 2, \ldots, n\}$.

Použijeme-li klasický pravděpodobnostní prostor s množinou elementárních jevů P_n (všechny permutace n-té třídy), modelujeme výrok „vydává zcela chaoticky" tak, že každá permutace vydávaných čísel je stejně pravděpodobná. Máme vypočítat pravděpodobnost náhodného jevu

$$\begin{aligned} A^{(n)} &= \{(k_1, \ldots, k_n) \in P_n : \text{existuje } 1 \le i \le n \text{ s vlastností } k_i = i\} \\ &= A_1 \cup A_2 \cup \cdots \cup A_n, \end{aligned}$$

kde

$$A_i = [i\text{-tý host obdržel číslo } i] = \{(k_1, \ldots, k_n) \in P_n : k_i = i\}.$$

Podle vzorce (1.15) je

$$\mathsf{P}(A^{(n)}) = 1 - \binom{n}{2}\frac{(n-2)!}{n!} + \binom{n}{3}\frac{(n-3)!}{n!} - \cdots + (-1)^{n+1}\frac{1}{n!},$$

protože

$$\mathsf{P}(A_i) = \frac{(n-1)!}{n!} \quad \text{pro} \quad 1 \le i \le n,$$

$$\mathsf{P}(A_i \cap A_j) = \frac{(n-2)!}{n!} \quad \text{pro} \quad 1 \le i \ne j \le n, \ldots$$

a

$$\mathsf{P}\left(\bigcap_{j=1}^{n} A_j\right) = \frac{1}{n!}.$$

Po úpravě dostaneme

$$\mathsf{P}(A^{(n)}) = 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n+1}\frac{1}{n!}.$$

Tabulka 1.1 udává několik hodnot $\mathsf{P}(A^{(n)})$.

n	1	2	3	4	5	6	7
$\mathsf{P}\left(A^{(n)}\right)$	1,0000	0,5000	0,6667	0,6250	0,6333	0,6319	0,6321

Tabulka 1.1: Pravděpodobnosti správného vydání kabátu

Dalším počítáním bychom zjistili, že velikost pravděpodobnosti $\mathsf{P}(A^{(n)})$ se již příliš nemění, což není žádné překvapení. Napíšeme-li Taylorův rozvoj pro funkci e^x v bodě $x = -1$, vidíme, že platí:

$$\lim_{n\to\infty} \mathsf{P}(A^{(n)}) = \lim_{n\to\infty} \left(1 - \sum_{j=0}^{n} \frac{(-1)^j}{j!}\right) = 1 - e^{-1} \doteq 0{,}6321. \quad \bigcirc$$

Příklad 1.4. Mnohokrát se každému stalo, že si znovu vyslechl anekdotu, kterou již někomu vyprávěl. Nalezneme jednoduchý model šíření anekdoty a vypočteme pravděpodobnost jejího návratu k osobě, která již anekdotu zná.

Uvažujme okruh osob $S = \{0, 1, \ldots, N\}$, mezi kterými se anekdota n krát od jedince k jedinci předává, přičemž osoba 0 je jejím autorem. Trajektorie šíření anekdoty $(0, k_1, k_2, \ldots, k_n)$ je posloupnost prvků množiny S délky $n+1$, která má na prvním místě prvek 0 a žádné dva po sobě jdoucí prvky nejsou totožné. Uvažme klasický pravděpodobnostní prostor, kde množina elementárních jevů Ω je složena právě z těchto posloupností. Matematickou indukcí podle n se můžeme snadno přesvědčit,, že $|\Omega| = N^n$, tj. že množina posloupností Ω je stejně mohutná jako krychle $K = \{1, 2, \ldots, N\}^n$. (Pokuste se zkonstruovat vzájemně jednoznačné zobrazení mezi K a Ω.)

Předpokládejme, že $n \leq N$. Snadno vypočteme pravděpodobnost náhodného jevu C, který spočívá v tom, že šíření anekdoty má cyklus. Doplněk náhodného jevu C tvoří právě ty posloupnosti v Ω, které mají různé prvky, tj.

$$\begin{aligned} \mathsf{P}(C) &= 1 - \frac{N(N-1)\ldots(N-n+1)}{N^n} \\ &= 1 - \prod_{j=1}^{n-1} \left(1 - \frac{j}{N}\right) = 1 - p_{n,N} = d_{n,N}. \end{aligned}$$

Je-li velikost výběru n malá v porovnání s velikostí populace N, můžeme si udělat hrubou představu o velikosti pravděpodobnosti $p_{n,N}$. Ve vztahu $\ln p_{n,N} = \sum_{j=1}^{n-1} \ln(1 - j/N)$ můžeme bez větší ztráty přesnosti použít známou aproximaci $\ln(1+t) \doteq t$ (použij Taylorův rozvoj funkce $\ln(1+t)$ v bodě $t = 0$, přesnost navržené aproximace plyne z $|\ln(1+t) - t| \leq t^2$ pro $|t| \leq \frac{1}{2}$). Odtud obdržíme přibližné vzorce

$$\ln p_{n,N} \doteq -\frac{1}{N} \sum_{j=1}^{n-1} j = -\frac{n(n-1)}{2N} \doteq -\frac{n^2}{2N}$$

a tedy

$$p_{n,N} \doteq \exp\left(-\frac{n^2}{2N}\right).$$

Tabulka 1.2 uvádí takto přibližně určené hodnoty pravděpodobností „zacyklení anekdoty“ $d_{n,N}$ při pevných hodnotách podílu $n/N = 0{,}1$ a $n/N = 0{,}05$: Podobným způsobem vypočteme pravděpodobnost $q_{n,N}$ toho, že se

$n/N = 0{,}1$				$n/N = 0{,}05$			
n	$d_{n,N}$	n	$d_{n,N}$	n	$d_{n,N}$	n	$d_{n,N}$
1	0,0488	14	0,5034	1	0,0247	27	0,4908
3	0,1393	15	0,5276	5	0,1175	28	0,5034
5	0,2212	20	0,6321	10	0,2212	29	0,5157
10	0,3935	25	0,7135	15	0,3127	30	0,5276
11	0,4231	30	0,7769	20	0,3935	35	0,5831
12	0,4512	35	0,8262	25	0,4647	40	0,6321
13	0,4780	40	0,8647	26	0,4780	45	0,6753

Tabulka 1.2: Závislost pravděpodobnosti zacyklení anekdoty na n a N.

anekdota někdy vrátí ke svému tvůrci. Doplňkový náhodný jev je sestaven ze všech posloupností v množině Ω, které používají prvek 0 pouze na prvém místě. Je tedy

$$\begin{aligned} q_{n,N} &= 1 - \frac{N(N-1)^{n-1}}{N^n} \\ &= 1 - \left(1 - \frac{1}{N}\right)^{n-1} \\ &= 1 - \left(1 - \frac{a}{n}\right)^{n-1}, \end{aligned}$$

kde $a = n/N$.

a	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
$q_{n,N}$	0,095	0,181	0,259	0,329	0,393	0,451	0,503	0,550	0,593

Tabulka 1.3: Aproximace pravděpodobnosti $q_{n,N}$ návratu anekdoty k jejímu tvůrci.

Je známo, že limita výrazu $(1 - a/n)^{n-1}$ při $n \to \infty$ je rovna číslu e^{-a}. Přicházíme k závěru, že rostou-li čísla n a N nade všechny meze tak, že podíl n/N zůstává konstantně roven číslu a, pak

$$q_{n,N} \doteq 1 - e^{-a} = 1 - (0{,}3679)^a.$$

Tabulka 1.3 uvádí takto určené pravděpodobnosti $q_{n,N}$ návratu anekdoty k jejímu tvůrci pro různé hodnoty podílu $a = n/N$. ○

1.2 Náhodná veličina

Velmi často se stává, že jsme schopni, nebo považujeme za užitečné, popisovat náhodný jev $\omega \in \Omega$ pouze částečně, prostřednictvím některé jeho číselné charakteristiky $X(\omega)$ (rozměr, hmotnost,...), kterou nazveme **náhodná veličina**. Přesněji a také stručněji: Náhodná veličina na klasickém pravděpodobnostním prostoru s množinou elementárních jevů Ω je libovolná reálná funkce X definovaná na Ω. Nás budou samozřejmě zajímat pravděpodobnosti, se kterými náhodná veličina X nabývá svých jednotlivých hodnot $x, y, \ldots \in (-\infty, \infty)$, tj. $\mathsf{P}[X = x] = \mathsf{P}(\{\omega \in \Omega : X(\omega) = x\})$ nebo pravděpodobnosti, se kterými nalezneme hodnoty náhodné veličiny X v některé podmnožině reálné přímky, například $\mathsf{P}[X < x] = \mathsf{P}(\{\omega \in \Omega : X(\omega) < x\})$. Poznamenejme, že funkce $\mathsf{P}[X = x]$ proměnné $x \in (-\infty, \infty)$ se nazývá **rozdělení náhodné veličiny** X.

Částečnou představu o „pravděpodobnostní velikosti" náhodné veličiny X si můžeme udělat tak, že spočítáme vážený průměr jejích hodnot, tj. číslo

$$\mathsf{E}X = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} X(\omega) = \sum_{x} x\, \mathsf{P}[X = x], \tag{1.16}$$

kde poslední sčítání je prováděno přes všechny možné hodnoty náhodné veličiny X. Číslo $\mathsf{E}X$ se nazývá **střední hodnota** náhodné veličiny X. Snadno se nahlédne, že se střední hodnota mění se změnou polohy a měřítka podle vzorce $\mathsf{E}(a + bX) = a + b\mathsf{E}X$, kde a, b jsou libovolné konstanty.

Pojmy a značení, které jsme právě zavedli, použijeme v následujících dvou příkladech.

Příklad 1.5. Házíme postupně šesti kostkami a obdržíme posloupnost $(k_1, k_2, \ldots, k_6)$ dosažených bodů. Jak velká je pravděpodobnost, že takto obdržíme monotonní posloupnost? Jaké je rozdělení pravděpodobností náhodné veličiny K, která označuje maximální počet bodů v sérii $(k_1, k_2, \ldots, k_6)$?

Úlohu budeme samozřejmě modelovat klasickým pravděpodobnostním prostorem s množinou elementárních jevů $\Omega = (6)^6$ všech posloupností délky 6 množiny $(6) = \{1, 2 \ldots, 6\}$. Označíme-li množinu všech monotonních posloupností M, je zřejmě $M = M_r \cup M_k$, kde M_r a M_k jsou neklesající, resp. nerostoucí posloupnosti v M. Podle (A3) platí

$$\mathsf{P}(M_r) = \mathsf{P}(M_k) = 6^{-6} \binom{6 - 1 + 6}{6} = 6^{-6} \binom{11}{6}.$$

Pravděpodobnost náhodného jevu M vypočteme nyní pomocí pravidla (1.14) z věty 1.1:

$$\mathsf{P}(M) = 2 \cdot 6^{-6} \binom{11}{6} - 6^{-6} \cdot 6 = 6^{-6} \cdot 918 = 0{,}0197,$$

protože náhodný jev $M_k \cup M_r$ zahrnuje právě šest konstantních posloupností $(k_1, k_2, \ldots, k_6)$. Vidíme, že náhoda dává dobře uspořádaným výsledkům jen malou naději.

Náhodná veličina K je funkce definovaná na množině $(6)^6$ formálně následujícím předpisem

$$K(k_1, k_2, \ldots, k_6) = \max(k_1, k_2, \ldots, k_6) \quad \text{pro} \quad (k_1, k_2, \ldots, k_6) \in (6)^6.$$

Máme vypočítat pravděpodobnosti náhodných jevů $[K = j]$ pro $1 \le j \le 6$. Snadněji určíme pravděpodobnosti náhodných jevů $[K \le j]$, které popisují ten výsledek experimentu, při kterém jsou všechny dosažené body nejvýše rovny číslu $j = 1, 2, \ldots, 6$. Je tudíž

$$\mathsf{P}[K \le j] = \frac{j^6}{6^6} \quad \text{a} \quad \mathsf{P}[K = 1] = \frac{1}{6^6},$$

$$\mathsf{P}[K = j] = \mathsf{P}[K \le j] - \mathsf{P}[k \le j - 1] = \left(\frac{j}{6}\right)^6 - \left(\frac{j-1}{6}\right)^6$$

pro $j = 1, 2, \ldots, 6$, a to podle (1.14) z věty 1.1 o vlastnostech pravděpodobnosti. Tabulka 1.4 udává rozdělení pravděpodobností náhodné veličiny K. Je možná zajímavé si všimnout, že střední hodnota maximálního počtu bodů K je rovna

$$\mathsf{E}K = \sum_{j=1}^{6} j\mathsf{P}[K = j] \doteq 5{,}56029,$$

kdežto nejpravděpodobnější hodnotou je číslo 6. ○

j	1	2	3	4	5	6
$\mathsf{P}[k = j]$	0,00002	0,00135	0,01425	0,07217	0,24711	0,66510

Tabulka 1.4: Rozdělení pravděpodobností maximálního počtu bodů v sérii šesti hodů hrací kostkou

Příklad 1.6. Významný polský matematik Stefan Banach (1892–1945) byl kuřák a měl vždy pro jistotu k disposici dvě krabičky zápalek, které při

použití náhodně střídal. Při jedné příležitosti, když poprvé v jedné z krabiček již nebyla žádná zápalka, si Banach položil otázku, jaké rozdělení pravděpodobností má náhodná veličina K_n vyjadřující počet zápalek ve zbývající krabičce.

Předpokládejme, že v každé krabičce je právě n zápalek. Nejpodrobnější záznam jejich postupného vyprazdňování je jistě dán posloupností nul a jedniček délky $2n+1$, kde 0, resp. 1 na k-tém místě posloupnosti znamená, že byla použita nultá, resp. první krabička (délka posloupnosti je stanovena tak, aby bylo možno vypočítat pravděpodobnosti všech možných obsahů $K_n = k$ zbývající krabičky pro $0 \le k \le n$). Jsme tedy v situaci příkladu 1.1, kde prostorem elementárních jevů je množina $\{0,1\}^{2n+1}$. Máme počítat pravděpodobnosti náhodných jevů

$$\begin{aligned}[K_n = k] &= [\text{při prvém objevení prázdné krabičky je ve zbývající krabičce právě } k \text{ zápalek}] \\ &= A_k^0 \cup A_k^1 \quad \text{pro} \quad 0 \le k \le n,\end{aligned}$$

kde

$A_k^i = [K_n = k] \cap [\text{právě } i\text{-tá krabička byla objevena jako prázdná}]$, $i = 0, 1$.

Podívejme se podrobněji na posloupnost $(j_1, \ldots, j_{2n+1}) \in A_k^1$. Významná je souřadnice s indexem $(n+1)+(n-k) = m$, který vyjadřuje, při kterém pokusu o zapálení cigarety nastal náhodný jev A_k^1. Naše posloupnost musí být taková, že v úseku $j_1, \ldots, j_{m-1}$ je právě n jedniček, $j_m = 1$ a konečně úsek $j_{m+1}, \ldots, j_{2n+1}$ je libovolná posloupnost v množině $\{0,1\}^k$. Je tedy

$$\mathsf{P}(A_k^1) = 2^{-(2n+1)} \binom{2n-k}{n} \cdot 1 \cdot 2^k$$

a protože A_k^1 a A_k^0 jsou dvě stejně mohutné disjunktní množiny, dostáváme konečný výsledek

$$\mathsf{P}[K_n = k] = 2\mathsf{P}(A_n^1) = \binom{2n-k}{n} \cdot 2^{k-2n} \quad \text{pro} \quad 0 \le k \le n.$$

Učiníme si opět představu o velikosti právě vypočtených pravděpodobností. Zvolíme $n = 10$ a vypočteme

$$p_k = \mathsf{P}[K_{10} = k] = \binom{20-k}{10} \cdot 2^{k-20} \quad \text{pro} \quad k = 0, 1, \ldots, 10$$

a dostáváme hodnoty uvedené v tabulce 1.5 a).

k	$\binom{20-k}{10}$	p_k	kp_k
0	184 756	0,1762	0,0000
1	92 378	0,1762	0,1762
2	43 758	0,1669	0,3338
3	19 448	0,1484	0,4451
4	8 008	0,1222	0,4888
5	3 003	0,0916	0,4582
6	1 001	0,0611	0,3666
7	286	0,0349	0,2444
8	66	0,0161	0,1289
9	11	0,0054	0,0483
10	1	0,0010	0,0098
součet		1,0000	2,7001

a) Pravděpodobnosti, že při prvním objevení se prázdné krabičky je ve zbývající krabičce k zápalek

n	$\binom{2n}{n}2^{-2n}$	$\frac{1}{\sqrt{\pi n}}$
1	0,5000	0,5642
5	0,2461	0,2523
10	0,1762	0,1784
15	0,1445	0,1457
20	0,1254	0,1262
25	0,1123	0,1128
30	0,1026	0,1030
31	0,1009	0,1012
32	0,0993	0,0997
33	0,0978	0,0982
34	0,0964	0,0968

b) Porovnání aproximace a přesné hodnoty výrazu $1/\sqrt{\pi n}$

Tabulka 1.5:

Zjistili jsme, že nejpravděpodobnější hodnoty veličiny K_{10} jsou nula a jedna. Vypočteme-li

$$\mathsf{P}[K_{10} > 6] = \mathsf{P}[K_{10} = 7] + \mathsf{P}[K_{10} = 8] + \mathsf{P}[K_{10} = 9] + \mathsf{P}[K_{10} = 10] = 0{,}0574,$$

vidíme, že při vyprázdnění jedné z krabiček bude druhá krabička obsahovat více než 6 zápalek jen s mizivou pravděpodobností. Součet čísel v posledním sloupečku (2,7001) je střední hodnota náhodné veličiny K_{10}. Při vyprázdnění jedné z krabiček je tedy střední (očekávaný) počet zápalek v krabičce druhé roven přibližně číslu 3.

Pro obecné n lze střední hodnotu náhodné veličiny K_n, tj. číslo $\mathsf{E}K_n$, vypočítat přesně. Je

$$\begin{aligned} n - \mathsf{E}K_n &= \mathsf{E}(n - K_n) = \sum_{k=0}^{n} (n-k) \binom{2n-k}{n} 2^{-(2n-k)} \\ &= \sum_{j=0}^{n} j \binom{n+j}{j} 2^{-(n+j)} = \frac{1}{2} \sum_{j=1}^{n} (n+j) \binom{n+j-1}{j-1} 2^{-(n+j-1)} \\ &= \frac{1}{2} \sum_{j=0}^{n} (n+j+1) \binom{n+j}{j} 2^{-(n+j)} - \frac{1}{2}(2n+1) \binom{2n}{n} 2^{-2n} \end{aligned}$$

$$= \frac{1}{2}\mathsf{E}\,(n + (n - K_n) + 1) - \frac{1}{2}(2n+1)\binom{2n}{n}2^{-2n}$$
$$= n - \frac{1}{2}\mathsf{E}K_n + \frac{1}{2} - \frac{1}{2}(2n+1)\binom{2n}{n}2^{-2n}$$

Po výpočtu dostaneme

$$\mathsf{E}K_n = (2n+1)\binom{2n}{n}2^{-2n} - 1.$$

Pro $n = 10$ vyjde $\mathsf{E}K_{10} = 2{,}7001$.

Při skutečně velkých n je obtížné stanovit číslo $\binom{2n}{n}2^{-2n}$, které se ve vzorci vyskytuje. Pomůže Stirlingův vzorec

$$n! \doteq n^n\sqrt{2\pi n}\mathrm{e}^{-n}$$

(limita při $n \to \infty$ podílu levé a pravé strany je rovna jedné). Prostým výpočtem dostaneme odhad

$$\binom{2n}{n}2^{-2n} \doteq \frac{1}{\sqrt{\pi n}}.$$

Přesnost, resp. nepřesnost této aproximace vidíme z tabulky 1.5 b).

Pro střední hodnotu náhodné veličiny K_n dostáváme tedy odhady

$$\mathsf{E}K_n \doteq \frac{2n+1}{\sqrt{\pi n}} - 1 \doteq 1{,}128\sqrt{n} - 1,$$

tj. $\mathsf{E}K_{10} \doteq 2{,}7467$, resp. 2,5670 (přesná hodnota je 2,7001) a $\mathsf{E}K_{50} \doteq 7{,}0586$, resp. 6,9762, přesně 7,0385.

Podobně se pokusíme odhadnout pravděpodobnost $\mathsf{P}[K_n = k]$. Platí

$$\mathsf{P}[K_n = k] = \binom{2n-k}{n}2^{-2n}2^k$$
$$= \binom{2n}{n}2^{-2n}2^k\frac{n(n-1)\dots(n-(k-1))}{2n(2n-1)\dots(2n-(k-1))}$$
$$= \binom{2n}{n}2^{-2n}2^k2^{-k}\frac{(1-\frac{1}{n})(1-\frac{2}{n})\dots(1-\frac{k-1}{n})}{(1-\frac{1}{2n})(1-\frac{2}{2n})\dots(1-\frac{k-1}{2n})}$$
$$\doteq \frac{1}{\sqrt{\pi n}}D_{n,k},$$

kde $D_{n,k}$ je podíl součinů v předposledním výrazu. O velikosti $D_{n,k}$ si můžeme učinit představu tak, že nejdříve vypočteme jeho přirozený logaritmus

$$\ln D_{n,k} = \sum_{j=1}^{k-1}\ln\left(1-\frac{j}{n}\right) - \sum_{j=1}^{k-1}\ln\left(1-\frac{j}{2n}\right).$$

Použijme nyní aproximaci $\ln(1+t) \doteq t$ jako v příkladu 1.4. Předpokládejme, že číslo k je malé ve srovnání s číslem n. Pak jsou malá také čísla $t = -j/n$ vyskytující se ve výrazu pro $\ln D_{n,k}$ a můžeme psát:

$$\ln D_{n,k} \doteq \sum_{j=1}^{k-1}\left(-\frac{j}{n}\right) - \sum_{j=1}^{k-1}\left(-\frac{j}{2n}\right) = -\frac{1}{2n}\sum_{j=1}^{k-1}j = -\frac{k(k-1)}{4n} \doteq -\frac{k^2}{4n}.$$

Dostáváme tedy aproximaci $D_{n,k} \doteq \mathrm{e}^{-\frac{k^2}{4n}}$ a konečně

$$\mathsf{P}[K_n = k] \doteq \frac{1}{\sqrt{\pi n}}\mathrm{e}^{-\frac{k^2}{4n}} \quad (k \ll n).$$

Vyzkoušíme tuto aproximaci pro $n = 10$ (porovnejte s přesnými hodnotami p_k v tabulce 1.5):

k	$\frac{1}{\sqrt{10\pi}}\mathrm{e}^{-\frac{k^2}{40}}$	k	$\frac{1}{\sqrt{10\pi}}\mathrm{e}^{-\frac{k^2}{40}}$
0	0,1784	6	0,0725
1	0,1740	7	0,0524
2	0,1614	8	0,0360
3	0,1425	9	0,0235
4	0,1196	10	0,0146
5	0,0955		

k	$\frac{1}{\sqrt{50\pi}}\mathrm{e}^{-\frac{k^2}{200}}$	k	$\frac{1}{\sqrt{50\pi}}\mathrm{e}^{-\frac{k^2}{200}}$
0	0,07979	25	0,003506
1	0,07939	30	0,00008864
5	0,07041	35	0,00001745
10	0,04839	40	0,00002677
15	0,02590	45	0,000003197
20	0,01080	50	0,000000297

Tabulka 1.6: Aproximace pro $n = 10$ (vlevo) a pro $n = 50$ (vpravo)

Pro $n = 50$, kdy již nejsme běžnými prostředky (kalkulačka) schopni pravděpodobnost $\mathsf{P}[K_{50} = k]$ vyčíslit přesně, jsou její přibližné hodnoty uvedeny také v tabulce 1.6. ○

1.3 Rozšíření klasické definice pravděpodobnosti

Jaké jsou nevýhody klasické definice pravděpodobnosti? Především druhá podmínka silně připomíná definici kruhem: pravděpodobnost definujeme za předpokladu, že jakési náhodné jevy jsou „stejně pravděpodobné“. Navíc, jak

jsme viděli v příkladu 1.1, ne vždy jsou elementární jevy takto symetrické. Stačí si představit krabičku od zápalek dodatečně opatřenou jedním až šesti puntíky, jaké má hrací kostka. Zůstaneme-li u předpokladu konečné množiny Ω, můžeme pak pravděpodobnost zavést tak, že i-tému elementárnímu jevu ω_i přiřadíme nezáporné číslo p_i tak, že je $\sum_{i=1}^{m} p_i = 1$, kde $m = |\Omega|$. Pravděpodobnost náhodného jevu A, kde $A \subset \Omega$, je pak dána předpisem

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

Uvedený postup je konzistentní s klasickou definicí pravděpodobnosti, pokud zvolíme $p_i = 1/m$, $1 \le i \le m$.

Náš model lze poněkud rozšířit, když budeme předpokládat, že množina Ω je spočetná. Pak můžeme její prvky (elementární jevy) očíslovat a přiřadit jim nezáporné hodnoty p_i splňující požadavek

$$\sum_{i=1}^{\infty} p_i = 1. \tag{1.17}$$

Protože je v (1.17) řada s nezápornými členy, znamená konvergence předpokládaná v (1.17) konvergenci *absolutní*, takže přerovnáním (přečíslováním elementárních jevů) se součet řady nezmění. Nic nás nenutí požadovat, aby množina A, jejíž pravděpodobnost počítáme, byla nutně konečná, byť i tentokrát musí být $A \in \mathcal{A}$. Pak ovšem náhodný jev A nemusí být *konečným* sjednocením elementárních jevů, což nás může vést k nutnosti rozšířit požadavek (1.1) také na spočetnou posloupnost množin.

Příklad 1.7. Opakujme hody symetrickou mincí, dokud nepadne první líc. Výsledkem pokusu je počet hodů nutných k dosažení prvního líce. Protože musíme (aspoň teoreticky) připustit posloupnost $(0, 0, \ldots)$, tedy situaci, že se líce vůbec nedočkáme, jako prostor elementárních jevů můžeme použít

$$\Omega = \mathbb{N} \cup \{\infty\},$$

což je spočetná množina.

Podívejme se na souvislost s klasickou definicí pravděpodobnosti. Zvolme nejprve pevné $n \in \mathbb{N}$ a hledejme pravděpodobnost elementárního jevu z Ω

$$\omega_k = \{\text{první líc nastal v } k\text{-tém hodu}\}$$

pomocí klasické pravděpodobnosti definované na Ω_n z příkladu 1.1. Pro $k \le n$ je uvažovaný elementární jev $\omega_k \subset \Omega$ ekvivalentní náhodnému jevu $C_k \in$

Ω_n, který je sjednocením elementárních jevů z Ω_n tvaru

$$(\underbrace{0, \ldots, 0}_{k-1}, 1, x_{k+1}, \ldots, x_n),$$

kde $x_{k+1}, \ldots, x_n$ jsou nuly nebo jedničky. Protože těchto $n - k$ posledních míst můžeme obsadit právě 2^{n-k} způsoby, je klasická pravděpodobnost náhodného jevu C_k rovna

$$\mathsf{P}_n(C_k) = \frac{2^{n-k}}{2^n} = 2^{-k}.$$

Všimněme si, že výsledná klasická pravděpodobnost *není* závislá na předem zvolené hodnotě n, tedy na tom, jak velkou konečnou délku posloupnosti rubů a líců jsme uvažovali. Zvolíme-li $\mathsf{P}(\omega_k) = \mathsf{P}_n(C_k)$, máme tak definovánu pravděpodobnost pro *každé* přirozené číslo k. Abychom rozšířili definici pravděpodobnosti na celou množinu $\Omega = \mathbb{N} \cup \{\infty\}$, spočítáme nejprve

$$\sum_{k=1}^{\infty} \mathsf{P}(\omega_k) = \sum_{k=1}^{\infty} \left(\frac{1}{2}\right)^k = 1.$$

Odtud plyne, že musí být $\mathsf{P}(\infty) = 0$. ○

Ani rozšíření na nekonečně mnoho elementárních jevů nemusí vždy stačit.

Příklad 1.8. Při kondičním běhání v Kunratickém lese jsem ztratil památeční kapesník bílé barvy. Pokud půjdu ke vzdálenější stanici metra, zopakuji si čtvrtinu z okruhu, na němž ke ztrátě došlo. Když předpokládám, že nikdo jiný (ani pes) můj kapesník nesebral, jaká je pravděpodobnost, že jej na této čtvrtině okruhu najdu, pokud je „stejně pravděpodobné", že jsem kapesník ztratil na libovolném místě okruhu? Známý zdravý selský rozum velí, že uvažovaná pravděpodobnost musí být rovna čtvrtině. Úsudek je zřejmě založen na porovnání délek dvou drah, tedy na analogii vzorce z klasické definice pravděpodobnosti. Pouze počet prvků konečné množiny byl nahrazen její velikostí (délkou, objemem). ○

Příklad 1.8 ukazuje na potřebu definovat pravděpodobnost pro množinu jako je interval reálných čísel. Naznačuje i cestu v podobě měření „velikosti" těchto množin.

1.4 Kolmogorovova definice pravděpodobnosti

V příkladu 1.7 jsme viděli, že je vhodné a možné dovolit více než konečně mnoho elementárních jevů. Dalším důvodem pro práci se spočetně mnoha

elementárními jevy bude potřeba popsat limitní chování a tedy pracovat s nekonečnou posloupností náhodných jevů. Proto poněkud rozšíříme definici algebry.

Definice 1.4. Nechť Ω je libovolná neprázdná množina. Neprázdný systém $\mathcal{A}$ podmnožin množiny Ω se nazývá σ-**algebra**, jestliže platí

$$A \in \mathcal{A} \quad \Rightarrow \quad A^c \in \mathcal{A}, \tag{1.18}$$

$$A_i \in \mathcal{A}, i = 1, 2, \ldots \quad \Rightarrow \quad \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}. \tag{1.19}$$

Vyjdeme z prostoru elementárních jevů Ω. Za náhodný jev budeme považovat každý prvek systému podmnožin $\mathcal{A}$ množiny Ω. O bohatosti tohoto systému vypovídá následující tvrzení.

Věta 1.2. Je-li Ω neprázdná množina výsledků náhodného pokusu a je-li $\mathcal{A}$ nějaká σ-algebra náhodných jevů definovaná na Ω, potom platí

$$\emptyset \in \mathcal{A}, \qquad \Omega \in \mathcal{A}, \tag{1.20}$$

$$A_1, \ldots, A_k \in \mathcal{A} \quad \Rightarrow \quad \bigcup_{i=1}^{k} A_i \in \mathcal{A}, \tag{1.21}$$

$$A_1, A_2, \ldots \in \mathcal{A} \quad \Rightarrow \quad \bigcap_{i=1}^{\infty} A_i \in \mathcal{A}, \tag{1.22}$$

$$A_1, \ldots, A_k \in \mathcal{A} \quad \Rightarrow \quad \bigcap_{i=1}^{k} A_i \in \mathcal{A}, \tag{1.23}$$

$$A, B \in \mathcal{A} \quad \Rightarrow \quad A - B \in \mathcal{A}. \tag{1.24}$$

D ů k a z: Protože systém $\mathcal{A}$ je neprázdný, můžeme pevně zvolit jednu množinu $A \in \mathcal{A}$. Podle (1.18) je také $A^c \in \mathcal{A}$, takže zvolíme-li $A_1 = A^c, A_2 = A_3 = \ldots = A$, dostaneme podle (1.19), že $\Omega \in \mathcal{A}$. Zbytek (1.20) plyne odtud pomocí (1.18). Máme-li dokázat (1.21), stačí doplnit posloupnost $A_1, \ldots, A_k$ posloupností prázdných množin $\emptyset$. Tvrzení (1.22) a (1.23) dostaneme pomocí známého de Morganova vzorce, (1.18) a (1.19), např.

$$\left(\bigcap_{i=1}^{\infty} A_i\right)^c = \bigcup_{i=1}^{\infty} A_i^c \in \mathcal{A}.$$

Poslední vztah (1.24) plyne z (1.23) a (1.18) pomocí

$$A - B = A \cap B^c.$$

□

Je dobré si uvědomit, že $\mathcal{A}$ nemusí být systémem *všech* podmnožin Ω, i když je tento systém nutně velmi bohatý. Pravděpodobnost definujeme pouze pro $A \in \mathcal{A}$ jako konečnou, nezápornou a σ-aditivní reálnou funkci.

Definice 1.5. (Kolmogorov) Nechť Ω je neprázdná množina, nechť $\mathcal{A}$ je σ-algebra náhodných jevů definovaných na Ω. **Pravděpodobností** se nazývá reálná funkce $\mathsf{P}(A)$ definovaná na $\mathcal{A}$, která pro $A \in \mathcal{A}$, $A_1, A_2, \ldots \in \mathcal{A}$, $A_i \cap A_j = \emptyset$ pro všechna $i \neq j$, splňuje

$$\mathsf{P}(\Omega) = 1, \tag{1.25}$$

$$\mathsf{P}(A) \geq 0, \tag{1.26}$$

$$\mathsf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathsf{P}(A_i). \tag{1.27}$$

Nejdůležitější vlastnosti takto zavedené pravděpodobnosti uvedeme v následující větě.

Věta 1.3. Je-li $A, B \in \mathcal{A}$, $A_1, \ldots, A_k \in \mathcal{A}$, $A_i \cap A_j = \emptyset$ pro všechna $i \neq j$, pak platí

$$\mathsf{P}(\emptyset) = 0, \tag{1.28}$$

$$0 \leq \mathsf{P}(A) \leq 1, \tag{1.29}$$

$$\mathsf{P}(A^c) = 1 - \mathsf{P}(A). \tag{1.30}$$

$$\mathsf{P}\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} \mathsf{P}(A_i), \tag{1.31}$$

$$A \subset B \quad \Rightarrow \quad \mathsf{P}(A) \leq \mathsf{P}(B), \tag{1.32}$$

$$A \subset B \quad \Rightarrow \quad \mathsf{P}(B - A) = \mathsf{P}(B) - \mathsf{P}(A). \tag{1.33}$$

Tvrzení o pravděpodobnosti sjednocení náhodných jevů shrneme do další věty.

Věta 1.4. Nechť je $A, B \in \mathcal{A}$, $A_i \in \mathcal{A}, 1 \leq i \leq n$. Potom platí

$$\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B), \tag{1.34}$$

$$\mathsf{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathsf{P}(A_i) - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathsf{P}(A_i \cap A_j) + \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} \mathsf{P}(A_i \cap A_j \cap A_k) \tag{1.35}$$

$$+\ldots+(-1)^{n-1}\mathsf{P}\left(\bigcap_{i=1}^{n}A_i\right).$$

D ů k a z: Tvrzení (1.34) plyne z vyjádření náhodného jevu $A\cup B$ ve tvaru

$$A\cup B=A\cup(B-A)=A\cup(B-(A\cap B)),$$

přičemž náhodné jevy A a $B-(A\cap B)$ jsou neslučitelné a náhodný jev $A\cap B$ je podjevem náhodného jevu B. Potom stačí použít (1.31) a (1.33). Tvrzení (1.35) se dokáže indukcí vzhledem k n. □

U monotonních posloupností náhodných jevů lze zaměňovat limitu a výpočet pravděpodobnosti, jak bude vidět z následující věty.

Věta 1.5. Nechť $A_1, A_2, \ldots \in \mathcal{A}$. Platí-li $A_1 \subset A_2 \subset \cdots$, pak je

$$\mathsf{P}\left(\bigcup_{i=1}^{\infty}A_i\right)=\lim_{i\to\infty}\mathsf{P}(A_i), \tag{1.36}$$

Platí-li $A_1 \supset A_2 \supset \cdots$, pak je

$$\mathsf{P}\left(\bigcap_{i=1}^{\infty}A_i\right)=\lim_{i\to\infty}\mathsf{P}(A_i). \tag{1.37}$$

Pro každou posloupnost náhodných jevů $\{A_i\}$ platí

$$\mathsf{P}\left(\bigcup_{i=1}^{\infty}A_i\right)\le\sum_{i=1}^{\infty}\mathsf{P}(A_i). \tag{1.38}$$

D ů k a z: K důkazu (1.36) zavedeme neslučitelné náhodné jevy

$$\begin{aligned} B_1 &= A_1,\\ B_2 &= A_2-A_1,\\ B_3 &= A_3-A_2,\\ &\cdots\ . \end{aligned}$$

Zřejmě platí $A=\bigcup_{i=1}^{\infty}A_i=\bigcup_{i=1}^{\infty}B_i$. Pro neslučitelné náhodné jevy B_i podle (1.27) z Kolmogorovovy definice pravděpodobnosti platí

$$\mathsf{P}(A)=\mathsf{P}\left(\bigcup_{i=1}^{\infty}B_i\right)=\sum_{i=1}^{\infty}\mathsf{P}(B_i)=\lim_{n\to\infty}\sum_{i=1}^{n}\mathsf{P}(B_i).$$

Pro konečné n je však vzhledem k předpokládané inklusi náhodných jevů A_i

$$\begin{aligned} \sum_{i=1}^{n}\mathsf{P}(B_i) &= \mathsf{P}(A_1)+\sum_{i=2}^{n}\mathsf{P}(A_i-A_{i-1})\\ &= \mathsf{P}(A_1)+\sum_{i=2}^{n}(\mathsf{P}(A_i)-\mathsf{P}(A_{i-1}))\\ &= \mathsf{P}(A_n), \end{aligned}$$

takže dohromady dostaneme

$$\mathsf{P}(A)=\lim_{n\to\infty}\mathsf{P}(A_n).$$

K důkazu (1.37) nejprve označme $A=\cap_{i=1}^{\infty}A_i$ a zaveďme náhodné jevy $C_i=A_i^c$, které splňují předpoklady první části věty. Pro náhodný jev C, definovaný vztahem

$$C=\bigcup_{i=1}^{\infty}C_i=\bigcup_{i=1}^{\infty}A_i^c=\left(\bigcap_{i=1}^{\infty}A_i\right)^c=A^c$$

platí podle prvního tvrzení věty $\mathsf{P}(C)=\lim_{i\to\infty}\mathsf{P}(C_i)$, takže podle (1.30) dostaneme

$$1-\mathsf{P}(A)=\lim_{i\to\infty}(1-\mathsf{P}(A_i))=1-\lim_{i\to\infty}\mathsf{P}(A_i).$$

□

Trojici $(\Omega,\mathcal{A},\mathsf{P})$ budeme opět nazývat **pravděpodobnostní prostor**.

1.5 Cvičení

1.1. Uvažujme rodiny, ve kterých jsou tři děti. Jaká je pravděpodobnost, že v náhodně vybrané rodině jsou dvě děvčata a jeden chlapec?

1.2. Osudí obsahuje celkem pět koulí, které se liší pouze barvou. Náhodně vytáhneme tři koule. Jaká je pravděpodobnost, že mezi nimi:
a) je bílá koule,
b) není modrá koule,
c) je bílá a není modrá koule?
(V osudí je bílá, modrá, černá, červená a zelená koule.)

1.3. V osudí je 8 bílých, 8 modrých a 8 červených koulí. Vytáhneme jednu kouli, zaznamenáme její barvu, kouli vrátíme, obsah osudí dobře promícháme a

opět vytáhneme jednu kouli. Jaká je pravděpodobnost, že obě koule mají stejnou barvu? Jak se tato pravděpodobnost změní, jestliže první vytaženou kouli do osudí nevracíme?

1.4. Házíme šesti hracími kostkami. Jaká je pravděpodobnost, že
a) padnou vesměs různá čísla,
b) padnou pouze lichá čísla?
Ještě než tyto pravděpodobnosti vypočítáte, zkuste předem uhodnout, která bude větší.

1.5. Třicet studentů bylo náhodně rozděleno do skupin po deseti osobách. Jaká je pravděpodobnost, že Yveta a Zdeněk se dostali do stejné skupiny? Jak se tato pravděpodobnost změní, jestliže nově vznikající skupinky budou mít 8, 12 a 10 osob?

1.6. Ze stovky očíslovaných vstupenek byly náhodně vylosovány tři. Jaká je pravděpodobnost, že tyto tři vstupenky lze uspořádat v aritmetickou posloupnost?

1.7. Jaká je pravděpodobnost, že se ve třídě, kde je n žáků, najde dvojice, která má narozeniny stejný den v roce? Pro jaké n je tato pravděpodobnost nejblíže hodnotě 0,5? (Neuvažujte přestupné roky, předpokládejte, že se během celého roku děti rodí rovnoměrně.)

1.8. Jaká je pravděpodobnost, že ve třídě, kde je n žáků, existuje spolužák, který má narozeniny stejný den jako třídní profesor?

2. Nezávislost

2.1 Podmíněná pravděpodobnost

Příklad 2.1. Mějme urnu s a černými a b bílými koulemi. Zajímá nás pravděpodobnost, s jakou *ve druhém* tahu bez vracení vytáhneme bílou kouli za předpokladu, že také v prvním tahu jsme vytáhli kouli bílou.

Označme jako B_i náhodný jev, že v i-tém tahu vytáhneme bílou kouli. Z klasické definice pravděpodobnosti plyne, že je

$$\mathsf{P}(B_1) = \frac{b}{a+b}.$$

Podobně v situaci, kdy jsme už vytáhli v prvním tahu bílou kouli, další bílou kouli vytáhneme s pravděpodobností

$$\mathsf{P}(B_2|B_1) = \frac{b-1}{a+b-1},$$

protože pro druhý tah je k disposici pouze $b-1$ bílých a a černých koulí. Označení $\mathsf{P}(B_2|B_1)$ jsme použili pro *podmíněnou pravděpodobnost náhodného jevu B_2 za podmínky výskytu náhodného jevu B_1*. Kromě toho je

$$\mathsf{P}(B_1 \cap B_2) = \frac{b(b-1)}{(a+b)(a+b-1)},$$

takže naši podmíněnou pravděpodobnost můžeme vyjádřit jako

$$\mathsf{P}(B_2|B_1) = \frac{\mathsf{P}(B_1 \cap B_2)}{\mathsf{P}(B_1)}. \qquad \bigcirc$$

Při výpočtu podmíněné pravděpodobnosti $\mathsf{P}(A|B)$ tedy omezujeme výchozí prostor elementárních jevů Ω na náhodný jev (množinu) B. Také z náhodného jevu (množiny) A bereme v úvahu jen tu jeho část, která je součástí B.

Definice 2.1. Mějme pravděpodobnostní prostor $(\Omega, \mathcal{A}, \mathsf{P})$, nechť $B \in \mathcal{A}$ splňuje podmínku $\mathsf{P}(B) > 0$. Potom **podmíněnou pravděpodobnost** náhodného jevu A za podmínky náhodného jevu B definujeme jako

$$\mathsf{P}(A|B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)}. \tag{2.1}$$

Definiční vztah (2.1) má za následek

$$\mathsf{P}(A \cap B) = \mathsf{P}(A|B)\mathsf{P}(B). \tag{2.2}$$

V definici jsme předpokládali, že je $\mathsf{P}(B) > 0$. Protože je však $A \cap B \subset B$, je pro $\mathsf{P}(B) = 0$ také $\mathsf{P}(A \cap B) = 0$. Proto má vztah (2.2) smysl i v případě, že je $\mathsf{P}(B) = 0$.

Výraz na levé straně vztahu (2.2) je symetrický v obou náhodných jevech, proto jej můžeme psát také jako

$$\mathsf{P}(A \cap B) = \mathsf{P}(B|A)\mathsf{P}(A). \tag{2.3}$$

Porovnání pravých stran (2.2) a (2.3) vede k

$$\mathsf{P}(A|B)\mathsf{P}(B) = \mathsf{P}(B|A)\mathsf{P}(A). \tag{2.4}$$

Definice 2.2. Mějme pravděpodobnostní prostor $(\Omega, \mathcal{A}, \mathsf{P})$. Řekneme, že náhodné jevy $A_1, A_2, \ldots \in \mathcal{A}$ tvoří **úplný systém jevů**, jestliže platí

$$A_i \cap A_j = \emptyset \text{ pro } i \neq j, \tag{2.5}$$

$$\bigcup_{i=1}^{\infty} A_i = \Omega. \tag{2.6}$$

Elementární jevy jsou také úplným systémem jevů, zřejmě tím nejjemnějším, o jakém lze v dané úloze uvažovat.

Věta 2.1. (Vzorec pro úplnou pravděpodobnost) Nechť $A_1, A_2, \ldots$ je úplný systém jevů v pravděpodobnostním prostoru $(\Omega, \mathcal{A}, \mathsf{P})$ takový, že platí

$$\mathsf{P}(A_i) > 0, \qquad i = 1, 2, \ldots. \tag{2.7}$$

Potom platí

$$\mathsf{P}(B) = \sum_{i=1}^{\infty} \mathsf{P}(B|A_i)\mathsf{P}(A_i). \tag{2.8}$$

D ů k a z: Uvažovanou pravděpodobnost můžeme s použitím definice úplného systému jevů postupně upravovat

$$\begin{aligned} \mathsf{P}(B) &= \mathsf{P}(B \cap \Omega) \\ &= \mathsf{P}(B \cap \bigcup_{i=1}^{\infty} A_i) \end{aligned}$$

$$\begin{aligned} &= \mathsf{P}(\bigcup_{i=1}^{\infty} (B \cap A_i)) \\ &= \sum_{i=1}^{\infty} \mathsf{P}(B \cap A_i). \end{aligned}$$

neboť náhodné jevy $B \cap A_1, B \cap A_2, \ldots$ jsou neslučitelné. Dokazované tvrzení dostaneme pomocí (2.3). □

Věta 2.2. (Bayesův vzorec) Nechť $A_1, A_2, \ldots$ je úplný systém jevů v pravděpodobnostním prostoru $(\Omega, \mathcal{A}, \mathsf{P})$ takový, že platí

$$\mathsf{P}(A_i) > 0, \qquad i = 1, 2, \ldots. \tag{2.9}$$

Jestliže je $\mathsf{P}(B) > 0$, pak platí

$$\mathsf{P}(A_j|B) = \frac{\mathsf{P}(B|A_j)\mathsf{P}(A_j)}{\sum_{i=1}^{\infty} \mathsf{P}(B|A_i)\mathsf{P}(A_i)}, \qquad j = 1, 2, \ldots \tag{2.10}$$

D ů k a z: Pomocí (2.4) můžeme pro zvolené j psát

$$\mathsf{P}(A_j|B) = \frac{\mathsf{P}(B|A_j)\mathsf{P}(A_j)}{\mathsf{P}(B)}.$$

Když do tohoto vztahu dosadíme podle vzorce pro úplnou pravděpodobnost, dostaneme dokazované tvrzení. □

Následující příklady ukazují, že někdy může být obtížné rozeznat podmíněnou pravděpodobnost od nepodmíněné, že není vždy snadné volit správný pravděpodobnostní model pro jejich výpočet.

Příklad 2.2. Tenista má prvé podání úspěšné s pravděpodobností 0,6, druhé s pravděpodobností 0,8. S jakou pravděpodobností p se tento tenista dopustí dvojchyby?

Označíme-li U_1, U_2 (N_1, N_2) náhodné jevy spočívající v úspěšném (neúspěšném) prvém a druhém podání, pak jediná správná interpretace pravděpodobností, které se objevují v zadání, vede k $\mathsf{P}(U_1) = 0{,}6, \mathsf{P}(U_2|N_1) = 0{,}8$, protože zvýšená pravděpodobnost úspěchu v druhém pokusu je vyvolána právě jen neúspěchem v pokuse prvém. Pravděpodobnost dvojchyby je možno nyní stanovit z definice podmíněné pravděpodobnosti:

$$p = \mathsf{P}(N_1 \cap N_2) = \mathsf{P}(N_2|N_1)\mathsf{P}(N_1) = 0{,}2 \cdot 0{,}4 = 0{,}08.$$

Zeptejme se nyní, který pravděpodobnostní prostor vlastně naši úlohu modeluje. Přirozenou odpovědí je prostor s množinou elementárních jevů $\Omega = \{U_1, (N_1, U_2), (N_1, N_2)\}$ s přiřazením pravděpodobností $\mathsf{P}(U_1) = 0{,}6$,

$$\mathsf{P}(N_1, U_2) = \mathsf{P}(U_2|N_1)\mathsf{P}(N_1) = 0{,}8 \cdot 0{,}4 = 0{,}32,$$

$\mathsf{P}(N_1, N_2) = p = 0{,}08$. Vzhledem k pravidlům tenisové hry je poněkud nepřirozené modelovat složitěji pomocí množiny $\Omega = \{(U_1, U_2), (U_1, N_2), (N_1, U_2), (N_1, N_2)\}$, přesto je to však možné, a to pomocí představy, že tenista podává po druhé i v případě, že byl v prvém podání úspěšný. Musí tedy být $\mathsf{P}(U_1, U_2) = \mathsf{P}(U_2|U_1) \cdot \mathsf{P}(U_1) = \mathsf{P}(U_1)\mathsf{P}(U_1) = 0{,}36$ a $\mathsf{P}(U_1, N_2) = \mathsf{P}(N_2|U_1)\mathsf{P}(U_1) = \mathsf{P}(N_1)\mathsf{P}(U_1) = 0{,}4 \cdot 0{,}6 = 0{,}24$. ○

Příklad 2.3. Favority dostihu jsou koně „a" a „b". Odborníci tipují, že „a" zvítězí s pravděpodobností 0,5, kůň „b" s pravděpodobností 0,3. Kůň „a" ztratil na startu tolik, že je již jisté, že nezvítězí. Jaká je nyní pravděpodobnost p, že zvítězí kůň „b"?

Buďte A, B náhodné jevy, které označují vítězství koně „a", resp. „b". Jediná správná interpretace pravděpodobnosti p je

$$p = \mathsf{P}(B|A^c) = \frac{\mathsf{P}(B \cap A^c)}{\mathsf{P}(A^c)} = \frac{\mathsf{P}(B)}{\mathsf{P}(A^c)} = \frac{0{,}3}{0{,}5} = 0{,}6,$$

protože $B \subset A^c$, když vyloučíme možnost mrtvého dostihu. ○

Příklad 2.4. Skříňka má tři zásuvky, v každé z nich jsou dvě mince, a to tak, že v jedné zásuvce jsou dvě zlaté, v další zlatá a stříbrná a ve zbývající zásuvce jsou dvě stříbrné mince. Náhodně otevřeme jednu zásuvku, náhodně z ní vybereme minci: je stříbrná. Jaká je nyní pravděpodobnost p, že v otevřené zásuvce zůstala zlatá mince?

Naivní rychlá odpověď $p = 1/2$ je nesprávná. Abychom pravděpodobnost p interpretovali dobře, jako pravděpodobnost podmíněnou, musíme zvážit, že pokus má dvě (nezávislé) náhodné fáze: volba zásuvky, volba mince. Je třeba také (uměle) rozlišovat individualitu jak jednotlivých zásuvek, tak i jednotlivých šesti mincí, aby naše konstrukce množiny elementárních jevů opravňovala užití klasického pravděpodobnostního modelu. Položíme-li

$$\Omega = \{(1, z_1), (1, z_2), (2, z_3), (2, s_1), (3, s_2), (3, s_3)\},$$

kde například $(1, z_2)$ označuje ten výsledek, že byla otevřena první zásuvka a z ní vyňata druhá zlatá mince, není důvod preferovat některý z šesti uvedených výsledků před ostatními a volba klasického modelu je oprávněná.

Označíme-li nyní jako Z náhodný jev spočívající v tom, že v otevřené zásuvce zůstane zlatá mince a jako S náhodný jev spočívající v tom, že z otevřené zásuvky byla vytažena mince stříbrná, je zřejmě $p = \mathsf{P}(Z|S)$. Protože $\mathsf{P}(S) = \mathsf{P}(\{(2, S_1), (3, S_2), (3, S_3)\}) = \frac{3}{6} = \frac{1}{2}$, $\mathsf{P}(Z \cap S) = \mathsf{P}(\{(2, S_1)\}) = \frac{1}{6}$, je $p = \frac{1}{3}$ a skutečně nikoliv $p = \frac{1}{2}$. ○

Příklad 2.5. Ve vězení očekávají tři lotři Alcapone, Babinský a Cimrman popravu. Popraveni budou však pouze dva, tuto dvojici již určil los, verdikt každému z nich však bude sdělen až za úsvitu. Alcapone se oklikou snaží posoudit své šance tak, že informovaného dozorce žádá: Jmenuj jednoho z mých spoluvězňů, který bude popraven! Dozorce je pravdomluvný, má-li více možností odpovědět, volí jméno náhodně. Tento dozorce odpoví – Babinský. Před rozhovorem věděl Alcapone, že bude popraven s pravděpodobností 2/3. Jaká je pravděpodobnost, řekněme p, nyní, po rozhovoru s dozorcem?

V průběhu celé této smutné příhody působí dva náhodné faktory – los a odpověď dozorce. Odpovídající prostor elementárních jevů je dán množinou $\Omega = \{(ab, b), (ac, c), (bc, b), (bc, c)\}$, kde prvá souřadnice jmenuje dvojici vylosovaných, druhá pak odpověď dozorce. Prvé dva elementární jevy jsou determinovány losem, proto je $\mathsf{P}(ab, b) = \mathsf{P}(ac, c) = \frac{1}{3}$. Zbývající dva reprezentují také možnou nahodilost dozorcovy odpovědi, proto je $\mathsf{P}(bc, b) = \mathsf{P}(bc, c) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$. Označíme-li nyní jako A náhodný jev spočívající v tom, že Alcapone bude popraven a B náhodný jev spočívající v tom, že dozorcova odpověď zní „Babinský", dostaneme $p = \mathsf{P}(A|B) = \frac{2}{3}$, protože $B = \{(ab, b), (bc, b)\}$ a $A \cap B = \{(ab, b)\}$, tj. $\mathsf{P}(B) = \frac{1}{3} + \frac{1}{6}$, $\mathsf{P}(A \cap B) = \frac{1}{3}$. Vidíme, že Alcapone žádnou informaci nezískal, že $\mathsf{P}(A) = \mathsf{P}(A|B) = \frac{2}{3}$ a jevy A, B jsou (možná překvapivě) nezávislé.

Modifikujme dozorcovo chování tak, že má-li více možností jak odpovědět, pak volí odpověď podle abecedy ($a > b > c$). Tím se ovšem mění pravděpodobnosti jednotlivých elementárních jevů: $\mathsf{P}(ab, b) = \mathsf{P}(ac, c) = \frac{1}{3}$, $\mathsf{P}(bc, b) = \frac{1}{3}$, $\mathsf{P}(bc, c) = 0$ a tudíž $\mathsf{P}(A|B) = \frac{1}{2}$. Dozorcova informace je tedy v tomto případě podstatná. ○

Ilustrujme nyní smysl a užitečnost Bayesova vzorce.

Příklad 2.6. Deseti bílými či černými koulemi byla urna naplněna tak, že bylo desetkrát hozeno symetrickou mincí. Padl-li rub (líc) mince, byla do urny vložena bílá (černá) koule. Takto náhodně naplněná urna je zkoumána pomocí pokusu, který spočívá v tom, že z urny je postupně taženo n koulí, každá z nich je však po zjištění barvy do urny vrácena. Výsledkem „diagnosy" je zjištění, že všech n tažených koulí má bílou barvu. Znalost stochastického mechanismu, který naplnil urnu, umožňuje ustanovit (apri-

orní) pravděpodobnosti jednotlivých hypotéz $B_0, \dots B_{10}$ o barevném složení urny: $\mathsf{P}(B_k) = \binom{10}{k} 2^{-10}$ (k bílých a $(10-k)$ černých koulí), $0 \le k \le 10$. Provedená diagnosa však přináší další informaci – zjištění, že nastal náhodný jev $B^{(n)}$ spočívající v tom, že všech n tažených koulí bylo bílých. Podmíněné (aposteriorní) pravděpodobnosti $\mathsf{P}(B_k|B^{(n)})$, $0 \le k \le 10$, tuto informaci respektují a můžeme je spočítat pomocí Bayesova vzorce. Označíme-li $p_{10}^{(n)} = \mathsf{P}(B_{10}|B^{(n)})$, pak

$$p_{10}^{(n)} = \frac{\mathsf{P}(B_{10})\mathsf{P}(B^{(n)}|B_{10})}{\sum_{k=0}^{10} \mathsf{P}(B_k)\mathsf{P}(B^{(n)}|B_k)} = \frac{2^{-10}}{\sum_{k=0}^{10} \binom{10}{k} 2^{-10} \left(\frac{k}{10}\right)^n} = \frac{1}{\sum_{k=0}^{10} \binom{10}{k} \left(\frac{k}{10}\right)^n},$$

protože $B_0, \dots, B_{10}$ je úplný systém jevů. Jelikož

$$\sum_{k=0}^{10} \binom{10}{k} \left(\frac{k}{10}\right)^n = \sum_{k=0}^{10} \binom{10}{k} \left(1 - \frac{k}{10}\right)^n \le \sum_{k=0}^{10} \binom{10}{k} \mathrm{e}^{-\frac{n}{10}} = \left(1 + \mathrm{e}^{-\frac{n}{10}}\right)^{10},$$

zjišťujeme, jak jsme asi očekávali, že $\lim_{n\to\infty} p_{10}^{(n)} = 1$, tj. že aposteriorní pravděpodobnost „zcela bílé urny“ při trvalém vytahování bílé koule je asymptoticky rovna jedné. Tabulka 2.1 udává hodnoty pravděpodobnosti $p_{10}^{(n)}$ pro některé počty tahů n. ○

n	10	20	30	40	50	60	80	100
$p_{10}^{(n)}$	0,0702	0,35233	0,6746	0,8667	0,9504	0,9823	0,9978	0,9997

Tabulka 2.1: Pravděpodobnost, že urna obsahuje právě 10 bílých koulí

Následující dvě úlohy ukáží, že aparát podmíněných pravděpodobností nebyl vytvořen pouze k řešení hravých úloh o tenise, dostizích atd., že je také schopen řešit i vážné aplikované úlohy.

Příklad 2.7. Dříve, než propukne nemoc D, lze její latentní existenci odhalit biologickým testem. Bohužel, toto zjištění není zcela jednoznačné. U skrytě nemocné osoby je test pozitivní s pravděpodobností 0,999, u zdravé osoby pouze s pravděpodobností 0,01. Test tedy onemocnění nemusí odhalit, na druhé straně může vyvolat falešný poplach. Předpokládáme, že sledovanou nemoc má 10 % vyšetřované populace. Zajímáme se o pravděpodobnost, že osoba s pozitivním testem má skutečně nemoc D. Zvolíme-li $A_1 = D$,

$A_2 = D^c$ a pro snazší čitelnost výsledných vztahů $B = P$, $B^c = N$, dostaneme podle Bayesova vzorce

$$\mathsf{P}(D|P) = \frac{\mathsf{P}(P|D) \cdot \mathsf{P}(D)}{\mathsf{P}(P|D) \cdot \mathsf{P}(D) + \mathsf{P}(P|D^c) \cdot \mathsf{P}(D^c)} \tag{2.11}$$

$$= \frac{0{,}999 \cdot 0{,}1}{0{,}999 \cdot 0{,}1 + 0{,}01 \cdot 0{,}9} \tag{2.12}$$

$$= 0{,}917\,355 \tag{2.13}$$

a podobně pro pravděpodobnost, že negativně reagující osoba je opravdu zdravá

$$\mathsf{P}(D^c|N) = \frac{\mathsf{P}(N|D^c) \cdot \mathsf{P}(D^c)}{\mathsf{P}(N|D^c) \cdot \mathsf{P}(D^c) + \mathsf{P}(N|D) \cdot \mathsf{P}(D)} \tag{2.14}$$

$$= \frac{0{,}99 \cdot 0{,}9}{0{,}99 \cdot 0{,}9 + 0{,}001 \cdot 0{,}1} \tag{2.15}$$

$$= 0{,}999\,888. \tag{2.16}$$

Ověřte si, že při apriorní pravděpodobnosti nemoci pouhé 1 % klesne pravděpodobnost, že pozitivně reagující je skutečně nemocný, na hodnotu 0,502 262, kdežto pravděpodobnost, že negativně reagující je skutečně zdravý, vzroste na hodnotu 0,999 990. ○

Příklad 2.8. O místo sekretářky se postupně uchází n dívek. Kdyby měl personální ředitel, který o přijetí rozhoduje, možnost posuzovat schopnosti uchazeček simultánně, dokázal by tuto skupinu uspořádat, a to od nejméně vhodné uchazečky se „jménem“ 1, přes jen o málo vhodnější dívku 2 až po dívku nejschopnější, se jménem n. Dívky se však o místo ucházejí postupně v náhodném pořadí $i_1, \dots, i_n$, odmítnutá uchazečka se již nikdy o místo znovu neucházi. Jednou z možností, jak se v této situaci chovat při výběru racionálně (nechceme-li se spolehnout na zkušenost a intuici personálního ředitele) spočívá v následujícím postupu: Je zvolena strategie s, $1 \le s \le n$, která je definována tak, že kandidátky $i_1, \dots, i_{s-1}$ jsou odmítnuty a přijata je prvá dívka mezi $i_s, i_{s+1}, \dots, i_n$, která je vhodnější, než všechny předchozí, tj. je přijata dívka i_{k_0}, kde $k_0 = \min\{k \ge s, i_k \ge i_1, i_2, \dots, i_{k-1}\}$. Tento postup má však svá úskalí. Nejen, že nemusí být vybrána nejlepší uchazečka n, může se také dokonce stát, že není vybrána uchazečka žádná; to právě tehdy, když se dívka n náhodně ocitne mezi uchazečkami $i_1, \dots, i_{s-1}$. Jak nyní zvolit strategii s, aby byla maximalizována pravděpodobnost $p(s,n)$ toho, že při strategii s je vybrána nejlepší dívka n. Je třeba mít také pod kontrolou pravděpodobnost $q(s,n)$ toho, že procedura skončí bez vybrané

kandidátky. Samozřejmě platí $p(1,n) = \frac{1}{n}$, $q(1,n) = 0$. Při volbě $n = 3$ a $s = 2$ jsou při jednotlivých náhodných pořadích příchodu kandidátek

$$1\,\underline{2}\,3 \quad 1\,\underline{3}\,2 \quad 2\,1\,\underline{3} \quad 2\,\underline{3}\,1 \quad 3\,1\,2 \quad 3\,2\,1$$

vybrány podtržené dívky. Je tedy $p(2,3) = \frac{3}{6} = \frac{1}{2}$, $q(2,3) = \frac{2}{6} = \frac{1}{3}$.

Počítejme nyní pravděpodobnosti obecně pro $s > 1$. Označíme-li jako A_k náhodný jev spočívající v tom, že nejschopnější kandidátka přišla jako k-tá, tj. $i_k = n$, pak

$$q(s,n) = \mathsf{P}\left(\bigcup_{k=1}^{s-1} A_k\right) = \sum_{k=1}^{s-1} \mathsf{P}(A_k) = (s-1)\frac{(n-1)!}{n!} = \frac{s-1}{n}.$$

Označíme-li jako B náhodný jev popisující situaci, že uchazečka n je vybrána, pak $\mathsf{P}(B|A_k) = 0$ pro $1 \le k \le s-1$ a

$$\mathsf{P}(B|A_k) = \frac{\binom{n-1}{k-1}(s-1)(k-2)!(n-k)!}{(n-1)!} = \frac{s-1}{k-1}$$

pro $s \le k \le n$. Zdůvodnění je prosté: za předpokladu A_k (na množině elementárních jevů A_k) dochází k optimálnímu výběru B právě tehdy, když nejlepší uchazečka mezi $i_1, \ldots, i_{k-1}$ je odmítnuta, tj. právě tehdy, když ji nalézáme mezi uchazečkami $i_1, \ldots, i_{s-1}$. Vzorec pro úplnou pravděpodobnost pak náš výpočet přivádí k závěru:

$$p(s,n) = \mathsf{P}(B) = \sum_{k=s}^{n} \mathsf{P}(A_k)\mathsf{P}(B|A_k) = \frac{1}{n}\sum_{k=s}^{n}\frac{s-1}{k-1} = \frac{s-1}{n}\sum_{k=s-1}^{n-1}\frac{1}{k}.$$

Protože posloupnost $p(1,n), p(2,n), \ldots, p(n,n)$ je jednovrcholová (dokažte!), je optimální strategie $s^* = s^*(n)$, $p(s^*,n) = \max\limits_{1\le s\le n} p(s,n)$, dána jako řešení nerovností

$$p(s^*-1,n) \le p(s^*,n) \ge p(s^*+1,n)$$

$$\Leftrightarrow$$

$$\frac{1}{s^*} + \frac{1}{s^*+1} + \ldots + \frac{1}{n-1} \le 1 \le \frac{1}{s^*-1} + \frac{1}{s^*} + \frac{1}{s^*+1} + \ldots + \frac{1}{n-1}.$$

Eulerův vzorec pro částečné součty harmonické řady

$$\sum_{k=1}^{N}\frac{1}{k} = c + \ln(N) + \mathrm{O}(N^{-1}) \quad \text{při } N \to \infty,$$

kde $c = 0{,}577215$ (viz [7, odst. 368, vzorec (4)]) umožňuje přibližný výpočet (přesnost roste s počtem uchazeček n):

$$p(s,n) \doteq \frac{s-1}{n}\ln\left(\frac{n-1}{s-2}\right) \doteq \frac{s}{n}\ln\frac{n}{s} = -\frac{s}{n}\ln\frac{s}{n}.$$

Optimální hodnota podílu s/n je tedy asymptoticky určena jako bod $x_0 \in (0,1)$, ve kterém funkce $f(x) = -x\ln(x)$ nabývá maxima, tj. $x_0 = \mathrm{e}^{-1} \doteq 0{,}3679$. Jelikož $f(\mathrm{e}^{-1}) = \mathrm{e}^{-1}$, přicházíme k závěru, že $s^*(n) \doteq n\mathrm{e}^{-1}$, $p(s^*(n),n) \doteq \mathrm{e}^{-1}$ a

$$q(s^*,n) \doteq \frac{s^*-1}{n} \doteq \mathrm{e}^{-1}.$$

n	s^*	$p(s^*,n)$	$q(s^*,n)$
3	2	0,5000	0,3333
4	2	0,4583	0,2500
5	3	0,4333	0,4000
10	4	0,3987	0,3000
20	8	0,3842	0,3500
50	19	0,3743	0,3600
100	38	0,3710	0,3700
500	185	0,3685	0,3680
1000	369	0,3682	0,3680

Tabulka 2.2: Ukázky pravděpodobností úspěšného vybrání nejlepší a nevybrání žádné uchazečky

Přesnější analýzou lze dokázat

$$\lim_{n\to\infty}\frac{s^*(n)}{n} = \mathrm{e}^{-1},$$
$$\lim_{n\to\infty} p(s^*(n),n) = \mathrm{e}^{-1}.$$

Tabulka 2.2 ilustruje naše předchozí úvahy numericky. ○

Učiňme ještě jednu velmi důležitou poznámku. Podmíněnou pravděpodobnost $\mathsf{P}(B|A_k)$ jsme v posledním příkladu počítali jako (nepodmíněnou) pravděpodobnost náhodného jevu $B \cap A_k$ v modelu klasického pravděpodobnostního prostoru s množinou elementárních jevů A_k. Tento v počtu pravděpodobnosti velmi obvyklý postup odráží jednoduché rozšíření zlomku:

$$\frac{|B\cap A_k|}{|A_k|} = \frac{\frac{|B\cap A_k|}{|\Omega|}}{\frac{|A_k|}{|\Omega|}},$$

kde Ω je základní prostor elementárních jevů pro naši úlohu, tj. prostor permutací řádu n.

2.2 Nezávislost náhodných jevů

Bude přirozené říkat, že náhodný jev A nezávisí na náhodném jevu B, jestliže platí současně

(2.17) $$\mathsf{P}(A) = \mathsf{P}(A|B), \quad \mathsf{P}(A) = \mathsf{P}(A|B^c).$$

Uvedené vztahy mají smysl, jen když je $0 < \mathsf{P}(B) < 1$. Použijeme-li definici podmíněné pravděpodobnosti, dostaneme ekvivalentní vyjádření

$$\mathsf{P}(A)\mathsf{P}(B) = \mathsf{P}(A \cap B) \tag{2.18}$$

a

$$\mathsf{P}(A)(1 - \mathsf{P}(B)) = \mathsf{P}(A \cap B^c) \tag{2.19}$$

Protože je $A \cap B^c = A \cap (\Omega - B) = A - A \cap B$ a $A \cap B \subset A$, můžeme poslední rovnici upravit na

$$\mathsf{P}(A) - \mathsf{P}(A)\mathsf{P}(B) = \mathsf{P}(A) - \mathsf{P}(A \cap B), \tag{2.20}$$

což je vztah ekvivalentní s (2.18). Všimněme si, že v tomto vztahu vystupují náhodné jevy A a B symetricky, že tedy jde o symetrický vztah mezi nimi. Ke vztahu (2.18) dojdeme, i když je $\mathsf{P}(B) = 0$ nebo $\mathsf{P}(B) = 1$. Proto budeme používat následující definici.

Definice 2.3. Mějme pravděpodobnostní prostor $(\Omega, \mathcal{A}, \mathsf{P})$. Náhodné jevy A a B se nazývají **nezávislé**, když platí

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

Náhodné jevy $A_1, A_2, \ldots$ se nazývají **nezávislé**, jestliže pro každé $k \in \mathbb{N}$ a pro každou k-tici náhodných jevů $A_{i_1}, \ldots, A_{i_k}$ platí

$$\mathsf{P}\Big(\bigcap_{j=1}^{k} A_{i_j}\Big) = \prod_{j=1}^{k} \mathsf{P}(A_{i_j})$$

Náhodné jevy $A_1, A_2, \ldots$ se nazývají **po dvou nezávislé**, jestliže každé dva z nich jsou nezávislé.

Na rozdíl od neslučitelnosti náhodných jevů, což je množinový pojem, nelze pojem nezávislosti náhodných jevů definovat bez zavedení pravděpodobnosti.

Příklad 2.9. V urně jsou 4 lístky označené po řadě 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 zjistí, že platí

$$\mathsf{P}(A_1) = \mathsf{P}(A_2) = \mathsf{P}(A_3) = \frac{1}{2},$$

$$\mathsf{P}(A_1 \cap A_2) = \mathsf{P}(A_1 \cap A_3) = \mathsf{P}(A_2 \cap A_3) = \frac{1}{4},$$

$$\mathsf{P}(A_1 \cap A_2 \cap A_3) = 0,$$

takže náhodné jevy A_1, A_2, A_3 jsou po dvou nezávislé, ale *nejsou nezávislé*. ○

Uvedeme dvě důležité vlastnosti spojené s pojmem nezávislost.

Věta 2.3. Nechť $A_1, A_2, \ldots, A_n$ jsou nezávislé náhodné jevy. Pak

(a) Libovolná posloupnost typu $\tilde{A}_1, \tilde{A}_2, \ldots, \tilde{A}_n$, kde $\tilde{A}_k = A_k$ nebo $\tilde{A}_k = A_k^c$, je posloupnost nezávislých náhodných jevů.

(b) Platí

$$\mathsf{P}\left(\bigcup_{k=1}^{n} A_k\right) = 1 - \prod_{k=1}^{n} (1 - \mathsf{P}(A_k)).$$

D ů k a z: Pro ověření (a) stačí prokázat implikaci

$$A_1, A_2 \text{ jsou nezávislé} \Rightarrow A_1, A_2^c \text{ jsou nezávislé}$$

a dále pak postupovat užitím matematické indukce. Platí však $\mathsf{P}(A_1 \cap A_2^c) = \mathsf{P}(A_1 - A_1 \cap A_2) = \mathsf{P}(A_1) - \mathsf{P}(A_1)\mathsf{P}(A_2) = \mathsf{P}(A_1)(1 - \mathsf{P}(A_2)) = \mathsf{P}(A_1)\mathsf{P}(A_2^c)$. Tvrzení (b) plyne z de Morganovy formule a z tvrzení (a) následovně:

$$1 - \mathsf{P}\left(\bigcup_{k=1}^{n} A_k\right) = \mathsf{P}\left(\left(\bigcup_{k=1}^{n} A_k\right)^c\right) = \mathsf{P}\left(\bigcap_{k=1}^{n} A_k^c\right) = \prod_{k=1}^{n} (1 - \mathsf{P}(A_k)).$$

□

Vzorec (b) ukazuje, jak propastný rozdíl je mezi pojmy *neslučitelné* náhodné jevy a *nezávislé* náhodné jevy. Pro neslučitelné jevy $A_1, \ldots, A_n$ platí $\mathsf{P}(\cup_{k=1}^{n} A_k) = \sum_{k=1}^{n} \mathsf{P}(A_k)$!

Příklad 2.10. Navážeme na příklad 2.1 s urnou, která obsahuje a koulí černých a b koulí bílých. Postupně vytáhneme dvě koule. Náhodný jev A_i spočívá ve vytažení černé koule v i-tém tahu, B_i spočívá ve vytažení bílé koule v i-tém tahu. Zřejmě je $A_i = B_i^c$, $i = 1, 2$. Dále rozlišíme dvě různé situace podle toho, zda se koule vytažená v prvním tahu do urny vrátí či nikoliv.

(a) Předpokládejme, že první vytažená koule se před druhým tahem vrací.

Potom dvojice (A_1, A_2), (A_1, B_2), (B_1, A_2), (B_1, B_2) jsou dvojicemi nezávislých náhodných jevů, neboť platí například

$$\begin{aligned} \mathsf{P}(A_1 \cap B_2) &= \frac{ab}{(a+b)^2} \\ &= \frac{a(a+b)}{(a+b)^2}\,\frac{(a+b)b}{(a+b)^2} = \mathsf{P}(A_1)\mathsf{P}(B_2). \end{aligned}$$

(b) Předpokládejme nyní, že kouli vytaženou v prvním tahu před druhým tahem do urny nevracíme. Potom množina Ω obsahuje celkem $|\Omega| = (a+b)(a+b-1)$ stejně pravděpodobných elementárních jevů. Platí například

$$\mathsf{P}(A_1 \cap B_2) = \frac{ab}{(a+b)(a+b-1)},$$

kdežto

$$\begin{aligned} \mathsf{P}(A_1)\mathsf{P}(B_2) &= \frac{a(a+b-1)}{(a+b)(a+b-1)}\,\frac{ab+b(b-1)}{(a+b)(a+b-1)} \\ &= \frac{a}{a+b}\,\frac{b}{a+b}, \end{aligned}$$

takže náhodné jevy A_1 a B_2 nejsou nezávislé. Totéž platí nutně o dvojicích náhodných jevů $(A_1, A_2 = B_2^c)$, $(B_1 = A_1^c, B_2)$, $(B_1 = A_1^c, A_2)$. ○

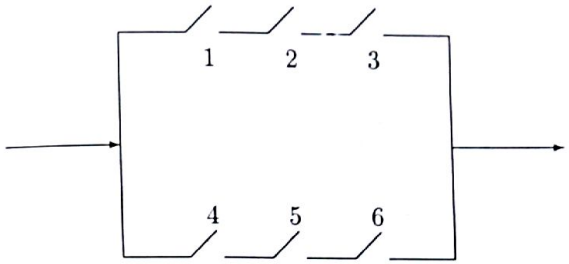
Příklad 2.11. Uvažme pokus, který má dva možné výsledky: zdar Z s pravděpodobností $1/2 + \varepsilon$, nezdar N s pravděpodobností $1/2 - \varepsilon$ $(\varepsilon > 0)$. Pokus dvakrát nezávisle opakujeme. Pokuste se odhadnout, který z náhodných jevů $A = \{(Z,Z),(N,N)\}$, $B = \{(Z,N),(N,Z)\}$ je pravděpodobnější. Přesná odpověď je dána následující úvahou: Je-li Z_i resp. N_i, $i = 1, 2$, náhodný jev spočívající ve zdaru resp. nezdaru i-tého pokusu, musí být náhodné jevy Z_1, Z_2 modelovány jako nezávislé s $\mathsf{P}(Z_1) = \mathsf{P}(Z_2) = 1/2 + \varepsilon$. Podle tvrzení (a) ve větě 2.3 platí

$$\begin{aligned} \mathsf{P}(A) &= \mathsf{P}(Z_1 \cap Z_2) + \mathsf{P}(N_1 \cap N_2) \\ &= \left(\frac{1}{2}+\varepsilon\right)^2 + \left(\frac{1}{2}-\varepsilon\right)^2 = \frac{1}{2} + 2\varepsilon^2 \\ &> \frac{1}{2} - 2\varepsilon^2 = 2\left(\frac{1}{2}+\varepsilon\right)\left(\frac{1}{2}-\varepsilon\right) \\ &= \mathsf{P}(Z_1 \cap N_2) + \mathsf{P}(N_1 \cap Z_2) = \mathsf{P}(B). \end{aligned}$$ ○

Příklad 2.12. Elektrický obvod je náhodně přerušován šesti nezávislými vypínači zapojenými podle schématu na obrázku 2.1. Každý z vypínačů je se stejnou pravděpodobností vypnut nebo zapnut. Určete pravděpodobnost p toho, že obvodem prochází proud.

Je-li A_i náhodný jev, který spočívá v tom, že vypínač i je sepnut, pak formulace úlohy vyvolává předpoklad, že $A_1, \ldots, A_6$ jsou nezávislé náhodné jevy s $\mathsf{P}(A_i) = \frac{1}{2}$. Proto je

$$\begin{aligned} p &= \mathsf{P}\left((A_1 \cap A_2 \cap A_3) \cup (A_4 \cap A_5 \cap A_6)\right) \\ &= \mathsf{P}\left(\cap_1^3 A_i\right) + \mathsf{P}\left(\cap_4^6 A_i\right) - \mathsf{P}\left(\cap_1^6 A_i\right) \\ &= 2\left(\frac{1}{2}\right)^3 - \left(\frac{1}{2}\right)^6 = \frac{15}{64}. \end{aligned}$$ ○



Obrázek 2.1: Schéma elektrického obvodu

2.3 Cvičení

2.1. Nechť platí $\mathsf{P}(A) = 0{,}3$, $\mathsf{P}(B) = 0{,}4$, $\mathsf{P}(A \cup B) = 0{,}6$. Spočítejte podmíněné pravděpodobnosti $\mathsf{P}(A|B)$ a $\mathsf{P}(B|A)$ a rozhodněte o závislosti či nezávislosti náhodných jevů A, B.

2.2. Náhodné jevy A, B, C jsou nezávislé a mají stejnou pravděpodobnost rovnou 0,1. Určete $\mathsf{P}(A \cup B \cup C)$.

2.3. Náhodné jevy A, B splňují $\mathsf{P}(B|A) = 0{,}216$, $\mathsf{P}(A) = 0{,}9$, $\mathsf{P}(B) = 0{,}25$. Určete $\mathsf{P}(A|B)$ a $\mathsf{P}(A - B)$.

2.4. Házíme dvěma hracími kostkami. Jev A znamená, že na modré kostce padlo liché číslo, jev B znamená, že na zelené kostce padlo sudé číslo, jev C znamená, že součet obou čísel je lichý. Jsou náhodné jevy A, B, C nezávislé? Jsou náhodné jevy A, B, C po dvou nezávislé?

2.5. Průměrně 90 % výrobků odpovídá požadavkům normy. Požadavkům zjednodušené zkoušky vyhoví standardní výrobek s pravděpodobností 0,95, kdežto nestandardní výrobek s pravděpodobností 0,20. S jakou pravděpodobností výrobek, který úspěšně prošel zjednodušenou zkouškou, splňuje také požadavky normy (je standardní)? S jakou pravděpodobností je výrobek, který neprošel zjednodušenou zkouškou, není standardní?

2.6. Otec chce povzbudit synovu zálibu (hrát tenis), proto nabídl hodnotný dar v případě, že syn zvítězí aspoň ve dvou po sobě jdoucích tenisových utkáních. Syn si může vybrat ze dvou možných strategií: hrát postupně s panem Novákem, s otcem a s panem Novákem (strategie A) nebo s otcem, s panem Novákem a s otcem (strategie B). Kterou strategii si vtipný syn zvolí, když ví, že pan Novák hraje podstatně lépe než jeho otec?

2.7. Dva korektoři četli nezávisle na sobě stejný text. První z nich objevil celkem a tiskových chyb, druhý celkem b tiskových chyb, z nichž c objevil také první korektor. Odhadněte, kolik neodhalených chyb v rukopisu ještě zůstalo.

2.8. Dva stejně silní hráči hrají opakovaně partie hry, v níž není možný nerozhodný výsledek. Rozhodněte, zda je pravděpodobnější, aby hráč A vyhrál právě tři partie ze čtyř nebo aby vyhrál právě pět partií z osmi. Jak se toto porovnání změní, žádáme-li, aby vyhrál aspoň tři resp. pět partií?

3. Některé klasické modely

Značná část úloh kombinatorické pravděpodobnosti může být modelována jako proces m tahů z urny, která obsahuje M rozlišitelných koulí. Použité kombinatorické pojmy a označení jsou shrnuty v dodatku A1

3.1 Výběr s vracením

Výběr s vracením je prováděn tak, že po každém tahu je vybraná koule vrácena do urny. Výsledek tohoto experimentu je popsán seznamem $a_1, \ldots, a_m$ tažených koulí ($1 \le a_i \le M$). Definice prostoru výsledků experimentu Ω podstatným způsobem závisí na tom, zda ku příkladu považujeme výsledky 4, 1, 2, 1 a 1, 4, 2, 1 za různé či nikoliv, tj. zda nás zajímá *pořadí*, v jakém byly koule taženy či se spokojujeme pouze se zjištěním kolikrát byla která koule tažena. V prvním případě je

$$\Omega = (M)^m = [\text{posloupnosti prvků množiny } \{1, \ldots, M\} \text{ o délce } m],$$

$|\Omega| = M^m$ a náhodný experiment se nazývá **uspořádaný výběr s vracením**. Jistě snadno nahlédneme, že není důvod preferovat některý takto definovaný výsledek experimentu před ostatními. Tento náhodný pokus tudíž modelujeme jako klasický pravděpodobnostní prostor s množinou výsledků Ω. V druhém případě, kdy nás zajímají četnosti (násobnosti) vytažených koulí, jsou výsledky 4, 1, 2, 1 a 1, 4, 2, 1 dostatečně popsány posloupností 1, 1, 2, 4 (koule 1 byla tažena dvakrát, dále byly taženy koule 2 a 4). Obecně je v tomto případě množina výsledků definována jako

$$\Omega = C'(M,m) = [\text{neklesající posloupnosti prvků množiny } \{1, 2, \ldots, M\} \text{ o délce } m],$$

tj. (viz (A3))

$$|\Omega| = \binom{M-1+m}{m}$$

a náhodný experiment se nazývá **neuspořádaný výběr s vracením**. Je velmi důležité si uvědomit, že neuspořádané výběry $C'(M,m)$ definují rozklad množiny v uspořádaných výběrů $(M)^m$ do $\binom{M+m-1}{m}$ částí (dvě posloupnosti z $(M)^m$ se nalézají v jedné části rozkladu právě tehdy, když je lze přerovnat do stejné neklesající posloupnosti z $C'(M,m)$). Je však zřejmé, že jednotlivé části rozkladu jsou různě početné. Uvažme například neuspořádaný výběr 1, 1, 1, 1, kterému odpovídá jediný výběr uspořádaný, tj. opět 1, 1, 1, 1. Uvážíme-li však neuspořádaný výběr 1, 1, 2, 4, zjistíme, že k němu

přináleží $4!/2! = 12$ výběrů uspořádaných. Neuspořádaný výběr s vracením *nemůže* tedy být modelován jako klasický pravděpodobnostní prostor na množině $\Omega = C'(M,m)$. Nemáme jinou možnost než definovat pravděpodobnost neuspořádaného výběru jako klasickou pravděpodobnost odpovídající části rozkladu množiny uspořádaných výběrů: neuspořádaný výběr lze jednoznačně charakterizovat celočíselným vektorem $(m_1,\dots,m_M)$, kde $0 \le m_i \le m$ je počet tahů koule $1 \le i \le M$ ($\sum_{i=1}^{M} m_i = m$). Označíme-li jako $R_{m_1,\dots,m_M} \subset (M)^m$ množinu uspořádaných výběrů, kde koule i byla tažena m_i-krát, pak

$$|R_{m_1,\dots,m_M}| = \binom{m}{m_1}\binom{m-m_1}{m_2}\cdots\binom{m_M}{m_M} = \frac{m!}{m_1!m_2!\cdots m_M!},$$

tj.

$$\mathsf{P}(R_{m_1,\dots,m_M}) = \frac{m!}{m_1!m_2!\cdots m_M!}M^{-m}.$$

Pravděpodobnosti

$$(3.1)\quad p_{m_1,\dots,m_M} = \frac{m!}{m_1!m_2!\cdots m_M!}M^{-m}, \quad 0 \le m_i \le M, \quad \sum_{i=1}^{M} m_i = m$$

definují pravděpodobnostní prostor s množinou elementárních jevů $\Omega = C'(M,m)$, který správně modeluje pokus, který jsme nazvali neuspořádaný výběr s vracením.

Oba výběry s vracením budou připomenuty v odst. 3.3 a 3.4 v souvislosti se dvěma modely rozmístění částic ve fyzikálním prostoru (Maxwellův-Boltzmannův, Boseův-Einsteinův).

3.2 Výběr bez vracení

Předpokládejme, že je $m \le M$ a že vytažené koule nejsou do urny vraceny. I v tomto případě uvažujme opět dvě definice výsledku pokusu:

Uspořádaný výběr bez vracení je náhodný pokus s množinou

$$\Omega = V(M,m) = [\text{posloupnosti prvků množiny } \{1,2,\dots M\} \text{ bez opakování o délce } m],$$

tj.

$$|\Omega| = \binom{M}{m}m! = \frac{M!}{(M-m)!}.$$

Neuspořádaný výběr bez vracení je náhodný pokus s množinou

$$\Omega = C(M,m) = [\text{podmnožiny v } \{1,2,,\dots,M\} \text{ s mohutností } m],$$

tj.

$$|\Omega| = \binom{M}{m}.$$

Uspořádaný výběr odpovídá potřebě sledovat nejen identitu, ale i pořadí tažených koulí, neuspořádaný výběr poskytuje pouze informaci o tom, které koule byly taženy. V případě uspořádaného výběru je nepochybně správný model klasického pravděpodobnostního modelu. Uvážíme-li, že každý neuspořádaný výběr (množina v $C(M,m)$) lze právě $m!$ způsoby permutovat do výběru uspořádaného, vidíme, že neuspořádané výběry reprezentují rozklad množiny uspořádaných výběrů do $\binom{M}{m}$ částí, z nichž každá má právě $m!$ prvků. Na rozdíl od výběru bez vracení platí, že *klasický pravděpodobnostní prostor je správný model jak pro uspořádaný, tak i pro neuspořádaný výběr bez vracení.*

Rekapitulaci uvádí tabulka 3.1.

Výběr s vracením		
	uspořádaný	neuspořádaný
Ω	posloupnosti délky m prvků v $\{1,2,\dots,M\}$	neklesající posloupnosti délky m prvků v $\{1,2,\dots,M\}$
$p(\omega)$	M^{-m}	$p_{m_1,\dots,m_M}$, viz (3.1)

Výběr bez vracení		
	uspořádaný	neuspořádaný
Ω	posloupnosti délky m prvků v $\{1,2,\dots,M\}$ bez opakování	podmnožiny v $\{1,2,\dots,M\}$ o mohutnosti m
$p(\omega)$	$\frac{(M-m)!}{M!}$	$\binom{M}{m}^{-1}$

Tabulka 3.1: Porovnání výběru bez vracení s výběrem s vracením

3.3 Maxwellův-Boltzmannův model

Uvažujme r částic, z nichž každá je právě v jedné z n přihrádek. Předpokládáme, že

1. Částice jsou *rozlišitelné*.
2. Pro každou částici umíme určit přihrádku, v níž je částice umístěna.
3. Stav systému lze udat tak, že pro každou částici udáme přihrádku, v níž je částice umístěna.
4. Všechny stavy jsou stejně pravděpodobné.

Uvedené předpoklady umožňují použít klasický pravděpodobnostní prostor s elementárními jevy tvaru

$$\omega = (s_1, \ldots, s_r),$$

kde každé z celých čísel $s_j, 1 \le j \le r$, nabývá některé z hodnot $1, \ldots, n$. Elementární jev je tedy totožný s vektorem „adres" jednotlivých částic, udávajících jejich umístění. Takovýchto uspořádaných n-tic je celkem $|\Omega| = n^r$ (r prvkové variace z n prvků s opakováním).

Zamyslíme-li se, jakým náhodným pokusem lze Maxwellovo-Boltzmannovo rozmístění realizovat, nabízí se model, který jsme nazvali uspořádaný výběr s vracením s $M = n$ a $m = r$. Z urny r krát vytáhneme s vracením přihrádku, do přihrádky tažené jako i-té v pořadí zařadíme předmět i a jistě obdržíme rozmístění, jehož stochastická konstrukce splňuje požadavky 1–4.

Uvážíme nyní náhodné veličiny $K_1, \ldots, K_n$, které označují počet předmětů v přihrádkách $1, 2, \ldots, n$. Nalezneme rozdělení náhodné veličiny K_1 (ostatní veličiny $K_2, \ldots, K_n$ mají zřejmě stejná rozdělení pravděpodobnosti). Platí

$$|[K_1 = k]| = \binom{r}{k}(n-1)^{r-k},$$

kde první činitel na pravé straně určuje počet způsobů, kolikrát lze vybrat k částic do sledované přihrádky, druhý činitel udává počet způsobů, kolikrát lze rozmístit ostatní částice ve zbývajících přihrádkách. Je tedy

$$p_{k,n,r} = \mathsf{P}[K_1 = k] = \binom{r}{k}\frac{(n-1)^{r-k}}{n^r}.$$

Nyní vyšetříme asymptotické chování této pravděpodobnosti v případě, že s rostoucím n roste také počet předmětů r_n tak, že v limitě je průměrný počet předmětů připadajících na jednu přihrádku stabilní:

$$\lim_{n\to\infty}\frac{r_n}{n} = \lambda.$$

Pomocí elementárního kalkulu dostaneme pro $k = 0, 1, \ldots$

$$\begin{aligned} p_k &= \lim_{n\to\infty} p_{k,n,r} \\ &= \lim_{n\to\infty}\binom{r_n}{k}\frac{(n-1)^{r_n-k}}{n^{r_n}} \\ &= \lim_{n\to\infty}\frac{r_n(r_n-1)\cdots(r_n-k+1)}{(n-1)^k}\frac{1}{k!}\left(1-\frac{1}{n}\right)^{r_n} \\ &= \frac{\lambda^k}{k!}\lim_{n\to\infty}\left(1-\frac{\frac{r_n}{n}}{r_n}\right)^{r_n} \\ &= \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda} \qquad (3.2) \end{aligned}$$

Snadno lze ověřit, že platí $\sum_{k=0}^{\infty} p_k = 1$, takže můžeme definovat prostor elementárních jevů $\Omega_P = \{0, 1, \ldots\}$ s pravděpodobnostmi p_k jednotlivých elementárních jevů.

Význam tohoto limitního přechodu spočívá v tom, že jsou-li n, r velká přirozená čísla, $\lambda = r/n$, pak lze pravděpodobnost $\mathsf{P}[K_1 = k] = p_{k,n,r}$ aproximovat pravděpodobností $p_k = \lambda^k \mathrm{e}^{-\lambda}/k!$. Přesnost této aproximace dokládá tabulka 3.2 počítaná pro $r = 500$ a $n = 365$ ($\lambda = 500/365$). (Tato tabulka udává například pravděpodobnost toho, že ve skupině 500 osob nemá nikdo narozeniny 2. února, je přibližně rovna 0,2537.) Pomocí binomické věty (po náhradě $j = k - 1$) snadno stanovíme společnou střední hodnotu náhodných veličin $K_1, \ldots, K_n$:

k	$p_{k,n,r}$	p_k
0	0,2537	0,2541
1	0,3484	0,3481
2	0,2388	0,2385
3	0,1089	0,1089
4	0,0372	0,0373
5	0,0101	0,0102
6	0,0023	0,0023

Tabulka 3.2: Porovnání přesné pravděpodobnosti $p_{k,n,r}$ s její aproximací p_k

$$\begin{aligned} \mathsf{E}K_1 &= \sum_{k=0}^{r} k\binom{r}{k}\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^{r-k} \\ &= \frac{r}{n}\sum_{j=0}^{r-1}\binom{r-1}{j}\left(\frac{1}{n}\right)^j\left(1-\frac{1}{n}\right)^{r-1-j} = \frac{r}{n} \end{aligned}$$

(průměrný počet předmětů na jednu přihrádku). Naprosto stejným způsobem jako v odstavci 3.1 vyšetříme nyní „sdružené rozdělení pravděpodobností" (podrobněji v 5. kapitole) náhodných veličin $K_1, \ldots, K_n$. Pro nezáporná celá čísla $r_1, \ldots, r_n$ taková, že $r_1 + \ldots + r_n = n$, platí

$$p_{r_1,\ldots,r_n} = \mathsf{P}[K_1 = r_1, \ldots, K_n = r_n]$$

$$= n^{-r}\binom{r}{r_1}\binom{r-r_1}{r_2}\binom{r-r_1-r_2}{r_3}\cdots\binom{r_n}{r_n}$$
$$= n^{-r}\frac{r!}{r_1!r_2!\ldots r_n!},$$

protože rozmístění $[K_1 = r_1, \ldots, K_n = r_n]$ je jednoznačně určeno některým *rozkladem* množiny (r) na části $(r_1), \ldots, (r_n)$.

Řešme nyní další úlohu, určit pravděpodobnost q toho, že existuje *prázdná* přihrádka. Označíme-li

$$B_i = \{i\text{-tá přihrádka je prázdná}\},$$

pak je

$$q = \mathsf{P}\left(\bigcup_{i=1}^{n} B_i\right).$$

Pozor, náhodné jevy $B_1, \ldots, B_n$ nejsou neslučitelné. Použijeme-li opět původní pravděpodobnostní prostor Ω, dostaneme snadno

$$\begin{aligned}
\mathsf{P}(B_i) &= \frac{(n-1)^r}{n^r}, \qquad i = 1, \ldots, n \\
\mathsf{P}(B_i \cap B_j) &= \frac{(n-2)^r}{n^r}, \qquad i \neq j, \quad i,j = 1, \ldots, n, \\
&\cdots \\
\mathsf{P}(B_1 \cap B_2 \cap \cdots \cap B_{n-1}) &= \frac{1}{n^r}, \\
\mathsf{P}(B_1 \cap \cdots \cap B_n) &= 0.
\end{aligned}$$

Použijeme-li (1.15) z věty 1.1, dostaneme postupně

$$\begin{aligned}
q &= \sum_{i=1}^{n} \mathsf{P}(B_i) - \sum\sum_{i<j} \mathsf{P}(B_i \cap B_j) \\
&\quad + \sum\sum\sum_{i<j<k} \mathsf{P}(B_i \cap B_j \cap B_k) - \cdots \\
&= \binom{n}{1}\left(1-\frac{1}{n}\right)^r - \binom{n}{2}\left(1-\frac{2}{n}\right)^r + \binom{n}{3}\left(1-\frac{3}{n}\right)^r - \cdots \\
&\quad \cdots (-1)^n \binom{n}{n-1}\left(1-\frac{n-1}{n}\right)^r.
\end{aligned}$$

Podle [16] popisuje Maxwellův-Boltzmannův model chování soustav molekul plynů.

3.4 Boseův-Einsteinův model

Opět uvažujme r částic, z nichž každá je právě v jedné z n přihrádek. Předpokládáme, že

1. Částice jsou *nerozlišitelné*.
2. Pro každou přihrádku známe počet částic, které jsou v ní umístěny.
3. Stav systému je určen počty částic v jednotlivých přihrádkách.
4. Všechny stavy jsou stejně pravděpodobné.

Opět můžeme použít klasický pravděpodobnostní prostor, tentokrát se stejně pravděpodobnými elementárními jevy

$$\omega = (r_1, \ldots, r_n),$$

kde každé z čísel $r_j, j = 1, \ldots, n$, udávající počet číslic v j-té přihrádce, může nabývat nezáporné celočíselné hodnoty tak, že platí $\sum_{i=1}^{n} r_i = r$. Ekvivalentně lze Boseovo-Einsteinovo rozmístění popsat jako neklesající posloupnost délky r prvků množiny $\{1, 2, \ldots, n\}$, která je elementárnímu jevu $\omega = (r_1, \ldots, r_n)$ jednoznačně přiřazena následovně:

$$\underbrace{1\,1\ldots1}_{r_1-\text{krát}} \quad \underbrace{2\,2\ldots2}_{r_2-\text{krát}} \quad \cdots \quad \underbrace{n\,n\ldots n}_{r_n-\text{krát}}.$$

Je tedy

$$|\Omega| = \binom{r+n-1}{r} = \binom{r+n-1}{n-1}.$$

Povšimněme si, že Boseova-Einsteinova rozmístění tvoří stejnou množinu elementárních jevů jako má pokus, který jsme nazvali neuspořádaný výběr s vracením. Rozdíl je však v tom, že zatímco stochastická logika přiřadila neuspořádaným výběrům nutně nestejné pravděpodobnosti $p_{r_1,\ldots,r_n}$, vynucují si předpoklady 1–4 v tomto případě klasický pravděpodobnostní model.

Stejně jako u Maxwellovy-Boltzmannovy statistiky nyní určíme pravděpodobnost toho, že daná přihrádka obsahuje právě k částic. Tento jev nastane, když zbývajících $r-k$ částic rozmisťujeme do zbývajících $n-1$ přihrádek, takže je

$$p_{k,n,r} = \mathsf{P}[K_1 = k] = \binom{r-k+n-2}{r-k} \Big/ \binom{r+n-1}{r}.$$

Za stejných podmínek jako u Maxwellova-Boltzmannova modelu vyšetříme limitní chování této pravděpodobnosti. Dostaneme tak vztahy

$$\begin{aligned} p_k &= \lim_{n\to\infty} p_{k,n,r} \\ &= \lim_{n\to\infty} \frac{(r_n-k+n-2)!(n-1)!r_n!}{(r_n-k)!(n-2)!(r_n+n-1)!} \\ &= \lim_{n\to\infty} \frac{(n-1)r_n(r_n-1)\cdots(r_n-k+1)}{(r_n+n-1)\cdots(r_n+n-k-1)} \\ &= \frac{\lambda^k}{(1+\lambda)^{k+1}}. \end{aligned}$$

Také tentokrát můžeme zavést prostor elementárních jevů $\Omega_G = \{0,1,\ldots\}$ s pravděpodobnostmi p_k těchto elementárních jevů. Jak lze snadno ověřit, platí $\sum_{k=0}^{\infty} p_k = 1$.

k	$p_{k,n,r}$
0	0,4220
1	0,2440
2	0,1400
3	0,0815
4	0,0471
5	0,0272
6	0,0157

Tabulka 3.3: Aproximace pravděpodobností pro Boseovu-Einsteinovu statistiku

Smysl této limitní úvahy opět spočívá v možnosti aproximovat pravděpodobnost $\mathsf{P}[K_1 = k] = p_{k,n,r}$ pravděpodobností $p_k = \lambda^k/(1+\lambda)^{k+1}$ pro velká r a n, $r/n = \lambda$. Pomocí uvedené aproximace spočítáme pravděpodobnosti $p_{k,n,r}$ pro $n = 365$, $r = 500$ a $k = 0,1,\ldots,6$, $((1+\lambda)^{-1} = 0{,}4220)$. Porovnejte tabulku 3.3 s pravděpodobnostmi $p_{k,n,r}$ uvedenými v tabulce 3.2.

I v tomto modelu můžeme počítat pravděpodobnost toho, že v nějaké přihrádce není žádná částice. Předpokládejme nejprve, že je $r \geq n$. K výpočtu použijeme velmi častý obrat – přejdeme k jevu opačnému, kdy je v každé přihrádce aspoň jedna částice. Pak vlastně rozdělujeme jen zbývajících $r-n$ částic:

$$\begin{aligned} q &= 1 - \mathsf{P}[\text{každá přihrádka je neprázdná}] \\ &= 1 - \frac{\binom{(r-n)+n-1}{n-1}}{\binom{r+n-1}{n-1}}. \end{aligned}$$

Pro $r < n$ musí být samozřejmě některá z přihrádek prázdná, je tedy pak $q = 1$.

Podle [6] Boseova-Einsteinova statistika dobře popisuje rozdělení fotonů.

3.5 Fermiův-Diracův model

Tento model dostaneme z Boseova-Einsteinova modelu, použijeme-li Pauliho princip, podle kterého může být v každé přihrádce nejvýše jedna částice. Předpokládáme tedy, že platí

1. Částice nejsou *rozlišitelné*.
2. V každé přihrádce je nejvýše jedna částice.
3. Stav systému je určen (neuspořádaným) seznamem přihrádek, které jsou obsazeny částicemi.
4. Všechny stavy jsou stejně pravděpodobné.

Opět použijeme klasický pravděpodobnostní prostor, tentokrát s elementárními jevy

$$\omega = (t_1,\ldots,t_r),$$

kde $t_1,\ldots,t_r$ je uspořádaný vektor různých celých čísel z intervalu $\langle 1,n)$. Jsou to vlastně „adresy" obsazených přihrádek. Pro $r \leq n$ je tedy (počet r-prvkových kombinací z n prvků)

$$\Omega = \binom{n}{r}.$$

3.6 Pólyovo urnové schéma

Pólyovo urnové schéma je určeno parametry $a,b,n \in \mathbb{N}$ a celočíselným parametrem Δ. Mějme urnu s a černými a b bílými koulemi, $a,b \in \mathbb{N}$. Z urny opakovaně n-krát táhneme po jedné kouli. Zjistíme barvu vytažené koule a kouli do urny vrátíme. Je-li $\Delta > 0$, přidáme do urny dalších Δ koulí stejné barvy, jako byla vytažena. Je-li $\Delta < 0$, pak $-\Delta$ koulí této barvy z urny ubereme (jsou-li tam ještě). Před každým tahem koule dostatečně promícháme tak, aby každá koule z urny měla stejnou pravděpodobnost, že bude vytažena.

Při zavádění prostoru elementárních jevů je vhodné rozlišovat mezi jednotlivými koulemi. Proto je (předpokládáme $a+b+(n-1)\Delta > 0$)

$$|\Omega| = (a+b)(a+b+\Delta)\cdots(a+b+(n-1)\Delta).$$

Abychom zjednodušili zápis, zaveďme pro reálné x a nezáporné celé k funkci

$$\begin{aligned} x^{[k]} &= x(x+\Delta)\cdots(x+(k-1)\Delta) && \text{pro } x+(k-1)\Delta > 0, k > 0 \\ &= 0 && \text{jinak}, k > 0 \\ x^{[0]} &= 1. \end{aligned}$$

Speciálně platí

$$x^{[n]} = x^n \qquad \text{pro } \Delta = 0,$$
$$x^{[n]} = \frac{x!}{(x-n)!} \qquad \text{pro } \Delta = -1.$$

Zaveďme náhodný jev

$$D_k = [\text{v } n \text{ tazích bylo taženo právě } k \text{ bílých koulí}], \qquad k = 0, 1, \ldots, n.$$

Uvážíme-li zejména počet způsobů, kolikrát lze umístit k koulí po jedné do n tahů, dostaneme pro $k = 0, 1, \ldots, n$

$$\begin{aligned} |D_k| &= \binom{n}{k} b(b+\Delta)\cdots(b+(k-1)\Delta)a(a+\Delta)\cdots(a+(n-k-1)\Delta) \\ &= \binom{n}{k} b^{[k]}a^{[n-k]}. \end{aligned}$$

Protože náhodné jevy $D_0, D_1, \ldots, D_n$ tvoří úplný systém jevů, musí platit

$$1 = \sum_{k=0}^{n} \mathsf{P}(D_k) = \sum_{k=0}^{n} \binom{n}{k} \frac{b^{[k]}a^{[n-k]}}{(a+b)^{[n]}}.$$

Odtud plyne **zobecněný binomický vzorec**

$$(b+a)^{[n]} = \sum_{k=0}^{n} \binom{n}{k} b^{[k]}a^{[n-k]}. \tag{3.3}$$

Obecné urnové schéma má zajímavou vlastnost. Označme

$$B_i = [\text{v } i\text{-tém tahu byla tažena bílá koule}].$$

Bez ohledu na hodnotu Δ platí následující tvrzení.

Věta 3.1. Je-li $a + b + (n-1)\Delta > 0$, pak platí

$$\mathsf{P}(B_i) = \frac{b}{b+a}, \qquad i = 1, 2, \ldots, n. \tag{3.4}$$

D ů k a z: Nejprve vyjádříme B_i jako sjednocení neslučitelných jevů. Označme jako B_{ij} náhodný jev „v i-tém tahu byla tažena bílá koule, v ostatních $n-1$ tazích bylo taženo celkem j bílých koulí". Zřejmě tedy platí pro $j = 0, 1, \ldots, n-1$

$$|B_{ij}| = \binom{n-1}{j} b^{[j+1]}a^{[n-1-j]}.$$

Protože náhodné jevy $B_{i0}, B_{i1}, \ldots, B_{i,n-1}$ jsou neslučitelné, platí

$$\begin{aligned} \mathsf{P}(B_i) &= \sum_{j=0}^{n-1} \binom{n-1}{j} \frac{b^{[j+1]}a^{[n-1-j]}}{(b+a)^{[n]}} \\ &= \frac{b}{b+a} \sum_{j=0}^{n-1} \frac{\binom{n-1}{j}(b+\Delta)^{[j]}a^{[n-1-j]}}{(b+a+\Delta)^{[n-1]}} = \frac{b}{b+a}. \quad \square \end{aligned}$$

Uveďme dva speciální případy Pólyova schématu.

$\Delta = 0$ **(Bernoulliovo schéma)** Pro tento případ se někdy používá označení **výběr s vracením**. Dosadíme-li za funkci $x^{[k]}$ obyčejnou k-tou mocninu, dostaneme pro $k = 0, 1, \ldots, n$

$$\begin{aligned} \mathsf{P}(D_k) &= \binom{n}{k} \frac{b^k a^{n-k}}{(b+a)^n} \\ &= \binom{n}{k} p^k (1-p)^{n-k}, \end{aligned} \tag{3.5}$$

když jsme zavedli nový parametr

$$p = \frac{b}{b+a}.$$

Pokud nás zajímá pouze počet bílých koulí v n výběrech, můžeme použít prostor elementárních jevů $\Omega_G = \{0, 1, \ldots, n\}$ s pravděpodobnostmi $p_k = \mathsf{P}(D_k)$ podle (3.5).

$\Delta = -1$ **(Pearsonovo schéma)** Tento model se nazývá také **výběr bez vracení**. Právě k bílých koulí vytáhneme v n tazích s pravděpodobností

$$\mathsf{P}(D_k) = \binom{n}{k} \frac{b(b-1)\cdots(b-k+1)a(a-1)\cdots(a-(n-k)+1)}{(a+b)(a+b-1)\cdots(a+b-n+1)}$$

3.7 Náhodná procházka

Uvažme částici, která se pohybuje po celočíselné přímce $\mathbb{Z}$ a označme jako S_k její polohu v časech $k = 0, 1, \ldots$. Předpokládáme, že na počátku je částice v bodě 0, tj. $S_0 = 0$ a že, je-li v některém časovém okamžiku k v bodě

a, pak ji v čase $k+1$ nalezneme s pravděpodobností $1/2$ v bodě $a+1$ a s pravděpodobností $1/2$ v bodě $a-1$, tj.

$$\mathsf{P}[S_{k+1} - S_k = 1] = \mathsf{P}[S_{k+1} - S_k = -1] = \frac{1}{2}.$$

Dále předpokládáme, že rozhodování o směru pohybu nezávisí na pohybech v časech $0, 1, \ldots, k$. Omezíme-li se na zkoumání pohybu $S_0, S_1, \ldots, S_n$ v konečném časovém úseku, je zřejmé, že vhodným modelem je klasický pravděpodobnostní prostor s množinou elementárních jevů $\Omega = \{0,1\}^n$, tj. n-členných posloupností $(y_1, \ldots, y_n)$ nul a jedniček. Hodnota $y_j = 1$ ($y_j = 0$) vysílá částici v čase $j-1$ o jednotku doprava (o jednotku doleva). Ekvivalentním modelem je klasický pravděpodobnostní prostor s množinou elementárních jevů

$$\Omega = \{(0, s_1, s_2, \ldots, s_n), |s_{j+1} - s_j| = 1\},$$

která je seznamem všech možných trajektorií pohybu naší částice.

Nyní určíme rozdělení pravděpodobností náhodné veličiny S_n, která označuje finální polohu částice. Tato finální poloha je určena jednoznačně počtem $0 \le k \le n$ pohybů doprava jako $k - (n-k) = 2k - n$. Je tedy

$$\mathsf{P}[S_n = 2k - n] = \binom{n}{k} 2^{-n} \quad \text{pro} \quad 0 \le k \le n,$$

protože náhodný jev $[S_n = 2k - n]$ je tvořen právě těmi posloupnostmi $(y_1, \ldots, y_n)$, které obsahují k jedniček.

Pomocí binomické věty se snadno přesvědčíme, že součet těchto pravděpodobností je roven jedné a uvážíme-li symetrii

$$\mathsf{P}[S_n = 2k - n] = \binom{n}{k} 2^{-n} = \binom{n}{n-k} 2^{-n} = \mathsf{P}[S_n = n - 2k], \tag{3.6}$$

zjistíme, že střední hodnota náhodné veličiny S_n je rovna nule. Pro sudé časové okamžiky, řekněme $2n$, je nejpravděpodobnější hodnotou náhodné veličiny S_{2n} nula. Použijeme-li Stirlingův vzorec $n! = n^n\sqrt{2\pi n}\mathrm{e}^{-n}(1+\epsilon_n)$, kde $\lim_{n\to\infty} \epsilon_n = 0$, obdržíme

$$\mathsf{P}[S_{2n} = 0] = \frac{1}{\sqrt{\pi n}}(1 + \delta_n), \tag{3.7}$$

kde $\lim_{n\to\infty} \delta_n = 0$, tj.

$$\mathsf{P}[S_{2n} = 0] \doteq \frac{1}{\sqrt{\pi n}}.$$

Připomeňme tabulku 1.5 b), která ukazuje přesnost této aproximace. Zejména si všimněme, že ať již zákonem velkých čísel rozumíme cokoliv, *nemůže* to být tvrzení: při opakovaném házení symetrickou mincí je pravděpodobnost toho, že rub se objevil v polovině případů, v limitě rovna jedné. *Právě naopak*, tato limita je rovna nule.

Postavme naší částici do cesty bariéru umístěnou v některém celočíselném bodě $1 \le a \le n$. Jaká je pravděpodobnost toho, že částice projde v některém okamžiku $k = 1, 2, \ldots, n$ touto bariérou?

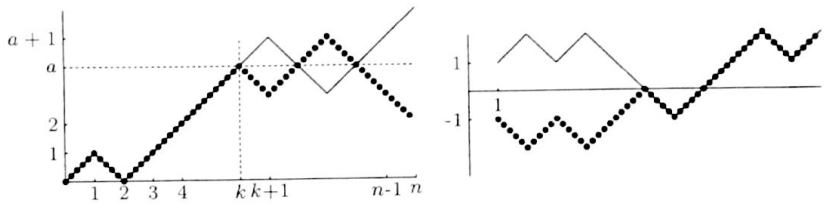
Chceme spočítat pravděpodobnost

$$P_{n,a} = \mathsf{P}[\max_{1\le k\le n} S_k \ge a].$$

Snadno dostaneme rovnost $P_{n,a} = \mathsf{P}(A_1) + \mathsf{P}(A_2) + \mathsf{P}(A_3)$, kde

$$\begin{aligned} A_1 &= [\max_{1\le k\le n} S_k \ge a, S_n > a] = [S_n > a], \\ A_2 &= [\max_{1\le k\le n} S_k \ge a, S_n < a], \\ A_3 &= [\max_{1\le k\le n} S_k \ge a, S_n = a] = [S_n = a]. \end{aligned}$$

(Vyšetřovaný náhodný jev jsme rozdělili do tří částí podle konečné polohy částice.)



a) b)

Obrázek 3.1: Dvojice trajektorií z množin A_1 a A_2

Platí ovšem $\mathsf{P}(A_1) = \mathsf{P}(A_2)$, protože počet trajektorií, které přispívají vlastnosti A_1 je roven počtu trajektorií, které přispívají vlastnosti A_2. Stačí zkonstruovat prosté zobrazení množiny A_1 na množinu A_2. Obrázek 3.1 a) toto zobrazení dobře definuje. Trajektorie kreslená souvislou čarou má vlastnosti A_1, k je první časový okamžik, kdy částice vstoupí do bariéry a. Trajektorie kreslená silnější tečkovanou čarou (do okamžiku k splývající se „souvislou" trajektorií a od okamžiku $k+1$ její zrcadlový obraz podle osy $y = a$)

má vlastnost A_2. Zobrazení, které přiřazuje „souvislé" trajektorii trajektorii „tečkovanou", je prosté zobrazení A_1 na A_2. Obdrželi jsme tedy rovnost

$$(3.8)\quad P_{n,a} = 2\,\mathsf{P}[S_n > a] + \mathsf{P}[S_n = a] = 2\,\mathsf{P}[S_n \geq a] - \mathsf{P}[S_n = a] \to 1$$

při $n \to \infty$, kde limitní přechod je důsledkem (3.7), protože

$$\mathsf{P}[S_{2n} = x] \leq \mathsf{P}[S_{2n} = 0] = \frac{1}{\sqrt{\pi n}}(1 + \delta_n)$$

pro $-2n \leq x \leq 2n$. V Kolmogorovově axiomatické teorii pravděpodobnosti lze vlastnost (3.8) interpretovat jako tvrzení:

(3.9) Částice konající náhodnou procházku jistě (tj. s pravděpodobností 1) vstoupí do každé bariéry $a \in \mathbb{Z}$, $a \neq 0$.

Studujme závěrem chování naší náhodné procházky podrobněji. Označme jako n_0^* první časový okamžik, kdy se částice navrátí do počátku celočíselné osy, pokud k návratu vůbec nikdy nedojde, pak volíme $n_0^* = \infty$. Implicitně předpokládáme nekonečný život částice. Určíme rozdělení pravděpodobností náhodného času návratu n_0^*, tj. pravděpodobnosti $p_{2n} = \mathsf{P}[n_0^* = 2n]$ (veličina n_0^* nemůže nabývat lichých hodnot): Pro $x \in \mathbb{N}$ a $y \in \mathbb{Z}$ označíme jako $N_{x,y}$ počet trajektorií částice, které spojují bod $(0,0)$ s bodem (x,y), tj. trajektorií, při kterých je částice v čase x v poloze y. Povšimněme si, že je-li $N_{x,y} \neq 0$, pak $x = p + q$, $y = p - q$, kde $p(q)$ je počet kroků, které částice učinila doprava (doleva) při svém životě v časovém intervalu $\langle 0, x\rangle$. Odtud $N_{x,y} = \binom{p+q}{p} = \binom{x}{\frac{x+y}{2}}$. Pro $x > 0, y > 0$ označme ještě jako $N^+_{x,y}$ počet trajektorií spojujících bod $(0,0)$ a (x,y) takových, že $S_1 > 0$, $S_2 > 0$, ..., $S_x = y > 0$, tj. počet trajektorií, které spojují body (0,0) a (x,y) cestou nad osou x. Platí

Věta 3.2. (Bertranova) Pro $x > 0, y > 0$ platí

$$N^+_{x,y} = \frac{y}{x} N_{x,y}.$$

D ů k a z: Důkaz stačí provést pro situaci $N_{x,y} \neq 0$: Postupně vyvozujeme:
$N^+_{x,y}$ = [počet trajektorií spojujících body (1,1) a (x,y), které neprotínají osu x]
= [počet všech trajektorií spojujících (1,1) a (x,y)]
- [počet trajektorií, které protínají osu x a spojují body (1,1) s (x,y)]
= $N_{x-1,y-1}$ − [počet všech trajektorií, které spojují body (1,-1) s (x,y)]
= $N_{x-1,y-1} - N_{x-1,y+1}$.

Je třeba podrobněji zdůvodnit, že $|A| = |B|$, kde
A = [trajektorie, které spojují body (1,1) a (x,y) a protínají osu x],
B = [trajektorie, které spojují body (1,-1) a (x,y)],
tj. zkonstruovat prosté zobrazení množiny A na množinu B. Na obrázku 3.1 b) je „souvislé" trajektorii z množiny A přiřazena „tečkovaná" trajektorie z množiny B.

Dokončíme nyní důkaz. Platí

$$N^+_{x,y} = N_{x-1,y-1} - N_{x-1,y+1} = \binom{x-1}{\frac{x+y-2}{2}} - \binom{x-1}{\frac{x+y}{2}}$$

$$= \binom{p+q-1}{p-1} - \binom{p+q-1}{q-1} = \binom{p+q}{p}\left(\frac{p}{p+q} - \frac{q}{p+q}\right)$$

$$= N_{x,y}\frac{p-q}{p+q} = N_{x,y}\frac{y}{x},$$

kde $x = p+q$ a $y = p-q$. Nyní již určíme pravděpodobnost $p_{2n} = \mathsf{P}[n_0^* = 2n]$ snadno: ze symetrie našeho pohybu a věty 3.2 plyne:

$$\begin{aligned} p_{2n} &= \mathsf{P}[S_1 = 1, S_2 > 0, \ldots, S_{2n-1} = 1, S_{2n} = 0] \\ &\quad + \mathsf{P}[S_1 = -1, S_2 < 0, \ldots, S_{2n-1} = -1, S_{2n} = 0] \\ (3.10)\qquad &= 2\,\frac{1}{2}\,\mathsf{P}[S_1 = 1, S_2 > 0, \ldots, S_{2n-2} > 0, S_{2n-1} = 1] \\ &= N^+_{2n-1,1}\,2^{-(2n-1)} = \frac{1}{2n-1}N_{2n-1,1}\,2^{-(2n-1)} \\ &= \frac{1}{2n-1}\binom{2n-1}{n}2^{-(2n-1)} = \frac{1}{2n}\binom{2n-2}{n-1}2^{-2n} \\ &= \frac{1}{2n}\mathsf{P}[S_{2n-2} = 0]. \end{aligned}$$

Všimněme si ještě, že také platí

$$\begin{aligned} p_{2n} &= \binom{2n-2}{n-1}2^{-(2n-2)} - \binom{2n}{n}2^{-2n} \\ &= \mathsf{P}[S_{2n-2} = 0] - \mathsf{P}[S_{2n} = 0]. \end{aligned}$$

Odtud

$$\begin{aligned} 1 - \sum_{k=1}^{n} p_{2k} &= 1 - (1 - \mathsf{P}[S_2 = 0]) - (\mathsf{P}[S_2 = 0] - \mathsf{P}[S_4 = 0]) \\ &\quad - \ldots - (\mathsf{P}[S_{2n-2} = 0] - \mathsf{P}[S_{2n} = 0]) \\ &= \mathsf{P}[S_{2n} = 0] \to 0 \quad \text{při } n \to \infty \end{aligned}$$

podle (3.7), takže je $\sum_{n=1}^{\infty} p_{2n} = 1$. Musí tedy být $\mathsf{P}[n_0^* = +\infty] = 0$, tj. došli jsme k závěru, že:

(3.11) Částice konající náhodnou procházku se jistě v konečném čase vrátí do bodu 0, tudíž jistě (s pravděpodobností 1) navštíví bod 0 nekonečněkrát.

Kombinací výroků (3.9) a (3.10) dostaneme:

(3.12) Částice konající náhodnou procházku jistě (s pravděpodobností 1) navštíví každý bod $a \in \mathbb{Z}$ nekonečněkrát.

Tvrzení (3.11) lze v Kolmogorovově axiomatice dokázat jako matematickou větu. V kontrastu s tvrzením (3.11) je následující výpočet střední hodnoty okamžiku prvního návratu n_0^*:

$$\begin{aligned} \mathsf{E}n_0^* &= \sum_{n=1}^{\infty} 2n\mathsf{P}[n_0^* = 2n] = \sum_{n=1}^{\infty} 2n\frac{1}{2n}\mathsf{P}[s_{2n-2} = 0] \\ &= \sum_{n=1}^{\infty} \binom{2n-2}{n-1} 2^{-(2n-2)} = \sum_{n=2}^{\infty} \frac{1}{\sqrt{\pi n}}(1+\delta_n) = +\infty, \end{aligned}$$

kde jsme postupně využili vzorce (3.10) a (3.7). K návratu částice do nuly sice jistě dojde, ale střední doba čekání na tuto událost je nekonečná! □

3.8 Geometrická pravděpodobnost

Začněme motivačním příkladem.

Příklad 3.1. Alena a Bohouš si smluvili schůzku na přesně určeném místě, ale v poněkud neurčitém čase. Mají se sejít ve zvolený den někdy mezi polednem a jednou hodinou odpolední, přičemž každý z nich je ochoten čekat celých dvacet minut, ovšem pouze během oné vymezené hodiny. Jaká je pravděpodobnost, že se opravdu sejdou, když předpokládáme, že každý z nich může přijít kdykoliv během udané hodiny a že přicházejí nezávisle na sobě?

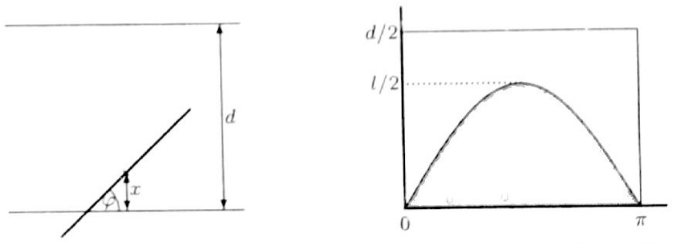
Úlohu si znázorníme geometricky. Označíme-li jako x okamžik příchodu Aleny a jako y okamžik příchodu Bohouše, můžeme se omezit pouze na dvojice (x, y), které leží v intervalu $\Omega = \{(x, y) : 0 \le x \le 1,\ 0 \le y \le 1\}$. Okamžiky příchodu tedy můžeme znázornit jako jakýkoliv bod jednotkového čtverce. Poslední větu zadání úlohy interpretujeme tak, že do nějaké části jednotkového čtverce padne bod (x, y) s pravděpodobností, která závisí na velikosti této části a nezávisí na jejím tvaru či umístění ve čtverci. Tento požadavek lze zapsat jako

(3.13) $$\mathsf{P}(A) = \frac{\mu(A)}{\mu(\Omega)}.$$

Samotné řešení úlohy je snadné. Náhodný jev A je dán nerovnostmi $0 \le x \le 1$, $0 \le y \le 1$, $|x - y| \le 1/3$, takže podle (3.13) dostaneme

$$\mathsf{P}(A) = 1 - (2/3)^2 = 5/9. \qquad \bigcirc$$

Podle (3.13) počítáme pravděpodobnost, když můžeme jako prostor elementárních jevů zvolit vhodný geometrický útvar, jehož velikost (plochu, objem) dokážeme určit. Jako jevy pak chápeme podmnožiny tohoto útvaru se stejnými vlastnostmi. Teorie míry zavádí pojem měřitelné množiny. Například na reálné přímce jsou měřitelné všechny prvky borelovské σ-algebry, která je nejmenší σ-algebrou nad intervaly tvaru $(-\infty, x)$, $x \in \mathbb{R}$. V úloze musí být splněn předpoklad symetrie, zaručující, že pravděpodobnost toho, že náhodně zvolený bod z Ω padne do množiny A, závisí pouze na velikosti této množiny a nikoliv na jejím umístění (tvaru atd.) v Ω.



a) Rovnoběžky s jehlou b) Geometrická pravděpodobnost

Obrázek 3.2: Buffonova úloha

Příklad 3.2. (Buffonova úloha) Uvažujme rovinu, v níž jsou v pravidelných vzdálenostech d umístěny rovnoběžky. Na tuto rovinu náhodně vrháme jehlu, jejíž délka l splňuje podmínku $l < d$. Jaká je pravděpodobnost, že tato jehla protne některou z rovnoběžek?

Označme jako x vzdálenost středu jehly od nejbližší rovnoběžky a jako φ úhel, který svírá jehla s touto přímkou (viz obrázek 3.2 a)). Jako geometrický objekt Ω zvolíme interval $\langle 0, d/2\rangle \times \langle 0, \pi)$. Množina A znamenající, že

jehla protnula některou přímku (více než jednu protnout nemůže), je dána požadavky (obrázek 3.2 b))

$$(l/2)\sin\varphi < x, \quad (x,\varphi) \in \Omega.$$

Podle (3.13) dostaneme

$$\mathsf{P}(A) = \frac{1}{\pi(d/2)} \int_0^\pi \frac{l}{2}\sin\varphi \mathrm{d}\varphi = \frac{2l}{\pi d}. \quad \bigcirc$$

Příklad 3.3. Stejně jako v předcházejícím příkladu mějme rovnoběžky ve vzdálenosti d rovnoměrně pokrývající rovinu a vrhejme na tuto rovinu souvislou konvexní množinu s obvodem délky a, jejíž průměr (maximální vzdálenost dvou bodů) je menší než d. Opět chceme určit pravděpodobnost, s jakou hranice této množiny protne některou z rovnoběžek.

Řešení začneme od speciálního případu, kdy množinu tvoří konvexní n-úhelník. Označme délky stran tohoto n-úhelníka jako $l_i, i = 1, \ldots, n$. Jev A, který znamená, že mnohoúhelník protne úsečku, lze vyjádřit jako sjednocení disjunktních množin $A = \cup_{i=1}^{n-1} \cup_{j=i+1}^{n} A_{ij}$, kde A_{ij} znamená, že rovnoběžka protnula právě i-tou a j-tou stranu. Vzhledem k tomu, že jde o *disjunktní* sjednocení, platí

$$\mathsf{P}(A) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathsf{P}(A_{ij}).$$

Použijeme-li toho, že je $\mathsf{P}(A_{ij}) = \mathsf{P}(A_{ji})$ a dále $\mathsf{P}(A_{ii}) = 0$, dostaneme dále

$$\mathsf{P}(A) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathsf{P}(A_{ij}).$$

S použitím výsledku předchozího příkladu dostaneme pro každé $i = 1, \ldots, n$, že platí (strana n-úhelníka je onou jehlou)

$$\sum_{j=1}^{n} \mathsf{P}(A_{ij}) = \frac{2l_i}{\pi d},$$

takže nakonec máme

$$\mathsf{P}(A) = \frac{1}{2} \sum_{i=1}^{n} \frac{2l_i}{\pi d} = \frac{a}{\pi d}.$$

Protože výsledná pravděpodobnost nezávisí na délkách jednotlivých stran n-úhelníka, ale pouze na jeho obvodu a, musí stejný vztah platit pro jakoukoliv konvexní množinu, neboť její hranici můžeme chápat jako limitu posloupnosti aproximujících mnohoúhelníků. $\bigcirc$

3.9 Cvičení

3.1. Máme šest druhů vína různých značek. Přiřadíme-li vínům etikety náhodně, jaká je pravděpodobnost, že
a) se „trefíme",
b) čtyři druhy označíme správně a dva nesprávně,
c) aspoň jeden druh označíme správně?

3.2. Při výrobě lahví se v roztavené sklovině mohou objevit malá tělíska (kamínky). Předpokládejme, že jedna láhev váží jeden kilogram a že zpracováváme celkem m kilogramů skloviny, v níž je λm kilogramů kamínků. Jaká je pravděpodobnost, že náhodně vybraná láhev neobsahuje kamínek, je-li $\lambda = 0{,}3$ nebo $\lambda = 0{,}1$? Jak se tyto pravděpodobnosti změní, když budeme vyrábět láhve pouze půlkilové? ([16, str. 114])

3.3. Máme n různých dopisů a n různých obálek. Dopisy byly do obálek umístěny náhodně. Jaká je pravděpodobnost, že aspoň jeden dopis se dostane do správné obálky?

3.4. Dva parníky, které používají jako jediné stejné přístaviště, mohou připlout kdykoliv během 24 hodin, jejich příjezdy jsou nezávislé. První parník obsadí přístaviště na jednu hodinu, druhý na dvě hodiny. Jaká je pravděpodobnost, že ani jeden parník nebude muset čekat na uvolnění přístaviště?

3.5. Osoby X a Y přijdou na smluvené místo kdykoliv mezi 12.00 a 13.00. Určete pravděpodobnost, že Y přijde až po X, jestliže přijde až po 12.30. Předpokládá se nezávislé chování obou osob, přičemž okamžiky příchodu jsou stejně možné kdykoliv během uvedené hodiny.

3.6. V rovině jsou narýsovány rovnoběžky vzdálené od sebe střídavě 15 a 80 milimetrů. Na tuto rovinu je náhodně vržen kruh o poloměru 25 milimetrů. Určete pravděpodobnost toho, že po dopadu nebude protínat žádnou z rovnoběžek.

3.7. Před kotoučem, který se otáčí konstantní rychlostí, je v rovině tohoto kotouče umístěna úsečka délky $2h$ tak, že přímka spojující střed kotouče se středem úsečky je na tuto úsečku kolmá. Po tečně ke kružnici odlétne v náhodný okamžik částice. Úsečka je od středu kotouče ve vzdálenosti l. Určete pravděpodobnost, že částice zasáhne úsečku.

3.8. Nechť $x, y \in (0,1)$ jsou náhodně zvolená čísla. Jaká je pravděpodobnost, že jejich součet je menší než 1 a součin menší než 0,09?

3.9. Na úsečce délky l jsou náhodně umístěny dva body, kterými je náhodně rozdělena na tři části. S jakou pravděpodobností lze z takto vzniklých tří úseček sestrojit trojúhelník?

3.10. Tyč dlouhá 200 mm je náhodně rozřezána na tři části. S jakou pravděpodobností je některá z těchto částí kratší než 10mm, jestliže dva řezy jsou stejně možné v každém místě tyče?

3.11. Pravoúhlá mříž je složena z válcových prutů o poloměru r. Vzdálenosti mezi osami prutů jsou rovny a a b. Kuličku o průměru d hodíme bez míření po dráze, která je kolmá k rovině mříže. Určete pravděpodobnost toho, že kulička zasáhne mříž.

4. Náhodná veličina

Má-li být naše teorie prakticky užitečná, musí se přizpůsobit tomu, že výsledkem pokusu nebývá ani tak zjištění, zda náhodný jev nastal či nenastal, ale mnohem častěji je výsledkem pokusu nějaké číslo: počet líců v posloupnosti hodů mincí; počet šestek, které padly na deseti hracích kostkách; doba, po kterou vydržela svítit zkoušená žárovka; počet bakterií v jednotkovém objemu vody; hmotnost pšenice z pokusného políčka atd. Vedle náhodného jevu je tedy třeba zabývat se náhodnou veličinou (též náhodnou proměnnou).

V našem modelu s pravděpodobnostním prostorem $(\Omega, \mathcal{A}, \mathsf{P})$ můžeme náhodnou veličinu X zavést jako reálnou funkci definovanou na náhodných jevech – podmnožinách z Ω, které jsou prvky $\mathcal{A}$. Podobně, jako jsme byli opatrní při definování pravděpodobnosti, kdy jsme do systému $\mathcal{A}$ zařadili jen některé podmnožiny Ω, totiž takové, kterým umíme přiřadit pravděpodobnost, měli bychom být opatrní i nyní. Je třeba zajistit možnost přiřazení (výpočtu) pravděpodobnosti všem rozumným množinám reálných čísel.

Od dvojice $(\Omega, \mathcal{A})$ potřebujeme přejít k podobné dvojici $(\mathbb{R}, \mathcal{B})$, kde $\mathbb{R}$ je reálná přímka, tak, aby podmnožinám $\mathbb{R}$, které jsou v σ-algebře $\mathcal{B}$, bylo možno přiřadit pravděpodobnost odvozenou z pravděpodobnostního prostoru $(\Omega, \mathcal{A}, \mathsf{P})$.

Při konstrukci σ-algebry na reálné přímce je vhodné zajistit, aby do $\mathcal{B}$ patřily všechny intervaly tvaru $\langle a, b\rangle$, $\langle a, b)$, $(a, b\rangle$, $(-\infty, b)$, $(-\infty, b\rangle$, (a, ∞), $\langle a, \infty)$. Nejmenší σ-algebra podmnožin $\mathbb{R}$, která obsahuje všechny intervaly zmíněných typů, se nazývá **borelovská σ-algebra**. K tomu, abychom uvedenou σ-algebru vytvořili, stačí vyjít pouze z intervalů tvaru $(-\infty, x)$. Například pro $a \le b$ jistě platí

$$\begin{aligned} \langle a, b\rangle &= (-\infty, b\rangle - (-\infty, a) \\ &= \bigcap_{n=1}^{\infty} (-\infty, b + \frac{1}{n}) - (-\infty, a), \end{aligned}$$

takže podle (1.22) a (1.24) je interval $\langle a, b\rangle$ nutně prvkem σ-algebry $\mathcal{B}$. Podobné tvrzení lze dokázat i pro ostatní typy intervalů.

Nechť X je reálná funkce definovaná na Ω, nechť a, b jsou libovolná reálná čísla resp. $-\infty, \infty$. Zaveďme užitečné označení pro některé podmnožiny Ω:

$$\begin{aligned} [X < b] &= \{\omega \in \Omega : X(\omega) < b\}, \\ [a < X < b] &= \{\omega \in \Omega : a < X(\omega) < b\}. \end{aligned}$$

Definice 4.1. Mějme pravděpodobnostní prostor $(\Omega, \mathcal{A}, \mathsf{P})$. Reálná funkce X definovaná na Ω, pro kterou platí

$$x \in \mathbb{R} \Rightarrow [X < x] \in \mathcal{A}, \tag{4.1}$$

se nazývá **náhodná veličina**.

Reálná funkce X splňující (4.1) se nazývá **měřitelná**, prvky σ-algebry $\mathcal{B}$ se nazývají **měřitelné množiny**.

Definice 4.1 zaručuje možnost určit pravděpodobnost, že náhodná veličina X nabude hodnoty z intervalu tvaru $(-\infty, x)$, $x \in \mathbb{R}$. Protože $\mathcal{B}$ je (nejmenší) σ-algebra nad těmito intervaly, ke každé množině $B \in \mathcal{B}$ existuje její vzor v $\mathcal{A}$:

$$B \in \mathcal{B} \Rightarrow X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{A}.$$

Proto pro každou množinu $B \in \mathcal{B}$ určuje náhodná veličina X pravděpodobnost $\mathsf{P}_X(B)$ náhodného jevu B pomocí vztahu

$$\begin{aligned} \mathsf{P}_X(B) &= \mathsf{P}(\{\omega \in \Omega : X(\omega) \in B\}) \\ &= \mathsf{P}(X^{-1}(B)). \end{aligned} \tag{4.2}$$

Definice 4.2. Množinová funkce

$$\mathsf{P}_X(B) = \mathsf{P}(X^{-1}(B)), \qquad B \in \mathcal{B}, \tag{4.3}$$

se nazývá **rozdělení pravděpodobnosti** náhodné veličiny X.

Pravděpodobnostní chování náhodné veličiny X je zřejmě určeno systémem pravděpodobností $\mathsf{P}[X < x]$, $x \in \mathbb{R}$. Symbolem $\mathsf{P}[X < x]$ jsme zkráceně zapsali $\mathsf{P}([X < x])$.

Definice 4.3. Nechť X je náhodná veličina definovaná na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, \mathsf{P})$. Reálná funkce

$$F_X(x) = \mathsf{P}[X < x]$$

se nazývá **distribuční funkce** náhodné veličiny X.

Nehrozí-li nedorozumění, dolní index se v označení distribuční funkce zpravidla vynechává.

Věta 4.1. Distribuční funkce je funkce
(a) neklesající,
(b) zleva spojitá,
(c) splňuje

$$\lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to \infty} F_X(x) = 1. \tag{4.4}$$

D ů k a z: Zvolme reálná čísla $a < b$. Zřejmě platí

$$[X < b] = [X < a] \cup [a \le X < b],$$

přičemž náhodné jevy (podmnožiny Ω) na pravé straně jsou neslučitelné. Proto podle (1.31) platí užitečné tvrzení

$$\mathsf{P}[a \le X < b] = F_X(b) - F_X(a), \tag{4.5}$$

jehož důsledkem je první dokazovaná vlastnost.

Pro $n \to \infty$ tvoří $(-\infty, x - \frac{1}{n})$ neklesající posloupnost intervalů. Proto je také

$$X^{-1}\left(-\infty, x - \frac{1}{n}\right) = \left\{\omega \in \Omega : X(\omega) < x - \frac{1}{n}\right\}$$

neklesající posloupnost množin z $\mathcal{A}$ s limitou $X^{-1}(-\infty, x)$. Podle věty 1.5 je pak také

$$\lim_{n \to \infty} F_X\left(x - \frac{1}{n}\right) = F_X(x),$$

což je právě tvrzení (b). Použijeme-li stejnou záměnu pořadí limity a integrálu (pravděpodobnosti), dostaneme jedno z tvrzení (c):

$$\lim_{x \to \infty} F_X(x) = \lim_{n \to \infty} \mathsf{P}[X < n] = \mathsf{P}\left(\bigcup_{n=1}^{\infty} [X < n]\right) = \mathsf{P}(\Omega) = 1.$$

Druhé tvrzení se dokáže podobně, jen pomocí nerostoucí posloupnosti náhodných jevů. □

Doporučuji, aby čtenář nepřehlédl důležitý vztah (4.5). Ještě se k němu vrátíme. Další obecné vlastnosti distribuční funkce uvádí následující tvrzení, v němž symbolem $F_X(x + 0)$ značíme limitu zprava funkce F_X v bodě x.

Věta 4.2. Pro distribuční funkci F_X platí

$$\mathsf{P}[X = x] = F_X(x + 0) - F_X(x), \qquad x \in \mathbb{R}, \tag{4.6}$$

D ů k a z: Vyjdeme z rozkladu náhodného jevu $[X \le x]$ na sjednocení dvou neslučitelných náhodných jevů $[X = x]$ a $[X < x]$. Přitom náhodný jev $[X \le x]$ lze zapsat jako limitu monotonní posloupnosti náhodných jevů $[X < x + \frac{1}{n}]$, takže je také podle věty 1.5

$$\mathsf{P}[X \le x] = \mathsf{P}\left(\bigcap_{j=1}^{\infty} \left[X < x + \frac{1}{n}\right]\right)$$

$$= \lim_{n\to\infty} \mathsf{P}\left[X < x + \frac{1}{n}\right]$$
$$= \lim_{n\to\infty} F_X\left(x + \frac{1}{n}\right)$$
$$= F_X(x+0).$$

Je tedy $\mathsf{P}[X = x] = \mathsf{P}[X \le x] - \mathsf{P}[X < x]$. □

Z dokázané věty plyne, že distribuční funkce je v bodě x spojitá, právě když je $\mathsf{P}[X = x] = 0$.

Věta 4.3. Distribuční funkce má nejvýše spočetně mnoho bodů nespojitosti (skoků).

D ů k a z: Označme jako C_n množinu bodů, v nichž má distribuční funkce F_X skok větší než je číslo $1/n$. Protože velikost skoku distribuční funkce v bodě x je dána pravděpodobností $\mathsf{P}[X = x]$, máme

$$C_n = \left\{x \in \mathbb{R} : \mathsf{P}[X = x] > \frac{1}{n}\right\}.$$

Protože hodnoty distribuční funkce leží v intervalu $\langle 0, 1\rangle$, může být v množině C_n nejvýše $n-1$ prvků. Množinu C bodů, v nichž má distribuční funkce F_X nějaký skok (tedy není spojitá), lze vyjádřit jako $C = \cup_{n=2}^{\infty} C_n$. Jako nejvýše spočetné sjednocení konečných množin je množina C nejvýše spočetná. □

Viděli jsme, že rozdělení náhodné veličiny lze určit pomocí distribuční funkce. Tvrzení věty 4.1 udává jakési minimální vlastnosti distribuční funkce. Lze dokázat, že každá neklesající zleva spojitá funkce G splňující

$$\lim_{x\to-\infty} G(x) = 0, \quad \lim_{x\to\infty} G(x) = 1$$

určuje nějakou náhodnou veličinu. V důkazu se zvolí pravděpodobnostní prostor $(\mathbb{R}, \mathcal{B}, \mathsf{P}_G)$ tak, aby pro každé $x \in \mathbb{R}$ bylo $\mathsf{P}(-\infty, x) = G(x)$. Náhodná veličina, jejíž existence se dokazuje, se zkonstruuje jako identické zobrazení.

V bohatém systému všech rozdělení jsou prakticky užitečné zejména dva speciální případy, tzv. spojité a diskrétní rozdělení.

4.1 Diskrétní rozdělení

Předpokládejme, že existují vesměs různá reálná čísla $x_1, x_2, \ldots$ taková, že je

$$\sum_{i=1}^{\infty} \mathsf{P}[X = x_i] = 1.$$

Seznam hodnot, kterých nabývá náhodná veličina s diskrétním rozdělením, a seznam pravděpodobností, s nimiž těchto hodnot náhodná veličina nabývá, udává diskrétní rozdělení pravděpodobností. Distribuční funkce je v tomto případě po částech konstantní se skoky právě v bodech $x_1, x_2, \ldots$, je tedy dána vztahem

$$F_X(x) = \sum_{i:x_i<x} \mathsf{P}[X = x_i].$$

Příklad 4.1. (Binomické rozdělení) Uvažujme n nezávislých pokusů, v každém může nastat *zdar* s pravděpodobností p, *nezdar* s pravděpodobností $1-p$. Můžeme zvolit $\Omega = \{0,1\}^n$. Elementární jev má pak tvar $\omega = (\omega_1, \ldots, \omega_n)$, kde ω_i je počet zdarů v i-tém pokusu. Binomické rozdělení má náhodná veličina

$$X(\omega) = \sum_{i=1}^{n} \omega_i,$$

tedy celkový počet zdarů v n pokusech. Pro každý z pokusů platí

$$\mathsf{P}(\omega_i) = p^{\omega_i}(1-p)^{1-\omega_i}.$$

Vzhledem k předpokládané nezávislosti pokusů dostaneme

$$\mathsf{P}(\omega) = \prod_{i=1}^{n} \mathsf{P}(\omega_i)$$
$$= p^{\sum\omega_i}(1-p)^{n-\sum\omega_i}.$$

Protože je celkem $\binom{n}{k}$ elementárních jevů, pro které je $\sum_{i=1}^{n}\omega_i = k$, dostáváme vztah

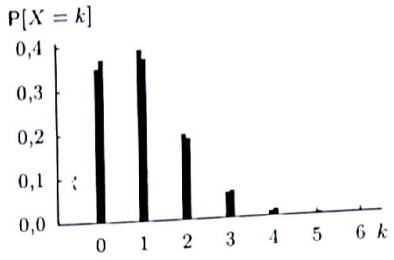
$$\mathsf{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n. \tag{4.7}$$

Vztah (4.7) udává rozdělení náhodné veličiny s binomickým rozdělením s parametry n, p, kde $n \in \mathbb{N}, 0 < p < 1$, což stručně zapisujeme jako $X \sim \mathsf{bi}(n,p)$. Pravděpodobnost binomického rozdělení jsme odvodili již dříve, v Bernoulliově schématu v odstavci 3.6. ○

V odstavci 3.3 věnovaném Maxwellovu-Boltzmannovu schématu s r rozlišitelnými předměty a n rozlišitelnými přihrádkami jsme odvodili, že rozdělení náhodné veličiny K_1, která udává počet předmětů v první přihrádce, je dáno pravděpodobnostmi

$$\mathsf{P}[K_1 = k] = \binom{r}{k}\frac{(n-1)^{r-k}}{n^r} = \binom{r}{k}\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^{r-k}$$

pro $k = 0, 1, \ldots, r$, tj. $K_1 \sim \mathsf{bi}(r, 1/n)$. Poté jsme se ujistili, že při *velkých* počtech přihrádek a předmětů v poměru $r/n \doteq \lambda \in (0, \infty)$ platí $\mathsf{P}[K_1 = k] \doteq \lambda^k \mathrm{e}^{-\lambda}/k!$ pro $k = 0, 1, \ldots$.



Obrázek 4.1: Porovnání pravděpodobností rozdělení bi(10, 0,1) vlevo a Po(1) vpravo

Řekneme, že náhodná veličina X má **Poissonovo rozdělení** s parametrem $\lambda > 0$, jestliže

$$\mathsf{P}[X = k] = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda} \text{ pro } k = 0, 1, \ldots,$$

což značíme $X \sim \mathsf{Po}(\lambda)$. Čtenář by se měl přesvědčit o platnosti vztahu

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda} = 1.$$

Podle následujícího tvrzení můžeme aproximovat při velkém počtu pokusů n a malé pravděpodobnosti p binomické rozdělení $\mathsf{bi}(n, p)$ pomocí Poissonova rozdělení $\mathsf{Po}(np)$.

Věta 4.4.(Poissonova) Nechť je $X_n \sim \mathsf{bi}(n, p_n)$, kde $\lim_{n\to\infty} np_n = \lambda \in (0, \infty)$ a $p_n \in (0, 1)$, nechť $X \sim \mathsf{Po}(\lambda)$. Potom platí

$$\lim_{n\to\infty} \mathsf{P}[X_n = k] = \mathsf{P}[X = k] \quad \text{pro } k = 0, 1, \ldots.$$

Rychlost konvergence ilustruje obrázek 4.1 a také tabulka 3.2.
D ů k a z: Pro $k = 0, 1, \ldots, n$ platí

$$\begin{aligned}\mathsf{P}[X_n = k] &= \binom{n}{k} p_n^k (1 - p_n)^{n-k} \\ &= \frac{1}{k!} \frac{np_n \cdot (n-1)p_n \cdots (n-k+1)p_n}{(1-p_n)^k} \left(1 - \frac{np_n}{n}\right)^n.\end{aligned}$$

Limita pravé strany při $n \to \infty$ je rovna $\mathsf{P}[x_n = k] = \lambda^k \mathrm{e}^{-\lambda}/k!$, protože $\lim_{n\to\infty} np_n = \lambda$, $\lim_{n\to\infty} p_n = 0$ a $(1 - x/n)^n$ je posloupnost funkcí, která konverguje k funkci e^{-x} stejnoměrně na každém omezeném intervalu. □

Příklad 4.2. Jako příklad situace, kdy se můžeme setkat s Poissonovým rozdělením, uveďme následující model. Pod mikroskopem sledujeme na Petriho misce výskyt nějakých elementů (např. krevních tělísek nebo bakterií). Pokusíme se nalézt jednoduchý zdůvodněný model pro pravděpodobnost, že

ve čtverci o jednotkové ploše se vyskytne právě k elementů. Rozdělme tento čtverec na n stejně velkých nepřekrývajících se částí o ploše $1/n$. Předpokládejme, že na takovéto malé plošce se právě jeden element vyskytne s pravděpodobností λ/n a více než jeden element s pravděpodobností, která je zanedbatelná. Předpokládejme navíc, že výskyty elementů v n malých částech jsou nezávislé. Potom má počet elementů v jednotkovém čtverci binomické rozdělení $\mathsf{bi}(n, \lambda/n)$. Parametr λ charakterizuje hustotu elementů, tedy jejich počet, průměrně připadající na jednotkovou plochu. ○

Obdobné použití Poissonova rozdělení ukazuje také následující příklad.

Příklad 4.3. Území Londýna bylo k jistému datu v průběhu druhé světové války zasaženo celkem 537 raketami typu V1, resp. V2. Mezi odborníky převládal názor, že rakety *nejsou* speciálně zaměřovány do některých městských částí. Tento názor „potvrdili" statistici následovně: Území města bylo rozděleno na $n = 24^2 = 576$ stejně velkých čtverců a byly zjištěny počty čtverců $n_0, n_1, \ldots, n_5$ s žádným, jedním, ..., pěti zásahy (tabulka 4.1). Přijmeme-li hypotézu, že rakety nejsou zaměřovány a zvolíme-li

k	0	1	2	3	4	≥ 5
n_k	229	211	93	35	7	1
$n\lambda^k \mathrm{e}^{-\lambda}/k!$	226,7	211,4	98,5	30,6	7,1	1,6

Tabulka 4.1: Porovnání skutečných a očekávaných počtů zásahů v Londýně

pevně libovolný z uvažovaných čtverců, řekněme j-tý ($1 \leq j \leq n$), pak je $X_j \sim \mathsf{bi}(537, 1/576)$, kde X_j je náhodná veličina označující počet zásahů čtverce j. Je tedy $\mathsf{P}[X_j = k] \doteq \mathrm{e}^{-\lambda}\lambda^k/k!$, kde $\lambda = np = 537/576 \doteq 0{,}9383$ pro $1 \leq j \leq 576$. Tyto pravděpodobnosti by měly být dobře odhadnuty relativními četnostmi n_k/n. Pokud platí naše hypotéza, že rakety nejsou zaměřovány, lze očekávat, že skutečné četnosti n_k budou přibližně rovny hodnotám $n\lambda^k \mathrm{e}^{-\lambda}/k!$. Tabulka 4.1 toto očekávání potvrzuje, takže závěr statistického šetření není v rozporu s názorem odborníků. Tato poslední formulace je správnější, než naše dřívější prohlášení o „potvrzení" hypotézy. K úloze se vrátíme ještě v příkladu 15.3, kdy zformulujeme skutečně statistické rozhodnutí. ○

Příklad 4.4. Vyšetřované události nastávají náhodně v čase, jejich výskyt se řídí následujícími matematickými pravidly: Existuje $\lambda \in (0, \infty)$ takové, že

(a) Pravděpodobnost výskytu alespoň jedné události v časovém úseku $\langle t, t + h)$ je rovna $\lambda h + \mathrm{o}(h)$, kde $\lim_{h\to 0+} h^{-1}\mathrm{o}(h) = 0$, $t > 0$.

(b) Pravděpodobnost výskytu více než jedné události v časovém úseku $(t, t+h)$ je rovna $o(h)$, $t > 0$.

(c) Je-li $N(t)$ náhodná veličina, která označuje počet událostí v časovém úseku $\langle 0, t)$, pak $[N(t) = j]$ a $[N(t+h) - N(t) = k]$ jsou nezávislé náhodné jevy pro $j, k = 0, 1, \ldots$ a pro $t, h > 0$.

Pro $t > 0$ a $k = 0, 1, \ldots$ označme $P_k(t) = \mathsf{P}[N_t = k]$ a definujme přirozeně $P_0(0) = 1$ a $P_k(0) = 0$ pro $k \in \mathbb{N}$. Předpoklady (a)–(c) implikují rovnosti

$$\begin{aligned} P_0(t+h) &= P[N_t = 0, N_{t+h} - N_t = 0] \\ &= P_0(t)\,(1 - \lambda h + o(h))\,, \\ P_k(t+h) &= \mathsf{P}[N_t = k, N_{t+h} - N_t = 0] \\ &\quad + \mathsf{P}[N_t = k-1, N_{t+h} - N_t = 1] + o(h) \\ &= P_k(t)\,(1 - \lambda h + o(h)) + P_{k-1}(t)\,(\lambda h + o(h)) + o(h), \\ &\quad t, h > 0, k \in \mathbb{N}. \end{aligned}$$

Odtud je

$$\begin{aligned} \frac{P_0(t+h) - P_0(t)}{h} &= -\lambda P_0(t) + \frac{1}{h} o(h), \\ \frac{P_k(t+h) - P_k(t)}{h} &= -\lambda P_k(t) + \lambda P_{k-1}(t) + \frac{1}{h} o(h). \end{aligned}$$

Limitním přechodem $h \to 0+$ pak dostaneme

$$\begin{aligned} &(4.8) \quad P_0'(t) = -\lambda P_0(t), \quad t > 0, \quad P_0(0) = 1, \\ &(4.9) \quad P_k'(t) = -\lambda P_k(t) + \lambda P_{k-1}(t), \quad t > 0, \quad P_k(0) = 0, \quad k \in \mathbb{N}. \end{aligned}$$

První diferenciální rovnice má při uvedené počáteční podmínce jediné řešení $P_0(t) = \mathrm{e}^{-\lambda t}$. Pro funkci $P_1(t)$ dostáváme tudíž rovnici

$$P_1'(t) = -\lambda P_1(t) + \lambda \mathrm{e}^{-\lambda t}, \quad t > 0,$$

s počáteční podmínkou $P_1(0) = 0$. Tato rovnice má jediné řešení, a to $P_1(t) = \mathrm{e}^{-\lambda t}$ (vyzkoušejte dosazením nebo řešte metodou variace konstant). Matematickou indukcí nyní již snadno ověříme, že nekonečná soustava diferenciálních rovnic (4.8),(4.9) má jediné řešení

$$P_k(t) = \mathrm{e}^{-\lambda t} \frac{(\lambda t)^k}{k!}, \quad t \geq 0, \quad k = 0, 1, \ldots.$$

Jinými slovy: Proces výskytu událostí s vlastnostmi (a), (b), (c) je nutně takový, že počet událostí v intervalu $(0, t)$ má Poissonovo rozdělení s parametrem λt. Můžeme se zamyslet, proč se konstanta λ obvykle nazývá intenzita výskytu událostí. ○

Příklad 4.5. (Negativně binomické rozdělení) Uvažujme podobnou posloupnost pokusů jako u binomického rozdělení: v dílčím pokusu *zdar* nastává s pravděpodobností p, *nezdar* s pravděpodobností $1 - p$. Nechť r je dané pevné přirozené číslo. Nezávislé dílčí pokusy opakujeme tak dlouho, dokud počet zdarů není roven číslu r. Náhodná veličina X, která je rovna počtu nezdarů, jež předcházejí r-tému zdaru, může nabývat všech nezáporných celočíselných hodnot. Náhodný jev $[X = k]$ nastane, právě když v prvních $k + r - 1$ pokusech nastane právě $r - 1$ zdarů a současně v $k + r$-tém pokusu nastane zdar. Hledanou pravděpodobnost tedy dostaneme jako

$$\begin{aligned} \mathsf{P}[X = k] &= \binom{k+r-1}{r-1} p^{r-1} (1-p)^k p, \qquad k = 0, 1, \ldots, \\ &= \binom{k+r-1}{r-1} p^r (1-p)^k, \qquad k = 0, 1, \ldots. \end{aligned} \tag{4.10}$$

Pro $r = 1$ se používá názvu **geometrické rozdělení**. Pravděpodobnosti geometrického rozdělení jsme již zjistili u Boseovy-Einsteinovy statistiky. ○

Příklad 4.6. (Hypergeometrické rozdělení) Mějme N předmětů, z nichž právě A má vlastnost $\mathcal{V}$. Z těchto N předmětů náhodně postupně vytáhneme (bez vracení) n předmětů. Náhodná veličina X udává počet tažených předmětů, které mají vlastnost $\mathcal{V}$. Tato náhodná veličina může nabývat pouze celočíselných hodnot z intervalu $\langle k_1, k_2\rangle$, kde je $k_1 = \max(0, A+n-N)$, $k_2 = \min(A, n)$. Jako elementární jev zvolíme n-tici vytažených předmětů. Protože nezáleží na pořadí, bude $|\Omega| = \binom{N}{n}$. Mezi vybranými předměty se může vyskytnout právě k předmětů s vlastností $\mathcal{V}$ celkem $\binom{A}{k}$ způsoby a $n - k$ předmětů bez této vlastnosti celkem $\binom{N-A}{n-k}$ způsoby, takže je nakonec

$$\mathsf{P}[X = k] = \frac{\binom{A}{k}\binom{N-A}{n-k}}{\binom{N}{n}}, \qquad k_1 \leq k \leq k_2. \tag{4.11}$$

○

4.2 Spojitá rozdělení

Distribuční funkce diskrétního rozdělení je schodovitá (po částech konstantní) se skoky pro taková x, která se mohou s kladnou pravděpodobností

vyskytnout jako hodnota náhodné veličiny. Ke spojitým rozdělením patří opačný extrém, kdy distribuční funkce nemá žádný bod nespojitosti, kdy ji lze navíc vyjádřit jako integrál s proměnnou horní mezí z nějaké jiné funkce.

Definice 4.4. Řekneme, že náhodná veličina X má **spojité rozdělení**, jestliže existuje funkce f_X, pro kterou platí

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t. \tag{4.12}$$

Nezáporná verze funkce f_X z (4.12) se nazývá **hustota rozdělení** náhodné veličiny X.

Vztahem (4.12) není funkce f_X dána jednoznačně. Distribuční funkce F_X je (absolutně) spojitá a má tudíž konečnou derivaci všude na $\mathbb{R}$. Platí $F_X'(x) = f_X(x)$ pro x, která nenáleží do nějaké množiny nulové míry, rozhodně to platí pro ta x, která jsou body spojitosti hustoty $f_X(x)$. Ovšem na množině nulové míry (například v nejvýše spočetně mnoha bodech) můžeme funkci $f_X(x)$ definovat libovolně, aniž narušíme požadavek (4.12). Požadavek na nezápornost hustoty není nijak omezující, jak plyne z následující věty.

Věta 4.5. Pro funkci $f_X(x)$ z (4.12) a pro $a < b$ platí

$$F_X(b) - F_X(a) = \int_a^b f_X(t)\mathrm{d}t, \tag{4.13}$$

$$\mathsf{P}[a \le X < b] = \int_a^b f_X(t)\mathrm{d}t, \tag{4.14}$$

$$\mathsf{P}[a < X < b] = \int_a^b f_X(t)\mathrm{d}t, \tag{4.15}$$

$$\mathsf{P}[a < X \le b] = \int_a^b f_X(t)\mathrm{d}t, \tag{4.16}$$

$$\mathsf{P}[a \le X \le b] = \int_a^b f_X(t)\mathrm{d}t. \tag{4.17}$$

Dále platí

$$\int_{-\infty}^{\infty} f_X(t)\mathrm{d}t = 1, \tag{4.18}$$

$$x \in \mathbb{R} \Rightarrow f_X(x) \ge 0 \quad \text{skoro všude.} \tag{4.19}$$

D ů k a z: Z vlastností distribuční funkce plyne řada vlastností hustoty.

Především, použijeme-li vztah (4.12) v rovnosti (4.5), dostaneme

$$\begin{aligned}\mathsf{P}[a \le X < b] &= \int_{-\infty}^{b} f_X(t)\mathrm{d}t - \int_{-\infty}^{a} f_X(t)\mathrm{d}t \\ &= \int_a^b f_X(t)\mathrm{d}t,\end{aligned}$$

když jsme použili aditivní vlastnost integrálu. Takto jsme dokázali (4.13) a (4.14). Vztahy (4.15)-(4.17) dostaneme, když využijeme skutečnost, že pro spojité rozdělení musí být distribuční funkce spojitá (dokonce absolutně spojitá), takže pro každé $x \in \mathbb{R}$ musí platit $\mathsf{P}[X = x] = 0$.

Vztah (4.18) je ekvivalentní s druhým vztahem z (4.4). Funkce $f_X(x)$ z (4.12) je nezáporná, protože distribuční funkce je neklesající. □

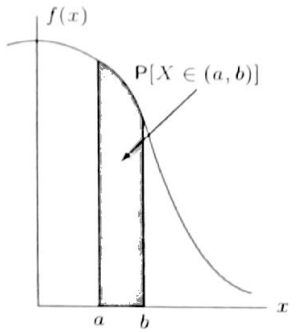
Lze dokázat podobné tvrzení jako o distribuční funkci. Je-li $g(x)$ nezáporná funkce splňující

$$\int_{-\infty}^{\infty} g(x)\mathrm{d}x = 1,$$

existuje náhodná veličina X se spojitým rozdělením, jejíž hustota je (až na množinu nulové míry) totožná s funkcí $g(x)$. Při důkazu se vyjde z toho, že funkce

$$G(x) = \int_{-\infty}^{x} g(t)\mathrm{d}t$$

má všechny vlastnosti distribuční funkce.



Obrázek 4.2: Význam plochy pod hustotou

Pro představu o interpretaci hustoty jsou asi nejdůležitější vztahy (4.14)–(4.17), graficky znázorněné také na obrázku 4.2. Navíc, zvolíme-li $a = x, b = x + \Delta x$, můžeme použít přibližnou rovnost

$$\int_x^{x+\Delta x} f_X(t)\mathrm{d}t \doteq f_X(x)\Delta x,$$

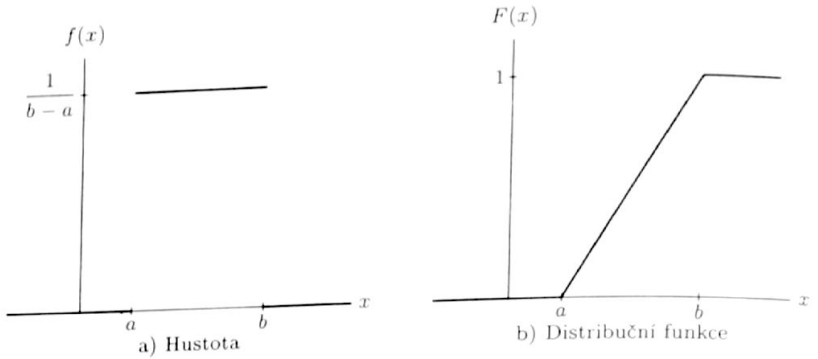
takže dostaneme

$$\mathsf{P}[x \le X < x + \mathrm{d}x] \doteq f_X(x)\mathrm{d}x. \tag{4.20}$$

Příklad 4.7. (Rovnoměrné rozdělení) Mějme reálná čísla $a < b$. Funkce

$$\begin{array}{rclrcll} f(x) &=& 0 & F(x) &=& 0 & x \le a \\ &=& \frac{1}{b-a} & &=& \frac{x-a}{b-a} & a < x < b \\ &=& 0 & &=& 1 & b \ge x \end{array} \tag{4.21}$$

určují hustotu a distribuční funkci náhodné veličiny s rovnoměrným rozdělením na intervalu $(0,1)$. Interpretace (4.20) je v tomto případě velmi názorná: jistota (jednotková pravděpodobnost) je na intervalu (a,b) rozprostřena *rovnoměrně*. ○



Obrázek 4.3: Rovnoměrné rozdělení

Příklad 4.8. (Exponenciální rozdělení) Uvažujme náhodný jev, který se vyskytuje v náhodných okamžicích, přičemž výskyty v nepřekrývajících se časových intervalech jsou nezávislé. Označme symbolem $Q(t)$ pravděpodobnost, že sledovaný jev *nenastane* během intervalu délky t. Jsou-li t_1, t_2 délky dvou na sebe navazujících časových intervalů, pak náš předpoklad lze zapsat jako

$$Q(t_1 + t_2) = Q(t_1)Q(t_2).$$

Předpokládejme navíc spojitost funkce Q a $Q(0) = 1$. Pro $t > 0, \Delta t > 0$ můžeme psát

$$\ln Q(t + \Delta t) \quad = \quad \ln Q(t) + \ln Q(\Delta t),$$

takže po úpravě máme

$$\lim_{\Delta t \to 0+} \frac{\ln Q(t + \Delta t) - \ln Q(t)}{\Delta t} = \lim_{\Delta t \to 0+} \frac{\ln Q(\Delta t)}{\Delta t} = -\lambda,$$

kde jsme označili jako $-\lambda$ derivaci funkce $\ln Q$ v bodě 0 zprava. Docházíme k diferenciální rovnici

$$\frac{\mathrm{d}}{\mathrm{d}t} \ln Q(t) = -\lambda,$$

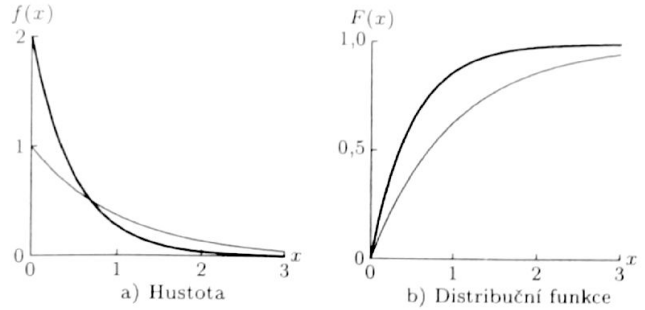
která má řešení $Q(t) = \mathrm{e}^{-\lambda t}$, splňující požadavek $Q(0) = 1$. Označme jako X náhodný okamžik, kdy nastane (poprvé!) sledovaný náhodný jev. Zřejmě je $F_X(t) = 1 - Q(t)$, tedy

$$F_X(t) = \begin{cases} 1 - \mathrm{e}^{-\lambda t}, & t > 0 \\ 0 & t \le 0. \end{cases} \tag{4.22}$$

Náhodná veličina X má **exponenciální rozdělení** s parametrem λ, což označíme $X \sim \mathrm{ex}(\lambda)$. Hustotu dostaneme snadno derivováním distribuční funkce

$$f_X(t) = \begin{cases} \lambda \mathrm{e}^{-\lambda t} & t > 0, \\ 0 & t \le 0. \end{cases} \tag{4.23}$$

○



Obrázek 4.4: Exponenciální rozdělení s parametry λ=1 a λ=2 (tučně)

4.3 Rozdělení funkce náhodné veličiny

Může být užitečné zajímat se o rozdělení *funkce* náhodné veličiny. Místo výčetní tloušťky kmene nás může například zajímat spíše její logaritmus. Uvedenou situaci musíme zřejmě modelovat pomocí složené funkce. Mějme náhodnou veličinu X definovanou na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, \mathsf{P})$, mějme dále reálnou funkci g. Kdy má smysl hovořit o $Y = g(X)$ jako o náhodné veličině? Pochopitelným (a jediným) požadavkem bude

$$[g(X) < x] \quad = \quad \{\omega \in \Omega : g(X(\omega)) < x\} \in \mathcal{A} \text{ pro všechna } x \in \mathbb{R}.$$

Tomuto požadavku měřitelnosti vyhovuje například spojitá funkce g.

Příklad 4.9. (Normální rozdělení) Nejdůležitější spojité rozdělení má hustotu

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2}. \tag{4.24}$$

Distribuční funkci tohoto **normovaného normálního** rozdělení

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2} \mathrm{d}t \tag{4.25}$$

nelze vyjádřit pomocí elementárních funkcí, je však podrobně tabelována v dostupných tabulkách a jsou známy poměrně přesné aproximace (viz [2]).

Mějme čísla $\mu \in \mathbb{R}, \sigma > 0$. Předpokládejme, že náhodná veličina Z má normální rozdělení s hustotou (4.24). Náhodná veličina $Y = \mu + \sigma Z$ má distribuční funkci

$$\begin{aligned} F_Y(y) &= \mathsf{P}[Y < y] = \mathsf{P}[\mu + \sigma Z < y] \\ &= \mathsf{P}[Z < \frac{y-\mu}{\sigma}] = \Phi\left(\frac{y-\mu}{\sigma}\right) \\ &= \int_{-\infty}^{\frac{y-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2} \mathrm{d}z. \end{aligned}$$

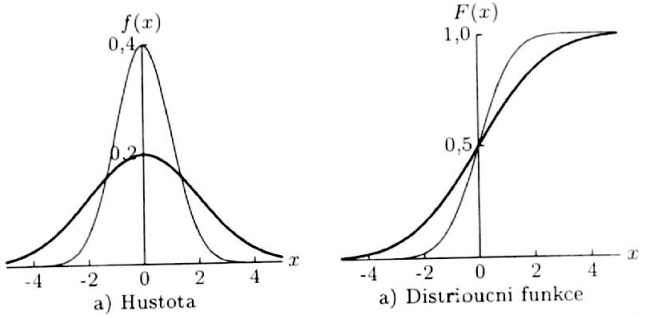
Po substituci $x = \mu + \sigma z$ dostaneme distribuční funkci náhodné veličiny Y ve tvaru

$$F_Y(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x,$$

takže hustota náhodné veličiny Y je nutně

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \tag{4.26}$$

Také rozdělení náhodné veličiny Y se nazývá normální. Symbolicky píšeme $Y \sim \mathsf{N}(\mu, \sigma^2)$, takže tvrzení, že Z má normované normální rozdělení zapíšeme jako $Z \sim \mathsf{N}(0,1)$. ○



Obrázek 4.5: Normální rozdělení $\mathsf{N}(0,1)$ a $\mathsf{N}(0,4)$ (tučně)

Příklad 4.10. Nechť má náhodná veličina X rovnoměrné rozdělení na intervalu $(0,1)$. Nalezněme distribuční funkci a hustotu náhodné veličiny $Y = X^2$. Pro $0 < y < 1$ je

$$\begin{aligned} F_Y(y) &= \mathsf{P}[Y < y] \\ &= \mathsf{P}[X^2 < y] \\ &= \mathsf{P}[X < \sqrt{y}] \\ &= \sqrt{y}, \end{aligned}$$

jinak je $F_Y(y) = 0$ pro $y \le 0$, $F_Y(y) = 1$ pro $y \ge 1$. Derivováním podle y dostaneme hustotu

$$\begin{aligned} f_Y(y) &= 0 && \text{pro } y \le 0, \\ &= \frac{1}{2\sqrt{y}} && \text{pro } 0 < y < 1, \\ &= 0 && \text{pro } y \ge 1. \end{aligned}$$ ○

Příklad 4.11. (Rozdělení $\chi^2(1)$) Nechť Z má normované normální rozdělení s hustotou $\varphi(z)$. Hledejme hustotu náhodné veličiny $X = Z^2$. Pro $x \le 0$ je distribuční funkce i hustota nulová, pro $x > 0$ dostaneme

$$\begin{aligned} F_X(x) &= \mathsf{P}[Z^2 < x] \\ &= \mathsf{P}[-\sqrt{x} < Z < \sqrt{x}] \\ &= \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2} \mathrm{d}z \\ &= \int_0^x \frac{1}{\sqrt{2\pi}} t^{-1/2} \mathrm{e}^{-t/2} \mathrm{d}t, \end{aligned}$$

takže po zderivování podle x dostaneme hustotu tvaru

$$f_X(x) = \frac{1}{\sqrt{2\pi}} x^{-1/2} \mathrm{e}^{-x/2}. \tag{4.27}$$

Rozdělení náhodné veličiny X s hustotou (4.27) se nazývá **χ^2-rozdělení** s jedním stupněm volnosti a značí se $X \sim \chi^2(1)$. ○

4.4 Kvantily

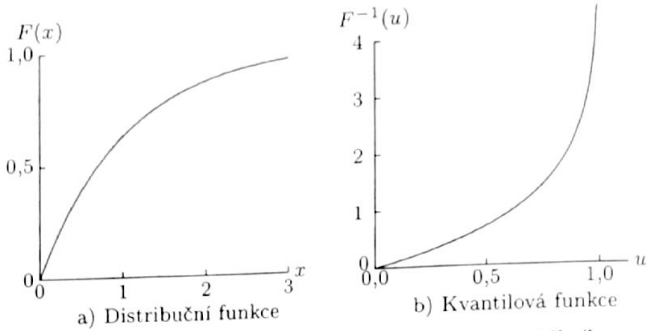
Zejména ve statistice se používá následující pojem.

Definice 4.5. Nechť $F(x)$ je distribuční funkce náhodné veličiny X. Funkce

$$F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}, \quad 0 < u < 1, \tag{4.28}$$

se nazývá **kvantilová** funkce náhodné veličiny X. Pro $0 < \alpha < 1$ se hodnota $F^{-1}(\alpha)$ nazývá **α-kvantil**.

V případě, že je distribuční funkce na množině $\{x \in \mathbb{R} : 0 < F(x) < 1\}$ spojitá a rostoucí, je kvantilová funkce totožná s obyčejnou inverzní funkcí k distribuční funkci na tomto intervalu.



a) Distribuční funkce b) Kvantilová funkce

Obrázek 4.6: Distribuční a kvantilová funkce exponenciálního rozdělení

Velmi populární je hodnota $\Phi^{-1}(0{,}975) = 1{,}96$, tedy 0,975-kvantil normovaného normálního rozdělení. Vzhledem k symetrii hustoty tohoto rozdělení kolem nuly pak totiž pro náhodnou veličinu Z s rozdělením $\mathsf{N}(0,1)$ platí, že

$$\mathsf{P}(|Z| < 1{,}96) = 0{,}95. \tag{4.29}$$

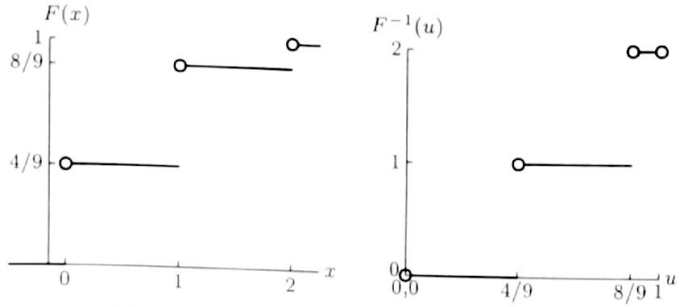
Pomocí kvantilové funkce lze vyjádřit také často používané **kritické hodnoty**. Například má-li náhodná veličina Z normované normální rozdělení s distribuční funkcí $\Phi(z)$, pak pro $\alpha \in (0,1)$ je kritická hodnota definována vztahem $z(\alpha) = \Phi^{-1}(1-\alpha)$. Platí tedy

$$\mathsf{P}[Z > z(\alpha)] = \alpha \tag{4.30}$$

Kritické hodnoty dalších používaných rozdělení zavedeme v oddílu 8.4.

Příklad 4.12. Připomeňme exponenciální rozdělení z příkladu 4.8. Protože jeho distribuční funkce je na intervalu $(0,\infty)$ rostoucí a zobrazuje jej na $(0,1)$, je kvantilová funkce inverzní funkcí k $1 - \mathrm{e}^{-\lambda x}$, tedy

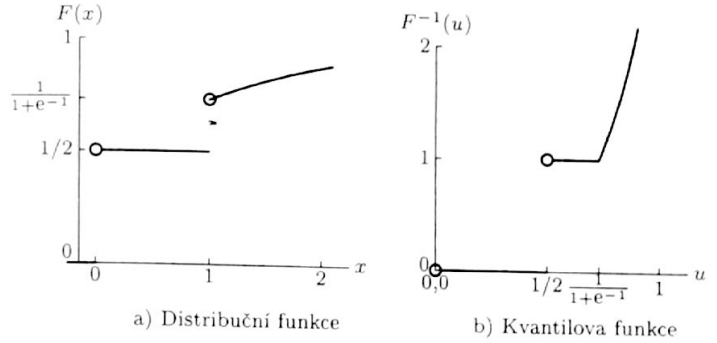
$$F^{-1}(u) = -\frac{1}{\lambda}\ln(1-u), \qquad u \in (0,1).$$



a) Distribuční funkce b) Kvantilová funkce

Obrázek 4.7: Distribuční a kvantilová funkce binomického rozdělení

Příklad 4.13. Náhodná veličina $X \sim \mathsf{bi}(2, 1/3)$ nabývá pouze tří hodnot s pravděpodobnostmi $\mathsf{P}[X=0] = 4/9$, $\mathsf{P}[X=1] = 4/9$ a $\mathsf{P}[X=2] = 1/9$. Kvantilová funkce $F^{-1}(u)$ je na intervalu $(0,1)$ (podobně jako distribuční funkce) zleva spojitá a má tvar uvedený na obrázku 4.7 b). ○



a) Distribuční funkce b) Kvantilova funkce

Obrázek 4.8: Distribuční a kvantilová funkce rozdělení z příkladu 4.14

Příklad 4.14. Uvažujme náhodnou veličinu X, jejíž distribuční funkce je dána vztahy (obr. 4.8)

$$F_X(x) = \begin{cases} 0 & x \leq 0, \\ 1/2 & 0 < x \leq 1, \\ 1/(1+\mathrm{e}^{-x}) & x > 1. \end{cases}$$

Kvantilová funkce je dána vztahy (obr. 4.8)

$$F^{-1}(u) = \begin{cases} 0 & 0 < u \le 1/2, \\ 1 & 1/2 < u \le 1/(1+\mathrm{e}^{-1}), \\ \ln(u/(1-u)) & u > 1/(1+\mathrm{e}^{-1}). \end{cases} \qquad \bigcirc$$

4.5 Moivreova-Laplaceova věta

Poissonova věta 4.4 určuje asymptotické chování binomického rozdělení při $n \to \infty$ a $p \to 0$. Vyšetření limity binomických pravděpodobností při $n \to \infty$, je-li pravděpodobnost zdaru p konstantní, je úlohou teorie pravděpodobnosti s velmi dlouhou a významnou historií.

Věta 4.6.(Moivreova-Laplaceova lokální věta – 1801) Pro $p \in (0,1)$, $q = 1-p$, při $n \to \infty$, platí

$$\max_{0 \le k \le n} \left| \binom{n}{k} p^k q^{n-k} - \frac{1}{\sqrt{npq}} \varphi(x_{n,k}) \right| = \mathrm{o}(n^{-1/2}), \tag{4.31}$$

kde $\varphi(x)$ je hustota normálního rozdělení daná (4.24) a

$$x_{n,k} = \frac{k-np}{\sqrt{npq}}.$$

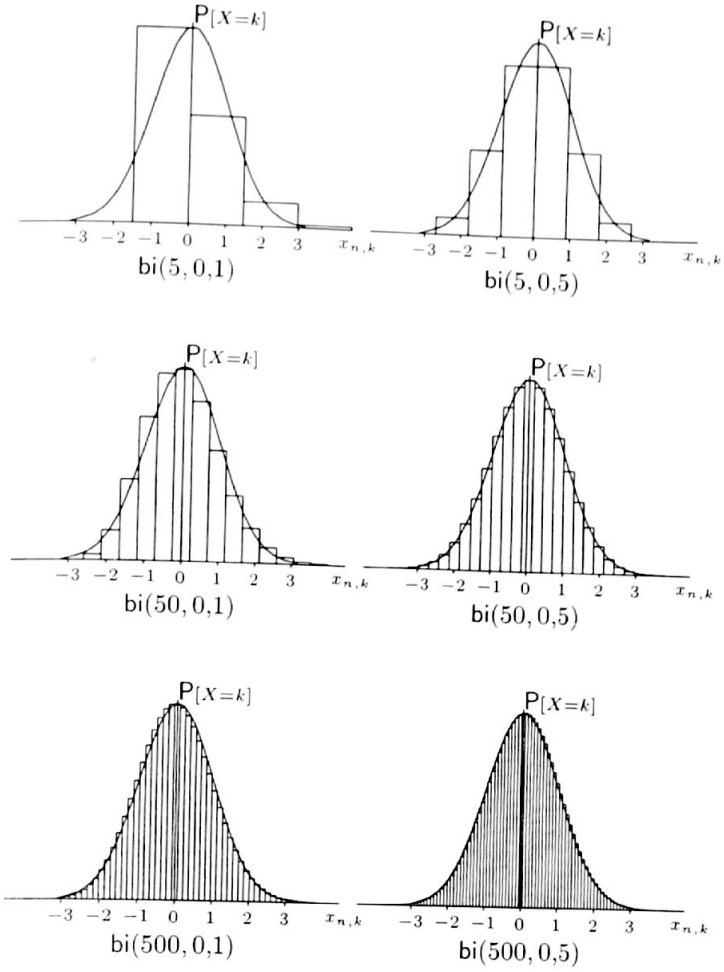
Je-li tedy $X_n \sim \mathsf{bi}(n,p)$, pak $\mathsf{P}[X_n = k] = \frac{1}{\sqrt{npq}} \varphi(x_{n,k}) + \mathrm{o}(n^{-1/2})$ při $n \to \infty$ stejnoměrně pro $0 \le k \le n$, volba $p = q = 1/2$ pak implikuje

$$\mathsf{P}[X_{2n} = n] = \binom{2n}{n} 2^{-2n} = \frac{1}{\sqrt{\pi n}} + \mathrm{o}(n^{-1/2}),$$

což je zpřesnění limitního přechodu, který jsme obdrželi pomocí Stirlingovy formule v odstavci 3.7 o náhodné procházce. Obecněji platí

$$\mathsf{P}[X_n = [np]] = \frac{1}{\sqrt{2\pi npq}} + \mathrm{o}(n^{-1/2}).$$

Grafická znázornění na obrázku 4.9 ilustrují platnost limitního přechodu dosti přesvědčivě. Pro názornost je na vodorovné ose použito vždy stejné měřítko pro veličinu $x_{n,k}$ a každé znázornění obsahuje graf hustoty rozdělení $\mathsf{N}(0,1)$. Jednotlivé obdélníky mají pak plochu úměrnou pravděpodobnosti veličiny $x_{n,k}$, jejíž hodnota odpovídá středu daného intervalu.



Obrázek 4.9: Souvislost binomického rozdělení s normálním

D ů k a z: Připomeňme definici komplexní exponenciely $e^{i\alpha} = \cos\alpha + i\sin\alpha$ a dva známé vztahy klasického integrálního počtu:

$$\varphi(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-iux}e^{-u^2/2}du, \quad x \in \mathbb{R}, \tag{4.32}$$

$$\delta_{jk} = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{it(j-k)}dt, \quad j,k \geq 0 \quad \text{(Kroneckerovo } \delta\text{)}.$$

Odtud plyne

$$\begin{aligned}\binom{n}{k}p^k q^{n-k} &= \sum_{j=0}^{n}\binom{n}{j}p^j q^{n-j}\delta_{jk}\\ &= \sum_{j=0}^{n}\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-itk}\binom{n}{j}p^j e^{itj}q^{n-j}dt\\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-itk}(pe^{it}+q)^n dt.\end{aligned}$$

Počítejme dále:

$$\begin{aligned}2\pi\sqrt{npq}\binom{n}{k}p^k q^{n-k} &= \sqrt{npq}\int_{-\pi}^{\pi} e^{-itk}(pe^{it}+q)^n dt\\ &= \int_{-\pi\sqrt{npq}}^{\pi\sqrt{npq}} e^{-\frac{iku}{\sqrt{npq}}}\left(pe^{\frac{iu}{\sqrt{npq}}}+q\right)^n du\\ &\quad \text{(provedli jsme substituci } u = t\sqrt{npq})\\ &= \int_{-\pi\sqrt{npq}}^{\pi\sqrt{npq}} e^{-iu\,x_{n,k}}\left(e^{-iu\frac{p}{\sqrt{npq}}}\right)^n\left(pe^{\frac{iu}{\sqrt{npq}}}+q\right)^n du\\ &= \int_{-\pi\sqrt{npq}}^{\pi\sqrt{npq}} e^{-iu\,x_{n,k}}\left(S\left(\frac{u}{\sqrt{n}}\right)\right)^n du,\end{aligned}$$

kde

$$S(y) = pe^{iy\sqrt{\frac{q}{p}}} + qe^{-iy\sqrt{\frac{p}{q}}}.$$

Použijeme-li nyní také rovnost (4.32), pak

$$\begin{aligned}&2\pi\sqrt{npq}\max_{0\leq k\leq n}\left|\binom{n}{k}p^k q^{n-k} - \frac{1}{\sqrt{npq}}\varphi(x_{n,k})\right| =\\ &= \max_{0\leq k\leq n}\left|\int_{-\pi\sqrt{npq}}^{\pi\sqrt{npq}} e^{-iu\,x_{n,k}}\left(S\left(\frac{u}{\sqrt{n}}\right)\right)^n du - \int_{-\infty}^{+\infty} e^{-iu\,x_{n,k}}e^{-u^2/2}du\right|\end{aligned}$$

$$\begin{aligned}&\leq \int_{-\pi\sqrt{npq}}^{\pi\sqrt{npq}}\left|\left(S\left(\frac{u}{\sqrt{n}}\right)\right)^n - e^{-u^2/2}\right|du + \int_{[|u|\geq\pi\sqrt{npq}]} e^{-u^2/2}du\\ &= I_n + J_n, \quad \text{(protože } |e^{-iux_{n,k}}| = 1\text{)}.\end{aligned}$$

Zřejmě platí $\lim_{n\to\infty} J_n = 0$ (zbytek konvergentního integrálu). Věta bude dokázána, ověříme-li, že je $\lim_{n\to\infty} I_n = 0$. Rozvineme nejprve funkci $S(y)$ do Taylorovy řady:

$$S(y) = p\sum_{k=0}^{\infty}\frac{\left(iy\sqrt{\frac{q}{p}}\right)^k}{k!} + q\sum_{k=0}^{\infty}\frac{\left(-iy\sqrt{\frac{p}{q}}\right)^k}{k!} = 1 - \frac{y^2}{2} + Z(y)$$

Odtud je

$$S\left(\frac{u}{\sqrt{n}}\right) = 1 - \frac{u^2}{2n} + Z\left(\frac{u}{\sqrt{n}}\right)$$

a

$$\begin{aligned}\left|Z\left(\frac{u}{\sqrt{n}}\right)\right| &\leq \sum_{k=3}^{\infty}\left(\frac{1}{\sqrt{n}}\right)^k\left(p\frac{\left(|u|\sqrt{\frac{q}{p}}\right)^k}{k!} + q\frac{\left(|u|\sqrt{\frac{p}{q}}\right)^k}{k!}\right)\\ &\leq \frac{1}{n^{3/2}}\sum_{k=3}^{\infty}\frac{|u|^k\left(\sqrt{\frac{q}{p}}+\sqrt{\frac{p}{q}}\right)^k}{k!} \leq n^{-3/2}e^{|u|C}\end{aligned}$$

pro nějakou konstantu $C \in (0,\infty)$, tj.

$$S\left(\frac{u}{\sqrt{n}}\right) = 1 - \frac{u^2}{2n} + o(n^{-1})$$

při $n \to \infty$. Odtud již plyne, že

$$\lim_{n\to\infty}\left(S\left(\frac{u}{\sqrt{n}}\right)\right)^n = \lim_{n\to\infty}\left(1 - \frac{u^2}{2n} + o(n^{-1})\right)^n = e^{-\frac{u^2}{2}}$$

pro $-\infty < u < \infty$. Zjistili jsme, že integrandy v integrálech I_n konvergují k nule a tudíž také $\lim_{n\to\infty} I_n = 0$. (Možnost záměny limity a integrálu lze po jisté námaze ověřit metodami, které jsme použili při vyšetřování funkce $(S(u/\sqrt{n}))^n$.) □

Řád konvergence (4.31) umožňuje přepsat naše tvrzení do integrální formy:

Věta 4.7. (Moivreova-Laplaceova integrální věta) Pro $p \in (0,1)$ a $-\infty < a < b < \infty$ nechť X_n jsou náhodné veličiny s rozdělením $\mathrm{bi}(n,p)$. Pak platí

$$\lim_{n\to\infty} \mathsf{P}\left[a < \frac{X_n - np}{\sqrt{npq}} < b\right] = \Phi(b) - \Phi(a), \tag{4.33}$$

kde $q = 1-p$ a $\Phi(z)$ je distribuční funkce normovaného normálního rozdělení.

Poznamenejme, že k limitě $\Phi(b) - \Phi(a) = \int_a^b \varphi(x)\mathrm{d}x$ konvergují i posloupnosti $\mathsf{P}[a \le \frac{X_n-np}{\sqrt{npq}} \le b]$, $\mathsf{P}[a < \frac{X_n-np}{\sqrt{npq}} \le b]$ atd., protože podle věty 4.6 je $\lim_{n\to\infty} \mathsf{P}[X_n = k] = 0$ pro každé celé číslo k. Tvrzení věty 4.7 lze velmi podstatně zobecnit (věta 9.6).

D ů k a z: Pomocí věty 4.6 dostáváme:

$$\begin{aligned}\mathsf{P}\left[a < \frac{X_n - np}{\sqrt{npq}} < b\right] &= \mathsf{P}\left[[np + a\sqrt{npq}] + 1 \le X_n \le [np + b\sqrt{npq}]\right] \\ &= \sum_{k=k_{a,n}}^{k_{b,n}} \mathsf{P}[X_n = k] \\ &= \sum_{k=k_{a,n}}^{k_{b,n}} \left(\frac{1}{\sqrt{npq}}\varphi(x_{n,k}) + \mathrm{o}(n^{-1/2})\right),\end{aligned}$$

kde

$$k_{a,n} = [np + a\sqrt{npq}] + 1, \quad k_{b,n} = [np + b\sqrt{npq}], \quad x_{n,k} = \frac{k - np}{\sqrt{npq}}.$$

Jelikož součet v mezích $k_{a,n}, k_{b,n}$ má celkem $k_{b,n} - k_{a,n} + 1 \le C\sqrt{n}$ sčítanců, platí $\lim_{n\to\infty} \sum_{k=k_{a,n}}^{k_{b,n}} \mathrm{o}(n^{-1/2}) = 0$. Odtud již plyne, že

$$\lim_{n\to\infty} \mathsf{P}\left[a < \frac{X_n - np}{\sqrt{npq}} < b\right] = \lim_{n\to\infty} \sum_{k=k_{a,n}}^{k_{b,n}} \frac{1}{\sqrt{npq}}\varphi(x_{n,k}) = \int_a^b \varphi(x)\mathrm{d}x,$$

protože je $x_{n,k+1} - x_{n,k} = \frac{1}{\sqrt{npq}}$ a součet v předcházejícím řádku je vytvořující Riemanova suma pro integrál uvedený na konci řádku, protože $x_{n,k_{b,n}} \to b$ a $x_{n,k_{a,n}} \to a$ při $n \to \infty$. □

Příklad 4.15. Při studiu náhodné procházky v odstavci 3.7 jsme jako S_n označili celočíselnou polohu částice v okamžiku n. Odvodili jsme její rozdělení pravděpodobností $\mathsf{P}[S_n = 2k - n] = \binom{n}{k}2^{-n}$ pro $k = 0, 1, \ldots, n$,

tedy $(n + S_n)/2 \sim \mathrm{bi}(n, 1/2)$. Věta 4.6 tudíž implikuje platnost limitního přechodu

$$\mathsf{P}[S_{2n} = 2(k-n)] = \mathsf{P}\left[\frac{S_{2n} + 2n}{2} = k\right] = \frac{1}{\sqrt{\pi n}} \exp\left(\frac{(k-n)^2}{n}\right) + \mathrm{o}(n^{-1/2})$$

při $n \to \infty$, $k = 0, 1, \ldots$. Použijeme-li větu 4.7, dostáváme $\mathsf{P}\left[a < \frac{S_n}{\sqrt{n}} < b\right] \to \Phi(b) - \Phi(a)$ při $n \to \infty$ pro každou dvojici $a < b$. ○

Tvrzení věty 4.7 lze zapsat ve tvaru

$$\lim_{n\to\infty} \left|\mathsf{P}[A \le X_n \le B] - \left(\Phi\left(\frac{B - np}{\sqrt{npq}}\right) - \Phi\left(\frac{A - np}{\sqrt{npq}}\right)\right)\right| = 0 \tag{4.34}$$

pro $-\infty < A < B < \infty$. Někdy, zejména při nepříliš velkých hodnotách n, se používá poněkud přesnější aproximace

$$\mathsf{P}[A \le X_n \le B] \doteq \Phi\left(\frac{B + 0{,}5 - np}{\sqrt{npq}}\right) - \Phi\left(\frac{A - 0{,}5 - np}{\sqrt{npq}}\right), \tag{4.35}$$

která přihlíží k celočíselnému charakteru hodnot náhodné veličiny X_n.

Příklad 4.16. Hodíme symetrickou kostku celkem 12 000-krát. Jaká je pravděpodobnost P toho, že počet šestek leží v intervalu $\langle 1800, 2100\rangle$? Bylo by velmi obtížné počítat součet $\sum_{k=1800}^{2100} \binom{12000}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{12000-k}$. Limitní přechod v (4.34), kde zvolíme $p = 1/6$, $A = 1800$, $B = 2100$, $n = 12000$, nabízí aproximaci

$$\begin{aligned}P &\doteq \Phi\left(\frac{2100 - 2000}{\sqrt{12000\frac{1}{6}\frac{5}{6}}}\right) - \Phi\left(\frac{1800 - 2000}{\sqrt{12000\frac{1}{6}\frac{5}{6}}}\right) \\ &= \Phi(\sqrt{6}) - \Phi(-2\sqrt{6}) \doteq \Phi(2{,}449) - \Phi(-4{,}899) \\ &\doteq 0{,}992.\end{aligned}$$

(Při výpočtu jsme použili tabulku hodnot distribuční funkce $\Phi(z)$.) ○

Někdy je užitečný jiný zápis tvrzení věty 4.7:

$$\lim_{n\to\infty} \left|\mathsf{P}\left[\left|\frac{X_n}{n} - p\right| < \delta\right] - \left(\Phi\left(\frac{n\delta}{\sqrt{npq}}\right) - \Phi\left(-\frac{n\delta}{\sqrt{npq}}\right)\right)\right| = 0 \tag{4.36}$$

pro $\delta \in (0,1)$.

Příklad 4.17. Nezávisle opakujeme pokus s výsledky 1 a 0, které mají neznámé pravděpodobnosti $p \in (0,1)$ a $q = 1 - p$. K odhadu parametru p

použijeme relativní četnost X_n/n, kde X_n označuje počet jedniček v sérii n pokusů. Protože je $X_n \sim \mathrm{bi}(n,p)$, věta 4.7 ve znění (4.36) umožňuje stanovit počet pokusů n potřebný k zajištění předepsané přesnosti odhadu $\delta > 0$ se spolehlivostí $1-\beta \in (0,1)$. Hledáme nejmenší přirozené číslo n, které splňuje nerovnost

$$\mathsf{P}\left[\left|\frac{X_n}{n} - p\right| < \delta\right] \geq 1-\beta,$$

kterou podle (4.36) můžeme aproximovat nerovností

$$\Phi\left(\frac{n\delta}{\sqrt{npq}}\right) - \Phi\left(-\frac{n\delta}{\sqrt{npq}}\right) = 2\Phi\left(\frac{n\delta}{\sqrt{npq}}\right) - 1 \geq 1-\beta,$$

která je ekvivalentní s požadavkem $n\delta/\sqrt{npq} \geq z(\beta/2)$, kde $z(p)$ je řešení rovnice $\Phi(z(p)) = 1-p$ (kritická hodnota normovaného normálního rozdělení, přehled označení kritických hodnot je uveden v tabulce 8.2, samotné kritické hodnoty jsou uvedeny v Appendixu, tab. B1). Použijeme-li odhad $pq \leq 1/4$, přicházíme k závěru, že potřebný počet pokusů je dán nerovností

$$n \geq \left(\frac{z(\beta/2)}{2\delta}\right)^2.$$

Volba $\delta = 0{,}05$ a $1-\beta = 0{,}9$ tudíž vyžaduje více než 270 pokusů. ○

Závěrem poznamenejme, že korektní odvození správnosti limit v (4.34) a (4.36) vyžaduje znalost té skutečnosti, že posloupnost pravděpodobností $\mathsf{P}\left[a < \frac{X_n - np}{\sqrt{npq}} < b\right]$ (viz (4.33)) konverguje k rozdílu $\Phi(b) - \Phi(a)$ stejnoměrně pro $-\infty < a < b < \infty$. Máme ovšem k disposici mnohem hlubší a přesnější zjištění:

Věta 4.8.(Berry-Essénova nerovnost) Buď $X_n \sim \mathrm{bi}(n,p)$ a $F_n(x)$ nechť označuje distribuční funkci náhodné veličiny $\frac{X_n - np}{\sqrt{npq}}$. Pak

$$M_n = \sup_{-\infty<x<\infty} |F_n(x) - \Phi(x)| \leq 0{,}7995\frac{p^2+q^2}{\sqrt{npq}},$$

tj. $M_n \leq 0{,}7995 n^{-1/2}$ při $p = q = 1/2$.

Důkaz této nerovnosti je mimo naše možnosti (viz například [18, str. 282]), nerovnost sama je velmi důležitá.

Aproximace binomického rozdělení rozdělením normálním prostřednictvím (4.33), (4.34) nebo (4.36) mají (přesně) řád $O(n^{-1/2})$ a jsou tím přesnější, čím menší je vzdálenost pravděpodobnosti od $1/2$. (Funkce $f(p) = \frac{p^2+(1-p)^2}{\sqrt{p(1-p)}}$ na intervalu $(0,1)$ nabývá svého minima v bodě $p = 1/2$, přičemž $f(1/2) = 1$.)

4.6 Cvičení

4.1. Určete rozdělení celkového počtu ok, která padnou při hodu třemi hracími kostkami.

4.2. Rozhodněte, které z následujících funkcí jsou hustotami:

a) $c\sin x$ pro $x \in (0,\pi/2)$,
b) $c\sin x$ pro $x \in (0,2\pi)$,
c) $cx^2\mathrm{e}^{-x^3}$ pro $x > 0$,
d) $cx^3\mathrm{e}^{-x^4}$,
e) $c|x^3|\mathrm{e}^{x^4}$,
f) $c\mathrm{e}^{-|x|}$,
g) $c\mathrm{e}^{x}$,
h) $c\mathrm{e}^{-x}$ pro $x \geq 0$, jinak $c(1+x)^{-1}$.

Mimo vymezený interval je nabízená funkce vždy rovna nule.

4.3. Pro náhodnou veličinu s hustotou

$$f(x) = \begin{cases} 3x^2 & \text{pro } 0 \leq x \leq 1, \\ 0 & \text{jinak.} \end{cases}$$

najděte distribuční funkci, kvantilovou funkci a pravděpodobnost $\mathsf{P}[1/3 < X < 2/3]$.

4.4. Pro náhodnou veličinu X s rovnoměrným rozdělením na intervalu $(-1,1)$ určete pravděpodobnosti

a) $\mathsf{P}\left(X^2 > \frac{1}{2}\right)$,
b) $\mathsf{P}\left(X^2 > \frac{1}{2} | X > 0\right)$.

4.5. Pro náhodnou veličinu X s rovnoměrným rozdělením na intervalu $(1,2)$ určete pravděpodobnosti

a) $\mathsf{P}\left(1/X < \frac{2}{3}\right)$,
b) $\mathsf{P}\left(1/X^2 < \frac{2}{3}\right)$.

4.6. Nechť

$$F(x) = \begin{cases} 0 & \text{pro } x \leq 0, \\ \sin x & \text{pro } 0 < x < \pi/2, \\ 1 & \text{pro } x \geq \pi/2. \end{cases}$$

Určete distribuční funkci a hustotu náhodné veličiny $W = \sin X$.

4.7. V lese, jehož hranice tvoří na mapě rovnostranný trojúhelník, se ztratilo dítě. Předpokládáme, že pravděpodobnost toho, že dítě je v určité části lesa, je úměrná pouze velikosti této části, nikoliv jejímu umístění.

a) Jaké je rozdělení vzdálenosti dítěte od zvolené strany lesa?
b) Jaké je rozdělení vzdálenosti dítěte od nejbližší strany lesa?

4.8. Najděte rozdělení objemu krychle, jejíž hrana má náhodnou délku s rovnoměrným rozdělením na intervalu $(0, 10)$.

4.9. Dva hráči košíkové střídavě házejí na koš tak dlouho, dokud jeden z nich nezasáhne. První hráč zasáhne koš s pravděpodobností 0,4, druhý s pravděpodobností 0,6. Určete rozdělení pravděpodobností počtu hodů, které provede každý z nich.

4.10. Terč tvoří kruh číslo 1 a dvě mezikruží s čísly 2 a 3. Zásah do kruhu číslo 1 znamená 10 bodů, zásah do mezikruží číslo 2 znamená 5 bodů a zásah do mezikruží číslo 3 znamená -1 bod. Kruh číslo 1, resp. mezikruží číslo 2 a 3 lze zasáhnout s pravděpodobnostmi, které jsou po řadě rovny 0,5, 0,3 a 0,2. Určete rozdělení pravděpodobností náhodného počtu bodů získaných po třech výstřelech.

5. Náhodný vektor

V praxi velmi často pracujeme nejen s jedinou náhodnou veličinou, ale také s náhodným vektorem. Zajímáme se, zda několik náhodných veličin spolu nějak souvisí, zda lze ze známé hodnoty jedné náhodné veličiny něco říci o rozdělení jiné náhodné veličiny. Snažíme se tedy vyšetřovat *závislost*. K tomu musíme mít odpovídající matematický model, který umožňuje pracovat nejen s rozděleními jednotlivých náhodných veličin, ale především s několika náhodnými veličinami *současně*.

Rozšiřme na mnohorozměrnou funkci $\boldsymbol{X}(\omega) = (X_1(\omega), \ldots, X_n(\omega))^T$ označení zavedené v (4.1):

$$\begin{aligned} [\boldsymbol{X} < \mathsf{x}] &= \{\omega \in \Omega : X_1(\omega) < x_1, \ldots, X_n(\omega) < x_n\} \\ [\mathsf{a} < \boldsymbol{X} < \mathsf{b}] &= \{\omega \in \Omega : a_1 < X_1(\omega) < b_1, \ldots, a_n < X_n(\omega) < b_n\} \end{aligned}$$

Jako $\mathcal{B}_n$ označme nejmenší σ-algebru nad intervaly tvaru

$$(-\infty, x_1) \times (-\infty, x_2) \times \cdots \times (-\infty, x_n),$$

pro všechny vektory $\mathsf{x} \in \mathbb{R}_n$. Nyní můžeme rozšířit definici náhodné veličiny také na náhodný vektor.

Definice 5.1. Mějme pravděpodobnostní prostor $(\Omega, \mathcal{A}, \mathsf{P})$. Reálná vektorová funkce $\boldsymbol{X}(\omega) = (X_1(\omega), \ldots, X_n(\omega))^T$ definovaná na Ω, pro kterou platí

$$\mathsf{x} \in \mathbb{R}_n \Rightarrow [\boldsymbol{X} < \mathsf{x}] \in \mathcal{A},$$

se nazývá **náhodný vektor**.

Místo náhodný vektor se někdy používá označení vektor náhodných veličin nebo vektorová náhodná veličina.

Náhodný jev $[\boldsymbol{X} < \mathsf{x}]$ lze psát také jako $\cap_{i=1}^{n}[X_i < x_i]$. Zvolíme-li pevné j z hodnot $1, \ldots, n$, bude nutně limita monotonní posloupnosti náhodných jevů

$$\lim_{x_j \to \infty} \bigcap_{i=1}^{n} [X_i < x_i] = \bigcap_{i \neq j} [X_i < x_i]$$

opět prvkem σ-algebry $\mathcal{B}$. To znamená, že $n-1$ rozměrný vektor, vzniklý vyloučením j-té složky z náhodného vektoru $\boldsymbol{X}$, je opět náhodným vektorem. Navíc, zapíšeme-li poslední vztah pomocí distribuční funkce, dostaneme

$$\begin{aligned} &\lim_{x_j \to \infty} F_{X_1, \ldots, X_n}(x_1, \ldots, x_j, \ldots, x_n) \\ &\qquad = F_{X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n) \end{aligned}$$

5.1 Diskrétní rozdělení

Jestliže náhodný vektor $\boldsymbol{X}$ nabývá pouze nejvýše spočetně mnoha hodnot, jde o diskrétní rozdělení. Podobně jako jsme diskrétní rozdělení náhodné veličiny popsali úplně seznamem hodnot, kterých veličina nabývá, a seznamem pravděpodobností těchto hodnot, popíšeme diskrétní rozdělení například dvourozměrného náhodného vektoru $(X,Y)^T$ pomocí seznamů hodnot

$$x_1, x_2, \ldots, \quad y_1, y_2, \ldots$$

a pravděpodobností všech možných dvojic těchto hodnot

$$\mathsf{P}[X = x_i, Y = y_j], \qquad i = 1, 2, \ldots, \quad j = 1, 2, \ldots,$$

kde

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \mathsf{P}[X = x_i, Y = y_j] = 1.$$

Předpokládáme při tom, že pro každou dvojici $i \neq j$ platí $x_i \neq x_j$ a $y_i \neq y_j$. Pokud některá z náhodných veličin X, Y nabývá pouze konečně mnoha hodnot, vystačíme s konečným seznamem a konečným součtem.

Nyní se věnujme souvislosti rozdělení náhodného vektoru $(X,Y)^T$ s rozděleními náhodných veličin X, Y. Náhodný jev $[X = x_i]$ můžeme vyjádřit jako $\cup_{j=1}^{\infty}[X = x_i, Y = y_j]$. Protože pro $k \neq j$ jsou náhodné jevy $[X = x_i, Y = y_k]$, $[X = x_i, Y = y_j]$ neslučitelné a náhodné jevy $[Y = y_1]$, $[Y = y_2], \ldots$ tvoří úplný systém jevů, platí

$$\mathsf{P}[X = x_i] = \sum_{j=1}^{\infty} \mathsf{P}[X = x_i, Y = y_j], \qquad i = 1, 2, \ldots. \tag{5.1}$$

Vztah (5.1) ukazuje vztah mezi **sdruženým** rozdělením náhodného vektoru $(X,Y)^T$ a **marginálním** rozdělením náhodné veličiny X. Podobně můžeme psát

$$\mathsf{P}[Y = y_j] = \sum_{i=1}^{\infty} \mathsf{P}[X = x_i, Y = y_j], \qquad j = 1, 2, \ldots. \tag{5.2}$$

Příklad 5.1. Uvažujme výsledné známky, které ze dvou předmětů získali studenti, kteří na konci semestru prospěli. Pro jednotlivé kombinace známek jsou pravděpodobnosti (na základě dlouholetých zkušeností expertem určené) uvedeny v tabulce 5.1. Všimněme si vztahu mezi sdruženým a marginálním rozdělením, například pravděpodobnost $\mathsf{P}[X = 1]$ dostaneme jako součet 0,15+0,10+0,05=0,30. ○

X	Y			
	1	2	3	marginální X
1	0,15	0,10	0,05	0,30
2	0,10	0,20	0,05	0,35
3	0,05	0,10	0,20	0,35
marginální Y	0,30	0,40	0,30	

Tabulka 5.1: Sdružené a marginální rozdělení pravděpodobností náhodných veličin X a Y

5.2 Spojité rozdělení

Pro **spojité rozdělení** můžeme provést podobnou úvahu. **Sdružená** distribuční funkce $F_{X,Y}$ náhodného vektoru $(X,Y)^T$ je dána předpisem

$$F_{X,Y}(x,y) = \mathsf{P}[X < x, Y < y]$$

a sdružená hustota $f_{X,Y}$ je pak takovou nezápornou funkcí, která pro každou dvojici reálných čísel x, y vyhovuje vztahu

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v)\mathrm{d}v\mathrm{d}u.$$

Všude, kde má sdružená distribuční funkce derivaci, platí

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y).$$

Protože náhodný jev $[X < x]$ můžeme pro libovolné reálné x vyjádřit (jak jsme viděli v úvodu této kapitoly) jako limitu monotonní posloupnosti náhodných jevů $\lim_{y\to\infty}[X < x, Y < y]$, dostáváme pro distribuční funkci náhodné veličiny X vztah

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x,y). \tag{5.3}$$

Vyjádříme-li tyto distribuční funkce pomocí hustot, dostaneme

$$\begin{aligned} \int_{-\infty}^{x} f_X(u)\mathrm{d}u &= \lim_{y\to\infty} \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v)\mathrm{d}v\mathrm{d}u \\ &= \int_{-\infty}^{x} \left(\int_{-\infty}^{\infty} f_{X,Y}(u,v)\mathrm{d}v \right) \mathrm{d}u. \end{aligned}$$

Je tedy

$$\int_{-\infty}^{\infty} f_{X,Y}(x,v)\mathrm{d}v \tag{5.4}$$

marginální hustota náhodné veličiny X.

Podobně jsou

$$F_Y(y) = \lim_{x\to\infty} F_{X,Y}(x,y),$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(u,y)\mathrm{d}u$$

distribuční funkce a hustota náhodné veličiny Y.

5.3 Nezávislost náhodných veličin

Pojem nezávislosti náhodných jevů jsme již zavedli. Je přirozené rozšířit jej na nezávislost náhodných veličin tak, aby náhodné veličiny X, Y byly nezávislé, právě když jsou nezávislé všechny náhodné jevy A_X a B_Y, kde A_X je tvrzení (náhodný jev) o náhodné veličině X, podobně A_Y je tvrzení (náhodný jev) o náhodné veličině Y. K tomu stačí, aby byly nezávislé náhodné jevy $[X < x]$ a $[Y < y]$ pro všechna reálná x, y.

Definice 5.2. Řekneme, že náhodné veličiny $X_1, X_2, \ldots$ jsou **nezávislé**, jestliže pro všechna $x_1, x_2, \ldots$ jsou nezávislé náhodné jevy $[X_1 < x_1]$, $[X_2 < x_2], \ldots$. Řekneme, že náhodné veličiny $X_1, X_2, \ldots$ jsou **po dvou nezávislé**, jestliže pro všechna $x_1, x_2, \ldots$ a pro libovolnou dvojici indexů $i \neq j$ jsou nezávislé náhodné jevy $[X_i < x_i], [X_j < x_j]$.

Speciálně, náhodné veličiny $X_1, \ldots, X_n$ jsou nezávislé, právě když platí

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) \equiv F_{X_1}(x_1)\cdots F_{X_n}(x_n). \tag{5.5}$$

Prakticky užitečná bude konkretizace pojmu nezávislosti náhodných veličin na případ sdruženého diskrétního a sdruženého spojitého rozdělení.

Mají-li náhodné veličiny X, Y diskrétní rozdělení, pak je definice nezávislosti totožná s požadavkem

$$\mathsf{P}[X = x_i, Y = y_j] = \mathsf{P}[X = x_i]\mathsf{P}[Y = y_j] \tag{5.6}$$

pro všechna i a j.

Příklad 5.2. V příkladu 5.1 uvedené náhodné veličiny X, Y jsou nutně závislé, neboť například je $\mathsf{P}[X = 1, Y = 1] = 0{,}15 \neq \mathsf{P}[X = 1]\mathsf{P}[Y = 1] = 0{,}30 \cdot 0{,}30$. ○

Podobně pro spojité rozdělení má požadavek na distribuční funkce za následek vztah mezi sdruženou a marginálními hustotami:

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y)$$

$$= \frac{\partial^2}{\partial x \partial y} F_X(x)F_y(y)$$
$$= \frac{\partial}{\partial y} f_X(x)F_Y(y)$$
$$= f_X(x)f_Y(y), \tag{5.7}$$

který silně připomíná vztah (5.6) pro nezávislé diskrétní náhodné veličiny.

Příklad 5.3. Uvažujme hod dvěma hracími kostkami. Jako elementární jev zvolíme dvojici (u, v), kde u je počet ok na první kostce, v je počet ok na druhé kostce. Protože bereme v úvahu uspořádané dvojice (u, v) (rozlišujeme mezi oběma kostkami), je $|\Omega| = 6^2$. Jednotlivé elementární jevy jsou stejně pravděpodobné. Zaveďme dvě náhodné veličiny: $X = u + v$ je celkový počet ok na obou kostkách, $Y = |u - v|$ je absolutní hodnota rozdílu těchto počtů. Sdružené rozdělení náhodných veličin X, Y ukazuje tabulka 5.2, v níž

Y \ X	2	3	4	5	6	7	8	9	10	11	12	marginální Y
0	1	0	1	0	1	0	1	0	1	0	1	6
1	0	2	0	2	0	2	0	2	0	2	0	10
2	0	0	2	0	2	0	2	0	2	0	0	8
3	0	0	0	2	0	2	0	2	0	0	0	6
4	0	0	0	0	2	0	2	0	0	0	0	4
5	0	0	0	0	0	2	0	0	0	0	0	2
marginální X	1	2	3	4	5	6	5	4	3	2	1	36

Tabulka 5.2: 36-násobky pravděpodobností v příkladu 5.3

jsou uvedeny pro úsporu místa 36-násobky pravděpodobností jednotlivých náhodných jevů. Nepřehlédněte marginální rozdělení v posledním pravém sloupci a ve spodním řádku. Z tabulky je zřejmé, že náhodné veličiny X, Y nejsou nezávislé, neboť je například

$$\mathsf{P}[X = 7, Y = 1] = \frac{2}{36} \neq \mathsf{P}[X = 7]\,\mathsf{P}[Y = 1] = \frac{6}{36}\frac{10}{36}.$$

○

5.4 Cvičení

5.1. Dvojice součástek má dobu života popsánu sdruženou hustotou

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{2}\mathrm{e}^{-x-y/2} & \text{pro } x > 0, y > 0, \\ 0 & \text{jinak.} \end{cases}$$

a) Jaká je pravděpodobnost toho, že druhá součástka přežije první?
b) Jaká je pravděpodobnost toho, že první součástka alespoň dvakrát přežije druhou součástku?
c) Určete distribuční funkci náhodné veličiny $Z = X + Y$.
d) Určete distribuční funkci náhodné veličiny $W = X - Y$.

5.2. Nechť X, Y jsou nezávislé náhodné veličiny s rovnoměrným rozdělením na intervalu $(0, 1)$. Najděte distribuční funkci a hustotu náhodné veličiny $Z = X + Y$.

6. Střední hodnota

Nyní se pokusíme o „nemožné", totiž o charakterizování náhodné veličiny X pomocí jediného čísla, které budeme značit $\mathsf{E}X$.

6.1 Diskrétní rozdělení

Vezměme náhodnou veličinu X s rozdělením určeným vztahem

$$\mathsf{P}[X = k] = \frac{1}{n}, \qquad k = 1, 2, \ldots, n.$$

Příkladem takové náhodné veličiny je počet ok na symetrické hrací kostce náhodně vrhané na pevnou podložku ($n = 6$). Průměr z možných hodnot náhodné veličiny X, totiž $(n + 1)/2$, jistě možné hodnoty charakterizuje docela dobře. Je však nepříjemné, že při sudém n této průměrné hodnoty náhodná veličina X vůbec nabývat nemůže.

V příkladu 4.1 jsme zavedli binomické rozdělení. Zde se průměr z možných hodnot, totiž $n/2$, k našemu účelu nehodí, protože vůbec nezávisí na parametru p. Můžeme však využít naší znalosti pravděpodobností jednotlivých hodnot a určit *vážený* průměr

$$\sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}.$$

Tuto myšlenku můžeme použít také pro negativně binomické rozdělení z příkladu 4.5, kdy náhodná veličina nabývá nekonečně mnoha hodnot. Použijme

$$\sum_{k=0}^{\infty} k \; \mathsf{P}[X = k], \tag{6.1}$$

nebo obecněji

$$\sum_{i=1}^{\infty} x_i \; \mathsf{P}[X = x_i], \tag{6.2}$$

kde $x_1, x_2, \ldots$ je nejvýše spočetná posloupnost navzájem různých možných hodnot náhodné veličiny X.

U vzorce (6.1) nehrozí problém, který by mohl nastat u obecnějšího vyjádření (6.2) v případě, že by náhodná veličina nabývala nejen kladných, ale i záporných hodnot, kde bychom po přerovnání členů řady (po přečíslování možných hodnot náhodných veličin) mohli dostat jiný součet. Proto

budeme obecně požadovat *absolutní* konvergenci řady (6.2), která zajistí jedinou hodnotu součtu (6.2), nezávislou na pořadí sčítanců.

Definice 6.1. Nechť X je náhodná veličina s diskrétním rozdělením, která nabývá hodnot $x_1, x_2, \ldots$. Jestliže řada (6.2) konverguje absolutně, označíme její součet symbolem $\mathsf{E}X$ a nazveme jej **střední hodnotou** náhodné veličiny X. Pokud řada (6.2) nekonverguje absolutně, pak říkáme, že náhodná veličina X nemá střední hodnotu.

Příklad 6.1. Vraťme se k binomickému rozdělení z příkladu 4.1. Střední hodnotu můžeme spočítat prostým dosazením do definice 6.1:

$$\begin{aligned}
\mathsf{E}X &= \sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k} \\
&= np\sum_{k=1}^{n}\frac{(n-1)!}{(n-k)!(k-1)!}p^{k-1}(1-p)^{n-k} \\
&= np\sum_{j=0}^{n-1}\frac{(n-1)!}{(n-1-j)!j!}p^{j}(1-p)^{n-1-k} \\
&= np.
\end{aligned}$$

Nejprve jsme ze součtu vyloučili nulový sčítanec, pak jsme sčítací index k nahradili indexem $j = k-1$ a nakonec jsme použili skutečnost, že součet pravděpodobností v binomickém rozdělení s parametry $n-1, p$ je roven 1.○

Zkusme nyní upravit pojem střední hodnoty pro funkci náhodné veličiny. Mějme náhodnou veličinu X s diskrétním rozdělením, která nabývá hodnot $x_1, x_2, \ldots$. Pro každou reálnou funkci g je $Y = g(X)$ náhodnou veličinou. Označme jako $y_1, y_2, \ldots$ navzájem *různé* hodnoty z hodnot $g(x_1), g(x_2), \ldots$. Rozdělení náhodné veličiny Y dostaneme pomocí

$$\begin{aligned}
\mathsf{P}[Y = y_j] &= \mathsf{P}\{\omega \in \Omega : g(X(\omega)) = y_j\} \\
&= \sum_{i:g(x_i)=y_j} \mathsf{P}[X = x_i]. \qquad (6.3)
\end{aligned}$$

Střední hodnotu náhodné veličiny Y (pokud existuje a tudíž nezáleží na pořadí sčítanců) můžeme spočítat:

$$\begin{aligned}
\mathsf{E}g(X) &= \sum_{j=1}^{\infty} y_j \mathsf{P}[Y = y_j] \\
&= \sum_{j=1}^{\infty} y_j \sum_{i:g(x_i)=y_j} \mathsf{P}[X = x_i] \\
&= \sum_{j=1}^{\infty} \sum_{i:g(x_i)=y_j} g(x_i)\mathsf{P}[X = x_i] \\
&= \sum_{i=1}^{\infty} g(x_i)\mathsf{P}[X = x_i]. \qquad (6.4)
\end{aligned}$$

Poslední způsob výpočtu střední hodnoty $\mathsf{E}g(X)$ je zpravidla nejsnazší.

6.2 Spojité rozdělení

Definice střední hodnoty náhodné veličiny musí i v případě spojité náhodné veličiny mít stejnou interpretaci. Opět musí nějak *vážit* možné hodnoty náhodné veličiny. Místo pravděpodobností jednotlivých hodnot (které jsou vesměs nulové!) použijeme ve shodě s (4.20) hustotu rozdělení.

Definice 6.2. Nechť náhodná veličina X má spojité rozdělení s hustotou $f_X(x)$. Pokud je

$$\int_{-\infty}^{\infty} |x| f_X(x)\mathrm{d}x < \infty,$$

nazveme integrál

$$\int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x \qquad (6.5)$$

střední hodnotou náhodné veličiny X, v opačném případě řekneme, že střední hodnota neexistuje.

Podobně jako u diskrétního rozdělení definujeme střední hodnotu funkce $g(x)$ náhodné veličiny X (tedy střední hodnotu náhodné veličiny $g(X)$) pomocí vztahu

$$\mathsf{E}g(X) = \int_{-\infty}^{\infty} g(x) f_X(x)\mathrm{d}x, \qquad (6.6)$$

pokud je

$$\int_{-\infty}^{\infty} |g(x)| f_X(x)\mathrm{d}x < \infty.$$

Příklad 6.2. Spočítáme střední hodnotu náhodné veličiny s rovnoměrným rozdělením, jehož hustota je uvedena v příkladu 4.7. Podle (6.5) dostaneme

$$\begin{aligned}
\mathsf{E}X &= \int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x \\
&= \int_a^b x\frac{1}{b-a}\mathrm{d}x
\end{aligned}$$

$$= \frac{b+a}{2}. \qquad \bigcirc$$

Příklad 6.3. Spočítejme střední hodnotu exponenciálního rozdělení $\mathrm{ex}(\lambda)$. Vzhledem k tomu, že hustota je v tomto případě kladná pouze pro kladný argument a použijeme-li metodu per partes, můžeme psát

$$\begin{aligned} \mathsf{E}X &= \int_0^\infty t\lambda \mathrm{e}^{-\lambda t}\mathrm{d}t \\ &= \left[-t\mathrm{e}^{-\lambda t}\right]_0^\infty + \int_0^\infty \mathrm{e}^{-\lambda t}\mathrm{d}t \\ &= \frac{1}{\lambda}\int_{-\infty}^\infty f_X(t)\mathrm{d}t \\ &= \frac{1}{\lambda}. \qquad \bigcirc \end{aligned}$$

6.3 Poznámka

Dříve, než se budeme věnovat vlastnostem střední hodnoty, pokusíme se ukázat, že definice 6.1 a 6.2 jsou jen speciálními příklady jediného pojmu. Informovaný čtenář jistě vytuší Lebesgueův integrál, ostatním se pokusíme základní myšlenku přiblížit.

Uvažujme nejprve náhodnou veličinu, která nabývá pouze kladných hodnot, tedy platí $\mathsf{P}[X > 0] = 1$ a jejíž střední hodnota existuje. Pokusíme se tuto střední hodnotu aproximovat obdobou váženého průměru pomocí

$$\sum_{k=1}^{n^2} \frac{k-1}{n}\mathsf{P}\left[\frac{k-1}{n} \le X < \frac{k}{n}\right]. \tag{6.7}$$

Vlastně tak nahrazujeme funkci $X(\omega)$ jednoduchou funkcí, která nabývá na množině $\left[\frac{k-1}{n} \le X < \frac{k}{n}\right]$ konstantní hodnoty $\frac{(k-1)}{n}$. Samotná funkce $X(\omega)$ (tedy náhodná veličina) je pak limitou těchto jednoduchých funkcí. Představme si nyní zmíněnou aproximující funkci například v levém okolí $x = 1$. Pro $n = 10$ dostaneme pro $k = 10$ aproximaci $(k-1)/n = 0{,}9$, kdežto pro $n = 11$ to bude $(k-1)/n = 10/11 = 0{,}90909\ldots$. Je zřejmé, že aproximující funkce je neklesající funkcí n. Tato monotonie aproximujících funkcí umožňuje zapsat střední hodnotu jako limitu výrazů (6.7). Obecnou

náhodnou veličinu (funkci $X(\omega)$) pak zapíšeme jako rozdíl dvou nezáporných náhodných veličin (nezáporných funkcí) $X = X^+ - X^-$, spočítáme střední hodnoty z funkcí X^+, X^- a výpočet dokončíme (pokud střední hodnota existuje) pomocí $\mathsf{E}X = \mathsf{E}X^+ - \mathsf{E}X^-$.

6.4 Vlastnosti střední hodnoty

Vedle náhodné veličiny jsme zavedli také náhodný vektor jako vektor náhodných veličin. Je tedy přirozené používat pojem střední hodnoty také pro náhodný vektor. Jako **střední hodnotu náhodného vektoru** použijeme vektor středních hodnot jednotlivých náhodných veličin tvořících tento vektor.

Pokud se v dále uvedených tvrzeních vyskytuje více náhodných veličin, vždy budeme předpokládat, že jsou definovány na stejném pravděpodobnostním prostoru, pokud se vyskytují střední hodnoty, vždy předpokládáme jejich existenci.

Věta 6.1. Nechť X je náhodná veličina, jejíž střední hodnota existuje, nechť $a, b \in \mathbb{R}$. Potom platí

$$\mathsf{E}a = a, \tag{6.8}$$

$$\mathsf{E}(a + bX) = a + b\mathsf{E}X. \tag{6.9}$$

D ů k a z: Konstantu a můžeme chápat jako náhodnou veličinu s diskrétním rozdělením, která nabývá pouze jediné hodnoty. Podle (6.2) má tedy střední hodnotu rovnou této jediné možné hodnotě a platí (6.8). K důkazu (6.9) použijeme vzorec pro střední hodnotu funkce náhodné veličiny. Například v případě spojitého rozdělení tak dostaneme posloupnost rovností

$$\begin{aligned} \mathsf{E}(a + bX) &= \int_{-\infty}^\infty (a + bx) f_X(x)\mathrm{d}x \\ &= a\int_{-\infty}^\infty f_X(x)\mathrm{d}x + b\int_{-\infty}^\infty x f_X(x)\mathrm{d}x \\ &= a \cdot 1 + b\mathsf{E}X. \end{aligned}$$

Pro diskrétní rozdělení je postup podobný. □

Věta 6.2. Platí

$$\mathsf{E}(X + Y) = \mathsf{E}X + \mathsf{E}Y, \tag{6.10}$$

jestliže střední hodnoty na pravé straně existují. Podobně, pro dva náhodné vektory stejné dimenze v případě, že střední hodnoty všech složek na pravé

straně existují, platí

(6.11) $$\mathsf{E}(\boldsymbol{X}+\boldsymbol{Y}) = \mathsf{E}\boldsymbol{X} + \mathsf{E}\boldsymbol{Y}.$$

D ů k a z: Tentokrát budeme předpokládat diskrétní rozdělení. Jestliže náhodné veličiny X, Y mají diskrétní rozdělení s hodnotami $x_1, x_2, \ldots$, resp. $y_1, y_2, \ldots$, střední hodnotu funkce $X+Y$ zapsanou podle (6.4) lze postupně s použitím vztahu mezi sdruženým a marginálním rozdělením (5.1) a (5.2) upravit na (nezáleží na pořadí sčítanců)

$$\begin{aligned}
\mathsf{E}(X+Y) &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(x_i+y_j)\mathsf{P}[X=x_i, Y=y_j] \\
&= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}x_i\mathsf{P}[X=x_i, Y=y_j] + \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}y_j\mathsf{P}[X=x_i, Y=y_j] \\
&= \sum_{i=1}^{\infty}x_i\sum_{j=1}^{\infty}\mathsf{P}[X=x_i, Y=y_j] + \sum_{j=1}^{\infty}y_j\sum_{i=1}^{\infty}\mathsf{P}[X=x_i, Y=y_j] \\
&= \sum_{i=1}^{\infty}x_i\mathsf{P}[X=x_i] + \sum_{j=1}^{\infty}y_j\mathsf{P}[Y=y_j].
\end{aligned}$$

Pro spojité rozdělení je důkaz analogický. Ještě k existenci střední hodnoty na levé straně. Ta plyne z nerovnosti $|X+Y| \le |X| + |Y|$ a z předpokladu existence středních hodnot na pravé straně. □

Nyní můžeme rozšířit tvrzení (6.9) věty 6.1 na náhodný vektor.

Věta 6.3. Nechť jsou $\mathbf{a}, \mathbf{B}$ vektor a matice konstant odpovídajících rozměrů a $\boldsymbol{X}$ náhodný vektor, pro jehož složky existují střední hodnoty. Potom platí

(6.12) $$\mathsf{E}(\mathbf{a}+\mathbf{B}\boldsymbol{X}) = \mathbf{a} + \mathbf{B}\mathsf{E}\boldsymbol{X}.$$

D ů k a z: Stačí ověřit tvrzení pro j-tou složku náhodného vektoru $\mathbf{a}+\mathbf{B}\boldsymbol{X}$. Použijeme-li vlastnosti (6.8), (6.11) střední hodnoty coby lineárního operátoru, dostaneme

$$\mathsf{E}(a_j + \sum_{i=1}^{n} b_{ji}X_i) = a_j + \sum_{i=1}^{n} b_{ji}\mathsf{E}X_i,$$

což je právě j-tá složka vektoru na pravé straně (6.12). □

Příklad 6.4. Střední hodnotu binomického rozdělení jsme určili v příkladu 6.1. Jinou možností pro výpočet této střední hodnoty je představit si

počet zdarů v n nezávislých pokusech jako součet počtů zdarů v jednotlivých pokusech ve tvaru

$$X = \sum_{i=1}^{n} Y_i,$$

kde Y_i má binomické rozdělení s parametry $1, p$ (tzv. alternativní rozdělení). Střední hodnota veličiny Y_i se počítá snadno:

$$\mathsf{E}Y_i = 0\cdot p + 1\cdot(1-p) = p.$$

Střední hodnota součtu je podle věty 6.2 rovna součtu středních hodnot, takže máme

$$\mathsf{E}X = \sum_{k=1}^{n}\mathsf{E}Y_i = np. \qquad \bigcirc$$

Předchozí dvě věty popisují linearitu střední hodnoty jako lineárního operátoru. Pro *nezávislé* náhodné veličiny platí podobný vztah také pro součin.

Věta 6.4. Jsou-li náhodné veličiny X, Y nezávislé, pak platí

(6.13) $$\mathsf{E}(XY) = \mathsf{E}X\mathsf{E}Y,$$

jestliže střední hodnoty na pravé straně existují.

D ů k a z: Zůstaňme opět u diskrétního rozdělení. S využitím vlastnosti nezávislých náhodných veličin

$$\mathsf{P}[X=x_i, Y=y_j] = \mathsf{P}[X=x_i]\mathsf{P}[Y=y_j], \qquad i=1,2,\ldots, \quad j=1,2,\ldots.$$

postupně upravíme střední hodnotu součinu XY

$$\begin{aligned}
\mathsf{E}XY &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(x_iy_j)\mathsf{P}[X=x_i, Y=y_j] \\
&= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(x_iy_j)\mathsf{P}[X=x_i]\mathsf{P}[Y=y_j] \\
&= \sum_{i=1}^{\infty}x_i\mathsf{P}[X=x_i]\sum_{j=1}^{\infty}y_j\mathsf{P}[Y=y_j] \\
&= \mathsf{E}X\mathsf{E}Y. \qquad \square
\end{aligned}$$

6.5 Cvičení

6.1. V daném roce se dividendy nebudou vyplácet s pravděpodobností 1/8, kdežto dividendy v hodnotě 2^j jednotek se budou vyplácet s pravděpodobností 2^{-j}, $j = 1, 2, 3$. Jakou výplatu lze v průměru očekávat?

6.2. Ověřte, zda vztah

$$\mathsf{P}(X = j) = \frac{j}{55}, \qquad j = 1, \ldots, 10,$$

určuje rozdělení náhodné veličiny X a spočítejte její střední hodnotu.

6.3. Spočítejte střední hodnotu náhodné veličiny Y s Poissonovým rozdělením, jejíž rozdělení je dáno vztahem

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

kde λ je kladný parametr.

6.4. Najděte střední hodnoty pro náhodné veličiny dané hustotami

a) $f(x) = 3x^2$ pro $0 < x < 1$,
b) $f(x) = 4x^3$ pro $0 < x < 1$,
c) $f(x) = \sin x$ pro $0 < x < \pi/2$,
d) $f(x) = \frac{1}{2}\sin x$ pro $0 < x < \pi$.

6.5. Hodíme bílou, zelenou a modrou hrací kostkou a číselná vyjádření výsledků označíme b, z, m. Zaveďme náhodnou veličinu

$$Y = \begin{cases} 0 & \text{je-li } b \neq z, b \neq m, z \neq m, \\ 1 & \text{platí-li právě jedna z rovností } b = z, b = m, z = m, \\ 2 & \text{je-li } b = z = m. \end{cases}$$

Určete rozdělení náhodné veličiny Y a její střední hodnotu.

6.6. Tři dorostenci kopou po jednom pokutovém kopu. První bude úspěšný s pravděpodobností 0,8, druhý s pravděpodobností 0,7 a třetí s pravděpodobností 0,9. Určete rozdělení pravděpodobnosti celkového počtu vstřelených branek a jeho střední hodnotu.

6.7. V loterii je m_i výher s hodnotou q_i, $i = 1, \ldots, k$. Má být vydáno N losů. Určete cenu losu tak, aby střední hodnota výhry na jeden los byla rovna polovině jeho ceny.

6.8. Dělník obsluhuje n strojů téhož typu, které jsou umístěny vedle sebe ve vzdálenostech a. Když skončí obsluhu stroje, přejde k tomu stroji, který v té chvíli nejdéle čeká na obsluhu, případně počká na první požadavek. Předpokládá se, že závada kteréhokoliv stroje je stejně pravděpodobná a že závady jsou nezávislé. Určete střední hodnotu dělníkovy cesty od stroje ke stroji.

6.9. Náhodná veličina nabývá hodnoty k, $k = 1, 2, \ldots$, s pravděpodobností, která je úměrná 3^{-k}. Určete střední hodnotu takové náhodné veličiny.

6.10. Náhodná veličina X má distribuční funkci

$$F(x) = \begin{cases} 0 & \text{pro } x \leq -1, \\ a + b\arcsin x & \text{pro } -1 < x < 1, \\ 1 & \text{pro } x \geq 1. \end{cases}$$

Určete konstanty a, b a střední hodnotu náhodné veličiny X.

7. Další charakteristiky

7.1 Rozptyl

Další užívanou charakteristikou rozdělení náhodné veličiny je rozptyl. Popisuje jinou vlastnost rozdělení – nikoliv nějakou střední či typickou hodnotu, ale velikost kolísání náhodné veličiny kolem střední hodnoty.

Definice 7.1. Nechť X je náhodná veličina s konečnou střední hodnotou. Potom výraz

$$\mathrm{var}X = \mathrm{E}(X - \mathrm{E}X)^2,$$

pokud střední hodnota na pravé straně existuje, se nazývá **rozptyl** náhodné veličiny X. Odmocnina z rozptylu ($\sqrt{\mathrm{var}X}$) se nazývá **směrodatná odchylka** náhodné veličiny.

Všude dále budeme existenci $\mathrm{var}X$ předpokládat. Místo rozptyl se někdy říká **variance**.

Věta 7.1. Pro náhodnou veličinu X s konečným rozptylem a libovolná reálná čísla a, b platí

$$\mathrm{var}X = \mathrm{E}X^2 - (\mathrm{E}X)^2, \tag{7.1}$$

$$\mathrm{var}(a + bX) = b^2\mathrm{var}X, \tag{7.2}$$

$$\sqrt{\mathrm{var}(a + bX)} = |b|\sqrt{\mathrm{var}X}. \tag{7.3}$$

D ů k a z: Nejprve dokažme tvrzení (7.2). Využijme při tom tvrzení věty 6.1. Dostaneme

$$\begin{aligned}\mathrm{var}(a + bX) &= \mathrm{E}\left((a + bX) - \mathrm{E}(a + bX)\right)^2 \\ &= \mathrm{E}\left(b(X - \mathrm{E}X)\right)^2 \\ &= b^2\mathrm{E}(X - \mathrm{E}X)^2 \\ &= b^2\mathrm{var}X.\end{aligned}$$

Nyní již snadno dokážeme první tvrzení věty:

$$\begin{aligned}\mathrm{var}X &= \mathrm{E}(X - \mathrm{E}X)^2 \\ &= \mathrm{E}(X^2 - 2X\mathrm{E}X + (\mathrm{E}X)^2) \\ &= \mathrm{E}X^2 - 2(\mathrm{E}X)^2 + (\mathrm{E}X)^2 \\ &= \mathrm{E}X^2 - (\mathrm{E}X)^2.\end{aligned}$$

Třetí tvrzení je triviálním důsledkem vztahu (7.2). □

Vztah (7.1) je velmi často pro výpočet rozptylu výhodnější, než samotná definice.

Někdy se k dané náhodné veličině s konečným nenulovým rozptylem (a tedy s konečnou a nenulovou směrodatnou odchylkou) hledá **normovaná** (standardizovaná) veličina, daná předpisem

$$Z = \frac{X - \mathrm{E}X}{\sqrt{\mathrm{var}X}}. \tag{7.4}$$

Z tvrzení (6.9) a (7.2) snadno plyne, že normovaná náhodná veličina má nulovou střední hodnotu a jednotkový rozptyl.

Příklad 7.1. Určíme střední hodnotu a rozptyl náhodné veličiny X, která má geometrické rozdělení (X je počet nezdarů před prvním zdarem v posloupnosti nezávislých pokusů s pravděpodobností zdaru p a nezdaru $1 - p$) s parametrem $p \in (0, 1)$, tj. $\mathrm{P}[X = k] = (1 - p)^k p$ pro $k = 0, 1, \ldots$ (viz příklad 4.5). Pro $p \in (0, 1)$ platí $p^{-1} = \sum_{k=0}^{\infty}(1 - p)^k$. Tuto řadu můžeme v intervalu $(0, 1)$ derivovat člen po členu (připomeňte si příslušné tvrzení z matematické analýzy) a dostaneme postupně

$$\begin{aligned}-p^{-2} &= -\sum_{k=0}^{\infty} k(1 - p)^{k-1}, \quad p \in (0, 1), \\ \frac{1 - p}{p} &= \sum_{k=0}^{\infty} k(1 - p)^k p, \quad p \in (0, 1), \\ \mathrm{E}X &= \frac{1}{p} - 1.\end{aligned}$$

Odtud derivováním člen po členu vyjde

$$\begin{aligned}\frac{1 - p}{p^2} &= \sum_{k=0}^{\infty} k(1 - p)^k, \quad p \in (0, 1), \\ \frac{2p - p^2}{p^4} &= \sum_{k=0}^{\infty} k^2(1 - p)^{k-1} p, \quad p \in (0, 1), \\ \frac{(2 - p)(1 - p)}{p^2} &= \sum_{k=0}^{\infty} k^2(1 - p)^k p, \quad p \in (0, 1),\end{aligned}$$

takže hledané momenty jsou

$$\mathrm{E}X^2 = \frac{(2 - p)(1 - p)}{p^2},$$

$$\begin{aligned} \mathsf{var}X &= \mathsf{E}X^2 - (\mathsf{E}X)^2 \\ &= \frac{(2-p)(1-p)}{p^2} - \left(\frac{1-p}{p}\right)^2 = \frac{1-p}{p^2}. \end{aligned}$$ ○

7.2 Kovariance

O závislosti dvou náhodných veličin do jisté míry vypovídá následující pojem.

Definice 7.2. Nechť pro náhodné veličiny X, Y existují rozptyly. Výraz

$$\mathsf{cov}(X,Y) = \mathsf{E}(X - \mathsf{E}X)(Y - \mathsf{E}Y) \tag{7.5}$$

se nazývá **kovariance** náhodných veličin X, Y.

Zřejmě platí $\mathsf{cov}(X,Y) = \mathsf{cov}(Y,X)$ a $\mathsf{cov}(X,X) = \mathsf{var}X$. Požadavek na existenci rozptylů zajišťuje existenci střední hodnoty definující kovarianci.

Věta 7.2. Nechť X, Y jsou náhodné veličiny, pro něž existují rozptyly, nechť a, b, c, d jsou reálná čísla. Potom platí

$$\mathsf{cov}(X,Y) = \mathsf{E}(XY) - (\mathsf{E}X)(\mathsf{E}Y), \tag{7.6}$$

$$\mathsf{cov}(a + bX, c + dY) = b\,d\,\mathsf{cov}(X,Y). \tag{7.7}$$

D ů k a z: První vztah, který bývá užitečný při výpočtu kovariance, plyne z následujících úprav:

$$\begin{aligned} \mathsf{cov}(X,Y) &= \mathsf{E}(X - \mathsf{E}X)(Y - \mathsf{E}Y) \\ &= \mathsf{E}\left(XY - X\mathsf{E}Y - (\mathsf{E}X)Y + (\mathsf{E}X)(\mathsf{E}Y)\right) \\ &= \mathsf{E}XY - (\mathsf{E}X)(\mathsf{E}Y). \end{aligned}$$

Důkaz druhého vztahu je analogický důkazu vztahu (7.2). □

Věta 7.3. Pro náhodné veličiny X, Y, jejichž rozptyly existují, platí

$$\mathsf{var}(X+Y) = \mathsf{var}X + \mathsf{var}Y + 2\mathsf{cov}(X,Y). \tag{7.8}$$

Jsou-li X, Y navíc nezávislé, platí

$$\mathsf{var}(X+Y) = \mathsf{var}X + \mathsf{var}Y. \tag{7.9}$$

D ů k a z: Vztah (7.8) dostaneme po malé úpravě

$$\begin{aligned} \mathsf{var}(X+Y) &= \mathsf{E}\left((X+Y) - \mathsf{E}(X+Y)\right)^2 \\ &= \mathsf{E}\left((X - \mathsf{E}X) + (Y - \mathsf{E}Y)\right)^2 \\ &= \mathsf{E}\left((X - \mathsf{E}X)^2 + 2(X - \mathsf{E}X)(Y - \mathsf{E}Y) + (Y - \mathsf{E}Y)^2\right) \\ &= \mathsf{var}X + 2\mathsf{cov}(X,Y) + \mathsf{var}Y. \end{aligned}$$

Jsou-li navíc náhodné veličiny X, Y nezávislé, jsou nezávislé také náhodné veličiny $(X - \mathsf{E}X), (Y - \mathsf{E}Y)$, jak plyne z definice nezávislosti náhodných veličin. Potom je ovšem

$$\mathsf{E}(X - \mathsf{E}X)(Y - \mathsf{E}Y) = \mathsf{E}(X - \mathsf{E}X)\mathsf{E}(Y - \mathsf{E}Y) = 0,$$

což znamená, že kovariance $\mathsf{cov}(X,Y)$ je v tomto případě nulová a zřejmě platí (7.9). □

Speciálním případem kovariance je **korelační koeficient**, který dostaneme, když spočítáme kovarianci dvou *normovaných* náhodných veličin:

$$\begin{aligned} \rho_{X,Y} &= \mathsf{cov}\left(\frac{X - \mathsf{E}X}{\sqrt{\mathsf{var}X}}, \frac{Y - \mathsf{E}Y}{\sqrt{\mathsf{var}Y}}\right) \\ &= \frac{\mathsf{cov}(X,Y)}{\sqrt{\mathsf{var}X\,\mathsf{var}Y}}. \end{aligned} \tag{7.10}$$

Snadno se dokáže následující tvrzení:

Věta 7.4. Nechť X, Y jsou náhodné veličiny s konečnými nenulovými rozptyly, nechť $a, c, b \neq 0, d \neq 0$ jsou reálná čísla. Potom platí

$$\rho_{a+bX,c+dY} = \mathrm{sign}(bd)\rho_{X,Y}.$$

Důležité vlastnosti korelačního koeficientu uvádí následující tvrzení.

Věta 7.5. Je-li korelační koeficient náhodných veličin X, Y definován, pak platí

$$|\rho_{X,Y}| \leq 1, \tag{7.11}$$

$$\rho_{X,X} = 1. \tag{7.12}$$

Jsou-li náhodné veličiny X, Y nezávislé, platí

$$\rho_{X,Y} = 0. \tag{7.13}$$

D ů k a z: Protože pro všechna t platí

$$0 \le \mathsf{var}\left(\frac{X-\mathsf{E}X}{\sqrt{\mathsf{var}X}} + t\frac{Y-\mathsf{E}Y}{\sqrt{\mathsf{var}Y}}\right) = 1 + 2t\rho_{X,Y} + t^2,$$

musí být diskriminant kvadratické rovnice v t v pravé části posledního vztahu nekladný, tedy

$$4\rho_{X,Y} - 4 \le 0,$$

což je ekvivalentní s (7.11). Důkaz tvrzení (7.13) plyne z věty 6.4, tvrzení (7.12) je triviální. □

Při práci s náhodným vektorem $\boldsymbol{X} = (X_1, \cdots, X_n)^T$ se používá následující zobecnění pojmu rozptyl.

Definice 7.3. Nechť pro náhodné veličiny $X_1, \cdots, X_n$ existují rozptyly. Potom výraz

$$\mathsf{var}\boldsymbol{X} = \begin{pmatrix} \mathsf{var}X_1 & \mathsf{cov}(X_1,X_2) & \cdots & \mathsf{cov}(X_1,X_n) \\ \mathsf{cov}(X_2,X_1) & \mathsf{var}X_2 & \cdots & \mathsf{cov}(X_2,X_n) \\ & & \cdots & \\ \mathsf{cov}(X_n,X_1) & \mathsf{cov}(X_n,X_2) & \cdots & \mathsf{var}X_n \end{pmatrix} \tag{7.14}$$

se nazývá **varianční matice** náhodného vektoru $\boldsymbol{X}$.

Po rozepsání po složkách lze ověřit, že varianční matici lze zapsat ve tvaru

$$\mathsf{var}\boldsymbol{X} = \mathsf{E}(\boldsymbol{X} - \mathsf{E}\boldsymbol{X})(\boldsymbol{X} - \mathsf{E}\boldsymbol{X})^T. \tag{7.15}$$

Věta 7.6. Nechť $\boldsymbol{X}$ je n-rozměrný náhodný vektor, pro jehož složky existují rozptyly, nechť $\mathbf{a} \in \mathbb{R}^m$ je vektor konstant, nechť $\mathbf{B}$ je matice konstant typu $m \times n$. Potom platí

$$\mathsf{var}(\mathbf{a} + \mathbf{B}\boldsymbol{X}) = \mathbf{B}\,\mathsf{var}\boldsymbol{X}\,\mathbf{B}^T. \tag{7.16}$$

D ů k a z: Použijeme-li zápis varianční matice ve tvaru (7.15) a vlastnosti střední hodnoty náhodného vektoru (6.12), dostaneme postupně

$$\begin{aligned} \mathsf{var}(\mathbf{a}+\mathbf{B}\boldsymbol{X}) &= \mathsf{E}\left((\mathbf{a}+\mathbf{B}\boldsymbol{X}) - \mathsf{E}(\mathbf{a}+\mathbf{B}\boldsymbol{X})\right)\left((\mathbf{a}+\mathbf{B}\boldsymbol{X}) - \mathsf{E}(\mathbf{a}+\mathbf{B}\boldsymbol{X})\right)^T \\ &= \mathsf{E}\left(\mathbf{B}(\boldsymbol{X}-\mathsf{E}\boldsymbol{X})\right)\left(\mathbf{B}(\boldsymbol{X}-\mathsf{E}\boldsymbol{X})\right)^T \\ &= \mathbf{B}\,\mathsf{E}(\boldsymbol{X}-\mathsf{E}\boldsymbol{X})(\boldsymbol{X}-\mathsf{E}\boldsymbol{X})^T\mathbf{B}^T \\ &= \mathbf{B}\,\mathsf{var}\boldsymbol{X}\,\mathbf{B}^T \qquad \square \end{aligned}$$

Příklad 7.2. Vraťme se k příkladu 5.1, kde jsme uvažovali výsledné známky, které ze dvou předmětů získali studenti, kteří na konci semestru prospěli. Pro jednotlivé kombinace známek jsou pravděpodobnosti uvedeny v tabulce 5.1. Postupně lze spočítat

$$\begin{aligned} \mathsf{E}X &= 1\cdot 0{,}30 + 2\cdot 0{,}35 + 3\cdot 0{,}35 = 2{,}05, \\ \mathsf{E}X^2 &= 1^2\cdot 0{,}30 + 2^2\cdot 0{,}35 + 3^2\cdot 0{,}35 = 4{,}85, \end{aligned}$$

takže je $\mathsf{var}X = 4{,}85 - 2{,}05^2 = 0{,}6475$. Podobně dostaneme $\mathsf{E}Y = 2{,}0$, $\mathsf{var}Y = 0{,}6$. Pro výpočet kovariance potřebujeme

$$\begin{aligned} \mathsf{E}XY &= 1\cdot 1\cdot 0{,}15 + 1\cdot 2\cdot 0{,}10 + 1\cdot 3\cdot 0{,}05 + \\ &+ 2\cdot 1\cdot 0{,}10 + 2\cdot 2\cdot 0{,}20 + 2\cdot 3\cdot 0{,}05 + \\ &+ 3\cdot 1\cdot 0{,}05 + 3\cdot 2\cdot 0{,}10 + 3\cdot 3\cdot 0{,}20 = 4{,}35, \end{aligned}$$

takže je $\mathsf{cov}(X,Y) = 4{,}35 - 2{,}05\cdot 2{,}0 = 0{,}25$. Nakonec spočítáme korelační koeficient

$$\rho_{X,Y} = \frac{0{,}25}{\sqrt{0{,}6475\cdot 0{,}6}} \doteq 0{,}401,$$

což nám dokumentuje, že náhodné veličiny X, Y nemohou být nezávislé. ○

Následující příklady ukazují, že náhodný vektor, řekněme $(X,Y)^T$, jako stochastický objekt *není* jednoduše dvojicí náhodných veličin X a Y. Jeho pravděpodobnostní struktura, tj. jeho rozdělení pravděpodobnosti, není zdaleka definována marginálními rozděleními veličiny X a veličiny Y. Pravděpodobnostní vztah mezi X a Y (částečně popsaný například jejich korelačním koeficientem) je skutečná esence pojmu náhodný vektor.

Příklad 7.3. Buďte X a Y náhodné veličiny, které nabývají hodnot 0 a 1. Pak

$$\mathsf{P}[X=1, Y=1] - \mathsf{P}[X=1]\mathsf{P}[Y=1] = \mathsf{E}(XY) - \mathsf{E}X\mathsf{E}Y = \mathsf{cov}(X,Y).$$

Odtud plyne, použijeme-li tvrzení (a) věty 2.3, že nula-jedničkové veličiny X a Y jsou nezávislé právě tehdy, jsou-li nekorelované. Příkladem této situace je náhodný vektor $(X,Y)^T$ s

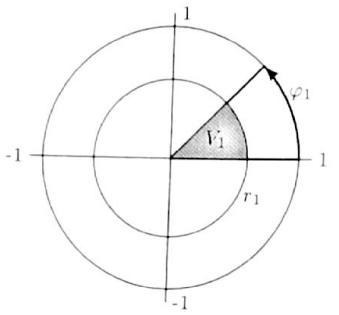
$$\begin{aligned} \mathsf{P}[(X,Y)^T = (0,0)^T] &= \mathsf{P}[(X,Y)^T = (1,0)^T] = \frac{1}{4}, \\ \mathsf{P}[(X,Y)^T = (0,1)^T] &= \mathsf{P}[(X,Y)^T = (1,1)^T] = \frac{1}{4}, \end{aligned}$$

tj. náhodný vektor s rovnoměrným rozdělením na vrcholech jednotkového čtverce. Platí $\mathsf{cov}(X,Y) = 0$, veličiny X, Y jsou nezávislé a obě mají rozdělení

bi$(1, 1/2)$. Také obě složky X' a Y' náhodného vektoru $(X', Y')^T$ s

$$\mathsf{P}\left[\begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right] = \mathsf{P}\left[\begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right] = \frac{1}{2}$$

mají rozdělení bi$(1, 1/2)$, platí však $\mathsf{cov}(X', Y') = 1/4 \neq 0$. Veličiny X' a Y' nezávislé nejsou. Není divu, vždyť je $\mathsf{P}[X' = Y'] = 1$. ○



Obrázek 7.1: Rovnoměrné rozdělení na jednotkovém kruhu

Příklad 7.4. Buď $(X, Y)^T$ náhodný vektor, který má rovnoměrné rozdělení na jednotkovém kruhu $K = \{(x, u)^T \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$, tj. náhodný vektor, jehož rozdělení pravděpodobností má hustotu $f(x, y) = \frac{1}{\pi}$ pro $(x, y)^T \in K$ a $f(x, y) = 0$ všude jinde na $\mathbb{R}^2$. *Veličiny X a Y nejsou samozřejmě nezávislé* (vždyť $\mathsf{P}[(X, Y)^T \in \check{C} - K] = 0$, kde $\check{C}$ je čtverec s vrcholy $[1, 1], [1, -1], [-1, -1], [-1, 1]$). Označíme jako $R = R(X, Y)$ a $\Phi = \Phi(X, Y)$ polární souřadnice náhodného vektoru X, Y a určíme hustotu sdruženého rozdělení pravděpodobností náhodného vektoru $(R, \Phi)^T$: pro $0 < r \leq 1$ a $0 < \varphi \leq 2\pi$ je

$$\begin{aligned} \mathsf{P}[R < r_1, \Phi < \varphi_1] &= \mathsf{P}[(X, Y)^T \in V_1] = \frac{1}{\pi}\pi r_1^2 \frac{\varphi_1}{2\pi} \\ &= \int_0^{r_1}\int_0^{\varphi_1} \frac{1}{2\pi} 2r \mathrm{d}\varphi \mathrm{d}r, \end{aligned}$$

(kde $V_1 = \{(r, \varphi)^T : 0 < r < r_1, 0 < \varphi < \varphi_1\}$, viz obrázek 7.1). Hustota $f(r, \varphi)$ rozdělení náhodného vektoru (R, Φ) je proto dána předpisem

$$f(r, \varphi) = \begin{cases} \frac{r}{\pi} & \text{pro } 0 < r \leq 1, 0 < \varphi \leq 2\pi, \\ 0 & \text{jinde.} \end{cases}$$

Marginální hustoty $g(r)$ a $h(\varphi)$ náhodných veličin R a Φ jsou podle (5.4) určeny jako

$$g(r) = \begin{cases} \int_{-\infty}^{\infty} f(r, \varphi)\mathrm{d}\varphi = \int_0^{2\pi} \frac{r}{\pi}\mathrm{d}\varphi = 2r & \text{pro } 0 < r \leq 1, \\ 0 & \text{jinde,} \end{cases}$$

$$h(\varphi) = \begin{cases} \int_{-\infty}^{\infty} f(r, \varphi)\mathrm{d}r = \int_0^1 \frac{r}{\pi}\mathrm{d}r = \frac{1}{2\pi} & \text{pro } 0 < \varphi \leq 2\pi, \\ 0 & \text{jinde.} \end{cases}$$

Všimněme si, že náhodná veličina Φ má rovnoměrné rozdělení na intervalu $(0, 2\pi)$, tj. $\mathsf{E}\Phi = \pi$, $\mathsf{var}\Phi = \frac{\pi^2}{3}$. Nyní určíme tyto momenty také pro náhodnou veličinu R:

$$\begin{aligned} \mathsf{E}R &= \int_0^1 r \cdot 2r\mathrm{d}r = \frac{2}{3}, \\ \mathsf{var}R &= \int_0^1 r^2 \cdot 2r\mathrm{d}r - \frac{4}{9} = \frac{1}{2} - \frac{4}{9} = \frac{1}{18}. \end{aligned}$$

Zajímavé ovšem je, že platí $f(r, \varphi) = g(r)h(\varphi)$ všude na $\mathbb{R}^2$, což podle (5.7) znamená nezávislost náhodných veličin R a Φ. Speciálně odtud plyne, že je $\mathsf{cov}(R, \Phi) = 0$. Pro čtenáře bude jistě užitečným cvičením spočítat také rozdělení, střední hodnoty, rozptyly a kovarianci původních náhodných veličin X a Y. ○

Příklad 7.5. Buďte α a X nezávislé náhodné veličiny, $X \sim \mathsf{N}(0, 1)$ a $\mathsf{P}[\alpha = 1] = \mathsf{P}[\alpha = -1] = 1/2$. Uvážíme-li náhodný vektor $(X, Y)^T = (X, \alpha X)^T$, platí $\mathsf{P}[X = Y] = \mathsf{P}[\alpha = 1] = 1/2 = \mathsf{P}[\alpha = -1] = \mathsf{P}[X = -Y]$ a X, Y jsou náhodné veličiny s rozdělením $\mathsf{N}(0, 1)$, protože

$$\mathsf{P}[Y < y] = \frac{1}{2}\mathsf{P}[X < y] + \frac{1}{2}\mathsf{P}[-X < y] = \Phi(y)$$

pro $y \in (-\infty, \infty)$. Počítejme (pomocí věty 6.4):

$$\mathsf{cov}(X, Y) = \mathsf{E}(XY) = \mathsf{E}(\alpha X^2) = \mathsf{E}\alpha\mathsf{E}X^2 = 0 \cdot 1 = 0.$$

Vytvořili jsme náhodný vektor $(X, Y)^T$, jehož souřadnice X a Y jsou nekorelované, nikoliv nezávislé náhodné veličiny s normovaným normálním rozdělením. Rozdělení pravděpodobností náhodného vektoru $(X, Y)^T$ přiděluje veškerou pravděpodobnostní hmotu do množiny

$$\{(x, y)^T \in \mathbb{R}^2 : x = y \text{ nebo } x = -y\},$$

nemá proto pravděpodobnostní hustotu, není to dvourozměrné normální rozdělení. ○

Následující dva příklady ilustrují v jednoduchých situacích jednu ze základních pravděpodobnostních myšlenek: Chceme-li informaci o některé konkrétní vlastnosti jistého rozdělení pravděpodobností, pokusíme se nalézt náhodnou veličinu, která toto rozdělení má, a navíc je taková, že se nám s ní při řešení konkrétní úlohy dobře pracuje.

Příklad 7.6. Vypočteme rozptyl náhodné veličiny X, která má binomického rozdělení bi(n, p). Podle příkladu 6.4 můžeme volit $X = \sum_{i=1}^n Y_i$, kde

$Y_1, Y_2, \ldots, Y_n$ jsou nezávislé náhodné veličiny s $\mathsf{P}[Y_i = 1] = p$ a $\mathsf{P}[Y_i = 0] = 1 - p$. Protože je

$$\mathsf{E}Y_i^2 = 1^2 p + 0^2(1-p) = p,$$

má každá z veličin Y_i rozptyl

$$\begin{aligned} \mathsf{var}Y_i &= \mathsf{E}Y_i^2 - (\mathsf{E}Y_i)^2 \\ &= p - p^2 \\ &= p(1-p), \end{aligned}$$

takže náhodná veličina X s binomickým rozdělením má rozptyl $np(1-p)$. Podobně, má-li X negativně binomické rozdělení s parametry r, p, můžeme ji podobně chápat jako součet nezávislých náhodných veličin $Y_1, \ldots, Y_r$, z nichž každá má geometrické rozdělení s parametrem p. Použijeme-li příklad 7.1, dostaneme

$$\mathsf{E}X = r\left(\frac{1}{p} - 1\right), \quad \mathsf{var}X = \frac{r(1-p)}{p^2}. \qquad \bigcirc$$

Příklad 7.7. Vraťme se k příkladu 1.2 o výběrových šetřeních a zabývejme se nejprve otázkou, jak realizovat náhodný pokus, který produkuje náhodný výběr s o rozsahu n z populace $S = \{1, 2, \ldots, N\}$ tak, aby všechny takové výběry byly stejně pravděpodobné, tj. $\mathsf{P}(s) = \binom{N}{n}^{-1}$. Přirozeně se nabízí model uspořádaného výběru s vracením (viz oddíl 3.1), kdy z urny $\{1, 2, \ldots, N\}$ vybíráme jednotky $j_1, j_2, \ldots, j_{T_n}$ právě tak dlouho, abychom získali n různých jednotek. Matematicky:

$$T_n = \min\{k \geq n : |(j_1, \ldots, j_k)| = n\}.$$

Snadno se přesvědčíme, že tato procedura – postupný náhodný výběr – má požadované vlastnosti. Jest tedy otázkou, jak dlouho budeme tento náhodný výběr (v průměru) pořizovat, kolik tahů z urny, kolik náhodných čísel bude třeba v počítači generovat. Jinými slovy chceme určit střední hodnotu náhodné veličiny T_n. Tuto veličinu však můžeme vyjádřit jako součet $T_n = \sum_{k=1}^{n} Y_k$, kde $Y_1 = 1$, Y_2 =[počet tahů potřebných k získání druhé jednotky výběru poté, co již byla vybrána první], ..., Y_k =[počet tahů potřebných k získání k-té jednotky výběru poté, co již bylo vybráno $k-1$ různých jednotek],.... Všimněme si, že Y_k je vlastně počet pokusů potřebných k registraci prvního zdaru v Bernoulliově modelu s pravděpodobností zdaru $p_k = (N-k+1)/N$ (vytažení některé z $k-1$ jednotek, které již byly taženy je nezdar). To ovšem znamená, že náhodná veličina $Y_k - 1$ má geometrické

rozdělení s parametrem p_k (viz příklad 7.1), tj. $\mathsf{E}Y_k = \frac{1-p_k}{p_k} + 1 = \frac{N}{N-k+1}$ pro $1 \leq k \leq n$. Celkem je tedy

$$\mathsf{E}T_n = \sum_{k=1}^{n} \mathsf{E}Y_k = N\left(\frac{1}{N-n+1} + \frac{1}{N-n} + \ldots + \frac{1}{N}\right)$$

a můžeme použít Eulerovu formuli pro částečné součty harmonické řady ($\sum_{k=1}^{m} \frac{1}{k} = c + \ln m + o(m^{-1})$ při $m \to \infty$), abychom obdrželi aproximaci $\mathsf{E}T_n \doteq \ln \frac{N}{N-n} = -\ln(1-\alpha)$, kde $\alpha = n/N$ je podíl rozsahu výběru a rozsahu populace.

Náš model umožňuje také výpočet rozptylu náhodné veličiny T_n, protože náhodné veličiny $Y_1, Y_2, \ldots, Y_n$ jsou zřejmě nezávislé. Podle příkladu 7.1 je

$$\mathsf{var}Y_k = \frac{1-p_k}{p_k^2} = \frac{N(k-1)}{(N-k-1)^2},$$

podle věty 7.3 je nakonec

$$\mathsf{var}T_n = \sum_{k=1}^{n} \mathsf{var}Y_k = N\sum_{k=1}^{n} \frac{k-1}{(N-k-1)^2}.$$

7.3 Další momenty

Někdy se zavádí řada dalších charakteristik. V definičních vztazích budeme vždy předpokládat existenci, konečnost, případně nenulovost jednotlivých středních hodnot. Charakteristika

$$\mu'_k = \mathsf{E}X^k \tag{7.17}$$

se nazývá **k-tý obecný moment** náhodné veličiny X, charakteristika

$$\mu_k = \mathsf{E}(X - \mathsf{E}X)^k \tag{7.18}$$

se nazývá **k-tý centrální moment** náhodné veličiny X. Charakteristika

$$\gamma_1 = \frac{\mu_3}{(\sqrt{\mathsf{var}X})^3} \tag{7.19}$$

se nazývá **koeficient šikmosti** náhodné veličiny X, charakteristika

$$\gamma_2 = \frac{\mu_4}{(\mathsf{var}X)^2} - 3 \tag{7.20}$$

se nazývá **koeficient špičatosti** náhodné veličiny X.

Ještě drobná poznámka ke koeficientu špičatosti γ_2. Jeho klasický název je v poslední době kritizován jako ne zcela výstižný. Pěkný článek na toto téma uveřejnili Čermák a Vodrážková [4].

Místo posledně uvedených charakteristik se někdy pracuje s podobně definovanými charakteristikami ([3], kap. 15.8)

$$\beta_1 = \gamma_1^2, \ \beta_2 = \gamma_2 + 3. \tag{7.21}$$

Výpočet momentů, tedy i střední hodnoty a rozptylu, bývá velmi pracný. Jsou však k disposici užitečné pomůcky, které navíc usnadňují i některé důkazy.

Definice 7.4. Reálnou funkci reálné proměnné

$$M_X(t) = \mathsf{E}\mathrm{e}^{tX} \tag{7.22}$$

nazveme **momentovou vytvořující funkcí** náhodné veličiny X.

Všimněte si, že tentokrát přiřazujeme náhodné veličině X *nenáhodnou* funkci. Do jisté míry, za splnění jistých požadavků regularity, jde o alternativní charakterizaci rozdělení náhodné veličiny (např. místo distribuční funkce). Bez důkazu uvedeme následující tvrzení (viz např. [18], II.7.24, [19], věta 4.37, [15], kap. 2).

Věta 7.7. Je-li momentová vytvořující funkce $M_X(t)$ náhodné veličiny X taková, že $M(b) < \infty$, $M(-b) < \infty$ pro některé $b > 0$, pak
a) M je konečná spojitá funkce na $\langle -b, b\rangle$.
b) M má spojité derivace všech řádů na $(-b, b)$.
c) Platí $\mathsf{E}|X|^r < \infty$ pro $r \geq 1$,

$$\mu'_r = \frac{\mathrm{d}^r}{\mathrm{d}t^r} M_X(t)|_{t=0}. \tag{7.23}$$

d) Pro $t \in \langle -b, b\rangle$ platí

$$M_X(t) = \sum_{r=0}^{\infty} \frac{t^r}{r!} \mu'_r, \tag{7.24}$$

přičemž řada konverguje absolutně.
e) Platí-li $M_X(t) = M_Y(t)$ pro všechna $t \in \langle -b, b\rangle$ pro nějaké $b > 0$, mají náhodné veličiny X, Y stejné rozdělení, tj. platí

$$F_X(x) \equiv F_Y(x). \tag{7.25}$$

Tvrzení c) ukazuje důvod pro slovní označení funkce $M_X(t)$. Derivováním vytvořující funkce můžeme dojít ke všem momentům. Vyjádříme-li

vytvořující funkci $M_X(t)$ jako mocninnou řadu v proměnné t, z koeficientů u jednotlivých členů můžeme najít obecné momenty také.

Všimněme si některých vlastností momentové vytvořující funkce.

Věta 7.8. Nechť a, b jsou reálná čísla, nechť X je náhodná veličina s momentovou vytvořující funkcí $M_X(t), t \in \langle -b, b\rangle, b > 0$. Potom momentová vytvořující funkce náhodné veličiny $Y = a + bX$ má tvar

$$M_Y(t) = \mathrm{e}^{at} M_X(bt). \tag{7.26}$$

D ů k a z: Důkaz spočívá v prostém výpočtu

$$\begin{aligned} M_Y(t) &= \mathsf{E}\mathrm{e}^{(a+bX)t} \\ &= \mathsf{E}\mathrm{e}^{at}\mathrm{e}^{(bt)X} \\ &= \mathrm{e}^{at} M_X(bt). \end{aligned}$$ □

Věta 7.9. Pro momentovou vytvořující funkci součtu dvou nezávislých náhodných veličin X, Y platí

$$M_{X+Y}(t) = M_X(t) M_Y(t), \tag{7.27}$$

pokud všechny střední hodnoty existují.

D ů k a z: Pro náhodnou veličinu W, která je součtem *nezávislých* náhodných veličin X, Y, dostaneme:

$$\begin{aligned} M_W(t) &= \mathsf{E}\mathrm{e}^{(X+Y)t} \\ &= \mathsf{E}\mathrm{e}^{Xt} e^{Yt} \\ &= \mathsf{E}\mathrm{e}^{Xt}\mathsf{E}\mathrm{e}^{Yt} \\ &= M_X(t) M_Y(t), \end{aligned}$$

když jsme použili zejména nezávislost náhodných veličin X, Y a tedy veličin $\mathrm{e}^{Xt}, \mathrm{e}^{Yt}$. □

Příklad 7.8. Spočítejme momentovou vytvořující funkci binomického rozdělení $\mathrm{bi}(n, p)$ (viz příklad 4.1). Pro všechna reálná t platí

$$\begin{aligned} M(t) &= \mathsf{E}\mathrm{e}^{tX} \\ &= \sum_{k=0}^{n} \mathrm{e}^{tk} \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^{n} \binom{n}{k} (p\mathrm{e}^t)^k (1-p)^{n-k} \\ &= \left(p(\mathrm{e}^t - 1) + 1\right)^n. \end{aligned} \tag{7.28}$$

Jinak (a snáze) jsme mohli funkci (7.28) nalézt z vytvořující funkce alternativního rozdělení s využitím věty 7.9. Derivováním (7.28) dostaneme

$$\frac{\mathrm{d}}{\mathrm{d}t}M(t) = n\left(p(\mathrm{e}^t - 1) + 1\right)^{n-1}\mathrm{e}^t p,$$

což po dosazení $t = 0$ dá ${\mu_1}' = np$. Dalším derivováním dostaneme

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}M(t) = n(n-1)\left(p(\mathrm{e}^t - 1) + 1\right)^{n-2}\mathrm{e}^{2}p^2 + n\left(p(\mathrm{e}^t - 1) + 1\right)^{n-1}\mathrm{e}^t p,$$

což po opětném dosazení dá ${\mu_2}' = np + n(n-1)p^2$. Rozptyl je tedy roven

$$\mathsf{var}X = np + n(n-1)p^2 - (np)^2 = np(1-p). \qquad \bigcirc$$

Příklad 7.9. Připomeňme normální rozdělení zavedené v příkladu 4.9. Spočítejme nejprve momentovou vytvořující funkci normovaného normálního rozdělení:

$$\begin{aligned} M_Z(t) &= \int_{-\infty}^{\infty} \mathrm{e}^{tz}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)\mathrm{d}z \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2 - 2tz + t^2 - t^2}{2}\right)\mathrm{d}z \\ &= \exp\left(\frac{t^2}{2}\right)\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(z-t)^2}{2}\right)\mathrm{d}z \\ &= \exp\left(\frac{t^2}{2}\right), \end{aligned}$$

když jsme použili skutečnost, že v posledním integrálu je na místě integrované funkce hustota tvaru (4.26), takže je tento integrál roven 1. Momentová vytvořující funkce je v okolí nuly konečná, takže podle věty 7.7 všechny momenty existují. Spočítejme střední hodnotu a rozptyl náhodné veličiny Z s hustotou $\varphi(z)$. Protože platí

$$\begin{aligned} {M'}_Z(t) &= t\exp\left(\frac{t^2}{2}\right), \\ {M''}_Z(t) &= t^2\exp\left(\frac{t^2}{2}\right) + \exp\left(\frac{t^2}{2}\right). \end{aligned}$$

Po dosazení $t = 0$ dostaneme $\mathsf{E}Z = 0, \mathsf{var}Z = 1$.

A nyní k náhodné veličině $Y = \mu + \sigma Z \sim \mathsf{N}(\mu, \sigma^2)$, jejíž hustotu jsme našli v (4.26). Vzhledem k pravidlům (6.9) a (7.2) musí platit $\mathsf{E}Y = \mu, \mathsf{var}Y = \sigma^2$. Momentová vytvořující funkce Y má vzhledem k tvrzení věty 7.8 tvar

$$M_Y(t) = \exp\left(\mu t + \sigma^2\frac{t^2}{2}\right). \qquad (7.29) \qquad \bigcirc$$

7.4 Cvičení

7.1. Nechť náhodné veličiny U, V mají diskrétní rozdělení určené tabulkou. Najděte marginální rozdělení obou náhodných veličin, jejich střední hodnoty, rozptyly a korelační koeficient.

	V		
U	1	2	3
1	0,1	0,2	0,3
2	0,2	0,1	0,1

7.2. Najděte střední hodnoty, rozptyl a korelační koeficient náhodných veličin X a Y z příkladu 5.3. Jsou tyto náhodné veličiny nezávislé?

7.3. Nechť X je počet líců při třech hodech korunovou mincí, nechť Y je počet líců při čtyřech hodech pětikorunovou mincí. Označme celkový počet líců v těchto pokusech jako W. Určete $\rho_{X,W}$.

7.4. Spočítejte korelační koeficient náhodných veličin X a $Y = X^2$, kde X má rovnoměrné rozdělení na intervalu $(-1, 1)$. Jsou tyto náhodné veličiny nezávislé?

7.5. Určete korelační koeficient složek náhodného vektoru $(X, Y)^T$, který má rovnoměrné rozdělení v trojúhelníku ohraničeném přímkami $x = 0$, $y = 0$, $x + y = c$, kde $c > 0$ je daná konstanta (uvnitř tohoto trojúhelníku je hustota rovna vhodné konstantě, jinak je hustota nulová).

7.6. Nechť X má rovnoměrné rozdělení na intervalu $(1, 2)$. Určete korelační koeficient $\rho_{X,1/X}$.

8. Některá rozdělení

Dříve, než uvedeme přehled používaných rozdělení, ukážeme několik metod hledání rozdělení funkcí náhodných veličin.

8.1 Konvoluce

O **konvoluci** hovoříme, když nová náhodná veličina vzniká jako součet *nezávislých* náhodných veličin. Někdy má součet rozdělení stejného typu jako mají jednotlivé sčítance, pouze s jinými parametry – v takovém případě se uplatní momentová vytvořující funkce. Jindy musíme hledat pravděpodobnosti či hustotu součtu přímo.

Věta 8.1. Jsou-li X, Y nezávislé náhodné veličiny s hustotami $f_X(x)$, $f_Y(y)$, pak má náhodná veličina $W = X + Y$ hustotu

$$f_W(w) = \int_{-\infty}^{\infty} f_X(x) f_Y(w-x)\mathrm{d}x. \tag{8.1}$$

D ů k a z: Pro distribuční funkci součtu dostaneme (substituce $t = x + y$)

$$\begin{aligned} F_W(w) &= \mathsf{P}[X+Y<w] \\ &= \int\int_{x+y<w} f_X(x) f_Y(y)\mathrm{d}y\mathrm{d}x \\ &= \int_{-\infty}^{w} \left(\int_{-\infty}^{\infty} f_X(x) f_Y(t-x)\mathrm{d}x \right) \mathrm{d}t, \end{aligned}$$

takže zřejmě (8.1) je hustotou náhodné veličiny W. □

Pro diskrétní nezávislé náhodné veličiny podobně dostaneme

$$\mathsf{P}[W = w_k] = \sum_{i=1}^{\infty} \mathsf{P}[X = x_i]\mathsf{P}[Y = w_k - x_i] \tag{8.2}$$

Příklad 8.1. Nechť každá z nezávislých náhodných veličin X, Y má rovnoměrné rozdělení na intervalu $(0, 1)$. Součet bude zřejmě nabývat hodnot z intervalu $(0, 2)$. Hustotu součtu dostaneme postupnou úpravou integrálu $(0 < w < 2)$

$$\begin{aligned} f_W(w) &= \int_{-\infty}^{\infty} f_X(x) f_Y(w-x)\mathrm{d}x \\ &= \int_{\max(0,w-1)}^{\min(1,w)} 1\mathrm{d}x, \end{aligned}$$

když jsme meze integrálu omezili na interval, v němž je součin hustot nenulový. Pro $0 < w \leq 1$ dostaneme

$$f_W(w) = \int_0^w 1\mathrm{d}x = w,$$

pro $1 \leq w < 2$ podobně

$$f_W(w) = \int_{w-1}^1 1\mathrm{d}x = 2 - w.$$

Pro $w \notin (0, 2)$ je zřejmě aspoň jedna z hustot $f_X(x), f_Y(y)$ nulová, takže je pak také $f_W(w) = 0$. ○

Příklad 8.2. Počítejme konvoluci náhodných veličin $X \sim \mathsf{N}(\mu_X, \sigma_X^2)$, $Y \sim \mathsf{N}(\mu_Y, \sigma_Y^2)$. Podle věty 7.9 dostaneme

$$\begin{aligned} M_W(t) &= \exp\left(\mu_X t + \sigma_X^2 \frac{t^2}{2}\right) \exp\left(\mu_Y t + \sigma_Y^2 \frac{t^2}{2}\right) \\ &= \exp\left((\mu_X + \mu_Y)t + (\sigma_X^2 + \sigma_Y^2)\frac{t^2}{2}\right), \end{aligned}$$

což ukazuje na to, že platí (viz (7.29))

$$X + Y \sim \mathsf{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

Pomocí vzorce (8.1) bychom došli ke stejnému výsledku, ale podstatně pracněji. ○

Příklad 8.3. Předpokládejme, že hustota náhodné veličiny X má pro $x > 0$ tvar

$$cx^{a-1}\mathrm{e}^{-bx},$$

kde $a > 0, b > 0$ jsou parametry, jinak je hustota nulová. Především určíme konstantu c. Musí platit

$$\begin{aligned} 1 &= \int_0^{\infty} cx^{a-1}\mathrm{e}^{-bx}\mathrm{d}x \\ &= \int_0^{\infty} c\left(\frac{t}{b}\right)^{a-1}\mathrm{e}^{-t}\frac{1}{b}\mathrm{d}t \\ &= \frac{c}{b^a}\Gamma(a), \end{aligned}$$

kde jsme použili známou Γ-funkci (viz Appendix A2). Je tedy (pro $x > 0$, jinak je hustota nulová)

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \mathrm{e}^{-bx} \tag{8.3}$$

hustota **gama rozdělení** s parametry $a > 0, b > 0$. Toto rozdělení budeme značit $\Gamma(a, b)$.

Spočítejme základní charakteristiky. Nejprve určíme momenty μ'_r. Použijeme při tom skutečnost, že integrál z hustoty rozdělení $\Gamma(a+r, b)$ je roven jedné:

$$\begin{aligned} \mu'_r &= \int_0^\infty x^r \frac{b^a}{\Gamma(a)} x^{a-1} \mathrm{e}^{-x} \mathrm{d}x \\ &= \frac{\Gamma(a+r)}{\Gamma(a) b^r} \int_0^\infty x^r \frac{b^{(a+r)}}{\Gamma(a+r)} x^{a+r-1} \mathrm{e}^{-bx} \mathrm{d}x \\ &= \frac{\Gamma(a+r)}{\Gamma(a) b^r}. \end{aligned}$$

Je tedy speciálně

$$\mathsf{E}X = \frac{\Gamma(a+1)}{b\Gamma(a)} = \frac{a}{b}$$

$$\mathsf{var}X = \frac{\Gamma(a+2)}{b^2\Gamma(a)} - \left(\frac{a}{b}\right)^2 = \frac{a}{b^2}.$$

○

Příklad 8.4. Momentová vytvořující funkce pro Γ-rozdělení z (8.3) má pro $|t| < b$ tvar

$$\begin{aligned} M(t) &= \int_0^\infty \mathrm{e}^{tx} \frac{b^a}{\Gamma(a)} x^{a-1} \mathrm{e}^{-bx} \mathrm{d}x \\ &= \frac{b^a}{(b-t)^a} \int_0^\infty \frac{(b-t)^a}{\Gamma(a)} x^{a-1} \mathrm{e}^{-(b-t)x} \mathrm{d}x \\ &= \left(\frac{b}{b-t}\right)^a. \end{aligned}$$

Jsou-li X, Y *nezávislé* náhodné veličiny s rozděleními $\Gamma(a_x, b)$ a $\Gamma(a_y, b)$, pak součet těchto náhodných veličin má zřejmě momentovou vytvořující funkci

$$\left(\frac{b}{b-t}\right)^{a_x} \left(\frac{b}{b-t}\right)^{a_y} = \left(\frac{b}{b-t}\right)^{a_x+a_y},$$

tedy opět Γ-rozdělení, tentokrát s parametry $a_x + a_y, b$. Předpoklad, že parametry b jsou u obou náhodných veličin *stejné*, je podstatný. ○

Proč jsme nepoužili při hledání rozdělení součtu dvou nezávislých náhodných veličin s rovnoměrným rozdělením jejich momentovou vytvořující funkci? Bylo by to vhodné, když má tato funkce tvar $M(t) = (\mathrm{e}^t - 1)/t$? Všimněte si, že součin dvou takovýchto funkcí se od $M(t)$ liší víc, než jen hodnotou parametru, jak to bylo například u normálního rozdělení.

Příklad 8.5. Připomeňme příklad 4.11, kde jsme určili hustotu náhodné veličiny Y, která byla druhou mocninou náhodné veličiny s normálním rozdělením $\mathsf{N}(0, 1)$. Porovnáme-li tuto hustotu (4.27) s hustotou (8.3), zjistíme, že jde o Γ-rozdělení s parametry $1/2, 1/2$. Nechť $Z_1, Z_2, \ldots, Z_n$ jsou nezávislé náhodné veličiny s normálním rozdělením $\mathsf{N}(0, 1)$, nechť je dále

$$Y = \sum_{i=1}^n Z_i^2.$$

Každý sčítanec má rozdělení $\Gamma(1/2, 1/2)$, sčítance jsou nezávislé, takže podle příkladu 8.4 má náhodná veličina Y rozdělení $\Gamma(n/2, 1/2)$. Tomuto speciálnímu případu Γ-rozdělení budeme říkat **χ^2 rozdělení** s n stupni volnosti (čte se chí-kvadrát) a budeme je značit symbolem $\chi^2(n)$. Hustota má tvar

$$\frac{1}{2^{n/2}\Gamma(n/2)} y^{n/2-1} \mathrm{e}^{-y/2}. \tag{8.4}$$

Toto rozdělení, jak jsme viděli, těsně souvisí s normálním rozdělením a ve statistice se používá velmi často. ○

8.2 Rozdělení odvozená od normálního

Věnujme se nyní jiným funkcím náhodných veličin. Odvodíme postupně tzv. výběrová rozdělení, která budeme později (v 12. a 14. kapitole) používat při zpracování náhodného výběru z normálního rozdělení (viz str. 157).

Věta 8.2. Nechť nezávislé náhodné veličiny X, Y mají hustoty f_X, f_Y, přičemž platí $f_Y(y) = 0$ pro $y \leq 0$, takže je $\mathsf{P}[Y > 0] = 1$, nechť $c > 0$ je daná konstanta. Potom má náhodná veličina $U = cX/Y$ rozdělení s hustotou

$$f_U(u) = \frac{1}{c} \int_0^\infty y f_X\left(\frac{uy}{c}\right) f_Y(y) \mathrm{d}y \tag{8.5}$$

pro $u > 0$, jinak je hustota nulová.

D ů k a z: Nejprve určíme distribuční funkci U, přičemž provedeme takovou substituci, aby horní mez integrálu byla rovna proměnné u.

$$\begin{aligned} F_U(u) &= \mathsf{P}[X < (u/c)Y] \\ &= \int_0^\infty \left(\int_{-\infty}^{uy/c} f_X(x) f_Y(y) \mathrm{d}x \right) \mathrm{d}y \\ &= \int_0^\infty \left(\int_{-\infty}^{u} \frac{1}{c} y f_X(ty/c) f_Y(y) \mathrm{d}t \right) \mathrm{d}y \\ &= \int_{-\infty}^{u} \left(\frac{1}{c} \int_0^\infty y f_X(ty/c) f_Y(y) \mathrm{d}y \right) \mathrm{d}t, \end{aligned}$$

když jsme použili Fubiniovu větu k záměně pořadí integrování. Z posledního vztahu již dokazované tvrzení bezprostředně plyne. □

Příklad 8.6. Mějme dvě *nezávislé* náhodné veličiny X, Y, které mají rozdělení $\chi^2(k)$ a $\chi^2(m)$. Hledáme rozdělení podílu

$$U = \frac{X/k}{Y/m}.$$

Použijeme větu 8.2, kde dosadíme $c = m/k$. Zvolíme $u > 0$ (jinak je hustota nulová) a dosadíme-li do (8.5) podle (8.4), dostaneme postupně

$$\begin{aligned} f_U(u) &= \frac{k}{m} \int_0^\infty y \frac{1}{2^{k/2}\Gamma(k/2)} \left(\frac{kuy}{m} \right)^{k/2-1} \\ &\quad \times \mathrm{e}^{-kuy/(2m)} \frac{1}{2^{m/2}\Gamma(m/2)} y^{m/2-1} \mathrm{e}^{-y/2} \mathrm{d}y, \\ f_U(u) &= \left(\frac{k}{m} \right)^{k/2} \frac{1}{2^{(k+m)/2}\Gamma(k/2)\Gamma(m/2)} \\ &\quad \times u^{k/2-1} \int_0^\infty y^{(k+m)/2-1} \mathrm{e}^{-y(1+ku/m)/2} \mathrm{d}y. \end{aligned}$$

Poslední integrál obsahuje až na konstantu

$$\frac{((1+ku/m)/2)^{(k+m)/2}}{\Gamma((k+m)/2)}$$

hustotu rozdělení $\Gamma((k+m)/2, (1+ku/m)/2)$, proto má výsledná hustota tvar

$$f_U(u) = \frac{\Gamma((k+m)/2)}{\Gamma(k/2)\Gamma(m/2)} \left(\frac{k}{m} \right)^{k/2} u^{k/2-1} \left(1 + \frac{k}{m} u \right)^{-(k+m)/2}. \tag{8.6}$$

Rozdělení s hustotou (8.6) se nazývá ***F*-rozdělení** (také Fisherovo-Snedecorovo rozdělení) s k a m stupni volnosti a značí se $\mathsf{F}(k, m)$. ○

Příklad 8.7. Mějme náhodnou veličinu X s rozdělením χ^2_n. Náhodná veličina $Y = \sqrt{X}$ bude nabývat stejně jako X pouze kladných hodnot. Distribuční funkce Y je pro $y > 0$ rovna

$$\begin{aligned} F_Y(y) &= \mathsf{P}[\sqrt{X} < y] \\ &= \mathsf{P}[X < y^2] \\ &= \int_0^{y^2} \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} \mathrm{e}^{-x/2} \mathrm{d}y \\ &= \int_0^{y} \frac{1}{2^{n/2-1}\Gamma(n/2)} t^{n-1} \mathrm{e}^{-t^2/2} \mathrm{d}t, \end{aligned}$$

takže hledaná hustota **rozdělení χ** s n stupni volnosti je rovna

$$f_Y(y) = \frac{1}{2^{n/2-1}\Gamma(n/2)} y^{n-1} \mathrm{e}^{-y^2/2}. \tag{8.7}$$ ○

Příklad 8.8. Velmi často budeme ve statistice pracovat s náhodnou veličinou, která je dána podílem nezávislých náhodných veličin

$$T = \frac{Z}{\sqrt{X/n}},$$

$Z \sim \mathsf{N}(0,1)$ a $X \sim \chi^2(n)$. K odvození hustoty použijeme tvrzení věty 8.2, kde dosadíme $c = \frac{1}{\sqrt{n}}$ a $Y = \sqrt{X}$. Hustota je pak dána výpočtem

$$\begin{aligned} f_T(t) &= \frac{1}{\sqrt{n}} \int_0^\infty y \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(ty)^2/(2n)} \frac{1}{2^{n/2-1}\Gamma(n/2)} y^{n-1} \mathrm{e}^{-y^2/2} \mathrm{d}y \\ &= \frac{1}{\Gamma(n/2) 2^{(n-1)/2} \sqrt{n\pi}} \int_0^\infty x^{(n+1)/2-1} \mathrm{e}^{-x(1+t^2/n)/2} \mathrm{d}x. \end{aligned}$$

Použijeme-li opět skutečnost, že pod integrálem je až na konstantu hustota rozdělení $\Gamma((n+1)/2, (1+t^2/n)/2)$, dostaneme výslednou hustotu Studentova ***t*-rozdělení** s n stupni volnosti (označení $\mathsf{t}(n)$):

$$f_T(t) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{\pi n}} \left(1 + \frac{t^2}{n} \right)^{-(n+1)/2}. \tag{8.8}$$ ○.

8.3 Mnohorozměrné normální rozdělení

Věta 8.3. Nechť složky $Z_1, \dots, Z_n$ náhodného vektoru $\boldsymbol{Z}$ jsou nezávislé a každá má rozdělení $\mathsf{N}(0,1)$. Nechť $\mathbf{Q}$ je ortonormální matice řádu n. Potom složky $U_1, \dots, U_n$ náhodného vektoru $\boldsymbol{U}$ definovaného vztahem $\boldsymbol{U} = \mathbf{Q}^T\boldsymbol{Z}$ jsou nezávislé a každá z nich má rozdělení $\mathsf{N}(0,1)$.

D ů k a z: Ukážeme, že sdružená distribuční funkce náhodného vektoru $\boldsymbol{U}$ je součinem marginálních distribučních funkcí jeho složek. Při úpravách integrálu přitom použijeme substituci $\mathbf{t} = \mathbf{Q}^T\mathbf{z}$. Inverzní transformace má vzhledem k ortonormalitě tvar $\mathbf{z} = \mathbf{Qt}$ a její jakobián je roven 1. Navíc ze stejné vlastnosti matice $\mathbf{Q}$ plyne, že je $\sum z_i^2 = \sum t_i^2$.

$$\begin{aligned}
F_{\boldsymbol{U}}(\mathbf{u}) &= \mathsf{P}[\boldsymbol{U} < \mathbf{u}] \\
&= \int\cdots\int_{\{\mathbf{z}:\mathbf{Q}^T\mathbf{z}<\mathbf{u}\}} (2\pi)^{-n/2}\mathrm{e}^{-\sum z_i^2/2}\mathrm{d}z_1\cdots\mathrm{d}z_n \\
&= \int\cdots\int_{\{\mathbf{t}:\mathbf{t}<\mathbf{u}\}} (2\pi)^{-n/2}\mathrm{e}^{-\sum t_i^2/2}\mathrm{d}t_1\cdots\mathrm{d}t_n \\
&= \prod_{i=1}^{n}\left(\int_{-\infty}^{u_i}(2\pi)^{-1/2}\mathrm{e}^{-t_i^2/2}\mathrm{d}t_i\right) \\
&= \prod_{i=1}^{n} F_{U_i}(u_i).
\end{aligned}$$

□

Pomocí (6.12) a (7.16) snadno zjistíme, že vektor $\boldsymbol{U}$ má, stejně jako $\boldsymbol{Z}$, nulovou střední hodnotu a jednotkovou varianční matici. Oba jsou zobecněním normovaného normálního rozdělení. Zobecnění normálního rozdělení zavedeme v následující definici.

Definice 8.1. Nechť $\boldsymbol{Z} = (Z_1, \dots, Z_n)^T$, kde $Z_i \sim \mathsf{N}(0,1), i = 1, \dots, n$. Nechť $\mathbf{a} \in \mathbb{R}^m$ je vektor konstant, nechť $\mathbf{B}$ je matice konstant typu (m,n). Označme $\mathbf{V} = \mathbf{BB}^T$. Řekneme, že náhodný vektor

$$\boldsymbol{U} = \mathbf{a} + \mathbf{B}\boldsymbol{Z} \tag{8.9}$$

má **m-rozměrné normální rozdělení** $\mathsf{N}_m(\mathbf{a}, \mathbf{V})$.

Na tomto místě se nebudeme zabývat možná překvapivým zjištěním, že mnohorozměrné normální rozdělení je určeno nikoliv maticí $\mathbf{B}$, ale maticí $\mathbf{V} = \mathbf{BB}^T$. Je-li $\mathbf{V}$ je pozitivně semidefinitní (nebo pozitivně definitní), odpovídající matice $\mathbf{B}$ existuje vždy, dokonce není maticí $\mathbf{V}$ dána jednoznačně (viz např. [21]).

Ze vztahů (6.11) a (7.16) plyne, že $\mathbf{a}$ je vektor středních hodnot a $\mathbf{V} = \mathbf{BB}^T$ je varianční matice náhodného vektoru $\boldsymbol{U}$. Pokud je matice $\mathbf{V}$ regulární, existuje hustota náhodného vektoru $\boldsymbol{U}$ a má tvar (viz např. [1], [2])

$$f(u_1, \dots, u_m) = (2\pi)^{-m/2}|\mathbf{V}|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{u}-\mathbf{a})^T\mathbf{V}^{-1}(\mathbf{u}-\mathbf{a})\right). \tag{8.10}$$

Pro další úvahy je důležitá následující vlastnost náhodného vektoru $\boldsymbol{U}$ s mnohorozměrným normálním rozdělením.

Věta 8.4. Nechť $\boldsymbol{U}$ má mnohorozměrné normální rozdělení $\mathsf{N}_m(\mathbf{a}, \mathbf{V})$, nechť $\mathbf{c} \in \mathbb{R}^k$ je vektor konstant, nechť $\mathbf{D}$ je matice konstant typu (k,m). Potom $\mathbf{c} + \mathbf{D}\boldsymbol{U}$ má mnohorozměrné normální rozdělení $\mathsf{N}_k\left(\mathbf{c} + \mathbf{Da}, \mathbf{DVD}^T\right)$.

D ů k a z: Matici $\mathbf{V}$ vyjádříme jako $\mathbf{V} = \mathbf{BB}^T$. Použijeme-li vyjádření (8.9), dostaneme

$$\begin{aligned}
\mathbf{c} + \mathbf{D}\boldsymbol{U} &= \mathbf{c} + \mathbf{D}(\mathbf{a} + \mathbf{B}\boldsymbol{Z}) \\
&= (\mathbf{c} + \mathbf{Da}) + (\mathbf{DB})\boldsymbol{Z} \\
&\sim \mathsf{N}_n\left(\mathbf{c} + \mathbf{Da}, \mathbf{DBB}^T\mathbf{D}^T\right).
\end{aligned}$$

□

Speciálně z právě dokázané věty plyne, že marginální rozdělení podvektoru vektoru s mnohorozměrným normálním rozdělením má opět mnohorozměrné normální rozdělení.

Zvolíme-li matici $\mathbf{D}$ o pouhém jediném řádku, zjistíme, že každá lineární funkce složek náhodného vektoru s mnohorozměrným normálním rozdělením má normální rozdělení. Lze dokázat, že náhodný vektor $\boldsymbol{U}$ má mnohorozměrné normální rozdělení, právě když každá lineární funkce náhodného vektoru $\boldsymbol{U}$ má mnohorozměrné normální rozdělení (viz např. [15]).

8.4 Přehled rozdělení odvozených od normálního

$Z_1, \dots, Z_k \sim \mathsf{N}(0,1)$, nezávislé

$X_k^2 = \sum_{j=1}^{k} Z_j^2 \sim \chi^2(k)$ — chí-kvadrát o k stupních volnosti

$X_k^2 \sim \chi^2(k), X_m^2 \sim \chi^2(m), Z \sim \mathsf{N}(0,1)$, nezávislé

$F_{k,m} = \dfrac{X_k^2/k}{X_m^2/m} \sim \mathsf{F}(k,m)$ — (Fisherovo-Snedecorovo) F rozdělení s k a m stupni volnosti

$T_k = \dfrac{Z}{\sqrt{X_k^2/k}} \sim \mathsf{t}(k)$ (Studentovo) t-rozdělení s k stupni volnosti

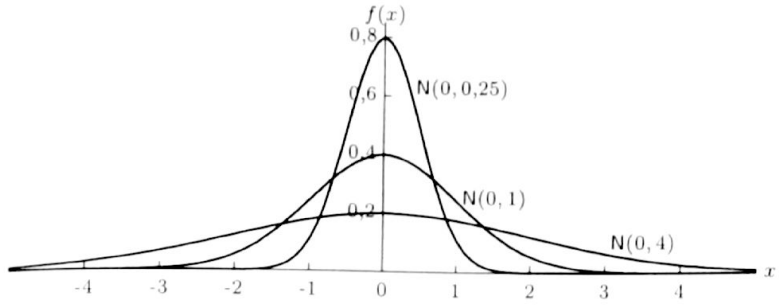
zřejmě platí $Z^2 \sim \chi^2(1), \quad (T_k)^2 \sim \mathsf{F}(1,k)$

Rozdělení	Označení	Hustota
normální	$\mathsf{N}(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-(x-\mu)^2/(2\sigma^2)}$
χ^2 rozdělení	$\chi^2(k)$	$\frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}\mathrm{e}^{-x/2} \quad (x>0)$
Studentovo t	$\mathsf{t}(k)$	$\frac{\Gamma((k+1)/2)}{\Gamma(k/2)\sqrt{\pi k}}\left(1+\frac{x^2}{k}\right)^{-(k+1)/2}$
Fisherovo F	$\mathsf{F}(k,m)$	$\frac{\Gamma((k+m)/2)}{\Gamma(k/2)\Gamma(m/2)}\left(\frac{k}{m}\right)^{k/2}x^{k/2-1}\left(1+\frac{k}{m}x\right)^{-(k+m)/2}$ $(x>0)$

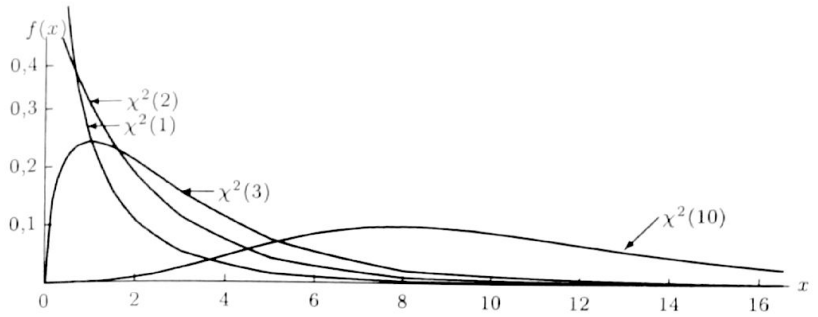
Tabulka 8.1: Hustoty rozdělení

Rozdělení	Krit. hodnota	Stř. hodnota	Rozptyl
$\mathsf{N}(\mu,\sigma^2)$	$\mathsf{P}[Z > z(p)] = p$	μ	σ^2
$\chi^2(k)$	$\mathsf{P}[X_k^2 > \chi_k^2(p)] = p$	k	$2k$
$\mathsf{t}(k)$	$\mathsf{P}[\lvert T_k\rvert > t_k(p)] = p$	$0 \quad (k>1)$	$\frac{k}{k-2} \quad (k>2)$
$\mathsf{F}(k,m)$	$\mathsf{P}[F_{k,m} > F_{k,m}(p)] = p$	$\frac{m}{m-2} \quad (m>2)$	$\frac{2m^2(k+m-2)}{k(m-2)^2(m-4)}$ $(m>4)$

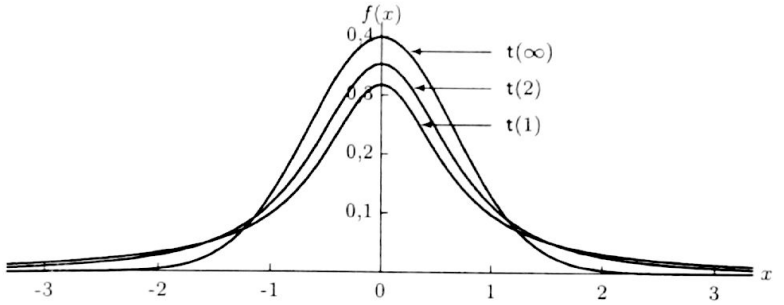
Tabulka 8.2: Označení kritických hodnot, střední hodnoty a rozptyly



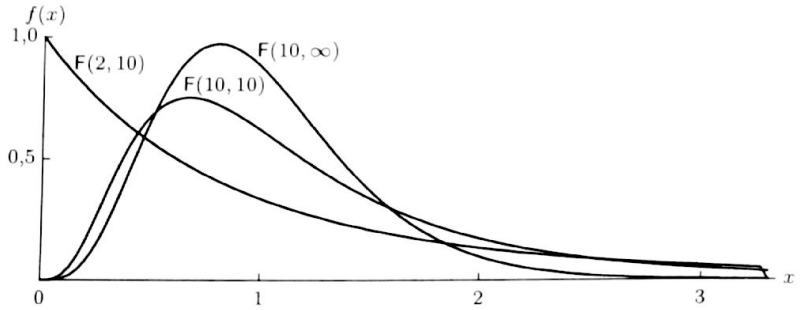
Obrázek 8.1: Hustoty normálního rozdělení s různými rozptyly



Obrázek 8.2: Hustoty χ^2 rozdělení



Obrázek 8.3: Hustoty rozdělení $\mathsf{t}(f)$



Obrázek 8.4: Hustoty rozdělení $\mathsf{F}(k, m)$

9. Asymptotické vlastnosti

9.1 Čebyševova nerovnost

Věta 9.1. Nechť $\mathsf{var}X < \infty$, nechť $\epsilon > 0$. Potom platí

$$\mathsf{P}[|X - \mathsf{E}X| \geq \epsilon] \leq \frac{\mathsf{var}X}{\epsilon^2}. \tag{9.1}$$

D ů k a z: Důkaz provedeme pro spojité rozdělení, pro diskrétní rozdělení by byl důkaz analogický. Označme $\mu = \mathsf{E}X$. Postupně budeme zdola odhadovat rozptyl náhodné veličiny X:

$$\begin{aligned}
\mathsf{var}X &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\mathrm{d}x \\
&= \int_{x:|x-\mu|\geq\epsilon} (x-\mu)^2 f(x)\mathrm{d}x + \int_{x:|x-\mu|<\epsilon} (x-\mu)^2 f(x)\mathrm{d}x \\
&\geq \int_{x:|x-\mu|\geq\epsilon} \epsilon^2 f(x)\mathrm{d}x \\
&= \epsilon^2 \mathsf{P}[|X-\mu| \geq \epsilon],
\end{aligned}$$

což je s dokazovanou nerovností ekvivalentní. □

Zvláště důležitý je speciální případ založený na binomickém rozdělení.

Věta 9.2. (Bernoulli) Nechť $X_n, n = 1, 2, \ldots$, jsou náhodné veličiny s binomickým rozdělením s parametry n, p, kde je $0 < p < 1$. Pro libovolné $\epsilon > 0$ platí

$$\lim_{n\to\infty} \mathsf{P}\left[\left|\frac{X_n}{n} - p\right| \geq \epsilon\right] = 0. \tag{9.2}$$

D ů k a z: Pro každé n platí

$$\begin{aligned}
\mathsf{E}\frac{X_n}{n} &= \frac{np}{n} = p, \\
\mathsf{var}\frac{X_n}{n} &= \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n},
\end{aligned}$$

takže po dosazení do (9.1) dostaneme

$$\mathsf{P}\left[\left|\frac{X_n}{n} - p\right| \geq \epsilon\right] \leq \frac{p(1-p)}{n\epsilon^2},$$

odkud je tvrzení zřejmé. □

Věta dává smysl tvrzení, že s rostoucím počtem nezávislých pokusů posloupnost relativních četností výskytu jevu A „nějak" konverguje k pravděpodobnosti výskytu tohoto jevu (předpokládá se, že je stejná ve všech pokusech). Zapíšeme-li tvrzení věty pomocí jevu opačného, dostaneme

$$\lim_{n\to\infty} \mathsf{P}\left[\left|\frac{X_n}{n} - p\right| < \epsilon\right] = 1.$$

Ať si tedy zvolíme jakkoli malé $\epsilon > 0$, vždy máme jednotkovou pravděpodobnost, že *v limitě* se relativní četnost od odhadované pravděpodobnosti liší o méně než o toto malé ϵ.

Předpoklad nezávislosti byl zbytečně silný, stačila by nekorelovanost. Podobně jako věta 9.2 se dokáže následují tvrzení.

Věta 9.3. Nechť $X_1, X_2, \ldots$ jsou po dvou nezávislé náhodné veličiny takové, že

$$\mathsf{E}X_i = a, \quad \mathsf{var}X_i < c, \qquad i = 1, 2, \ldots,$$

nechť je $|a| < \infty, c < \infty$. Pro libovolné $\epsilon > 0$ potom platí

$$\lim_{n\to\infty} \mathsf{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - a\right| < \epsilon\right] = 1. \tag{9.3}$$

Platí-li pro libovolné $\epsilon > 0$ vztah (9.3), pak říkáme, že posloupnost průměrů

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

konverguje **podle pravděpodobnosti** ke konstantě a.

Zákon velkých čísel ve formě věty 9.3 nemusí uspokojovat naši představu o tomto přírodním zákonu, která v případě Bernoulliovy posloupnosti nezávislých náhodných veličin $X_i \sim \mathsf{bi}(1, p)$ je vyjádřena výrokem: *Relativní četnost zdaru* $n^{-1}\sum_1^n X_i$ *po* n *pokusech se při* $n \to \infty$ *blíží teoretické pravděpodobnosti* p, a to s pravděpodobností 1. Nulovou pravděpodobnost intuitivně ponecháváme logicky možným ale pravděpodobnostně nemožným sériím, jako je 1,1,1,.... Takové tvrzení jsme schopni dokázat.

Věta 9.4. (Borelův-Cantelliův nula-jedničkový zákon) Buďte F_1, $F_2, \ldots$ náhodné jevy v pravděpodobnostním prostoru $(\Omega, \mathcal{A}, \mathsf{P})$. Označíme $[F_n, \infty \times n] = \{\omega \in \Omega : \omega$ je prvkem nekonečně mnoha množin $F_n\}$. Pak platí

$$\sum_{k=1}^{\infty} \mathsf{P}(F_k) < \infty \quad \Rightarrow \quad \mathsf{P}[F_n, \infty \times n] = 0. \tag{9.4}$$

Jsou-li náhodné jevy $F_1, F_2, \ldots$ nezávislé, pak

$$\mathsf{P}[F_n, \infty \times n] = 0 \quad \Leftrightarrow \quad \sum_{n=1}^{\infty} \mathsf{P}(F_n) < \infty, \tag{9.5}$$

tj.

$$\mathsf{P}[F_n, \infty \times n] = 1 \quad \Leftrightarrow \quad \sum_{n=1}^{\infty} \mathsf{P}(F_n) = \infty, \tag{9.6}$$

D ů k a z: Platí

$$[F_n, \infty \times n] = \{\omega \in \Omega : \forall_{n\in\mathbb{N}} \exists_{k(n)\geq n} : \omega \in F_{k(n)}\} = \bigcap_{n-1}^{\infty} \bigcup_{k\geq n} F_k.$$

Tento vztah předně ukazuje, že množina $[F_n, \infty \times n]$ je náhodný jev v σ-algebře $\mathcal{A}$ a má tudíž pravděpodobnost. Dokažme (9.4). Je-li $\sum_{k=1}^{\infty} \mathsf{P}[F_k] < \infty$, plyne odtud, že

$$\mathsf{P}[F_n, \infty \times n] = \lim_{n\to\infty} \mathsf{P}\left(\bigcup_{k\geq n} F_k\right) \leq \lim_{n\to\infty} \sum_{k=n}^{\infty} \mathsf{P}(F_k) = 0,$$

kde jsme nejprve použili (1.37) z věty 1.5 pro množiny $A_n = \cup_{k\geq n} F_k$, pak odhad (1.38) a nakonec tu skutečnost, že zbytky konvergentní řady konvergují k nule.

Je lhostejné, kterou z ekvivalencí (9.5) nebo (9.6) budeme dokazovat. Vzhledem k (9.4) stačí (pro nezávislé jevy $F_1, F_2, \ldots$) ukázat, že platí implikace

$$\sum_{k=1}^{\infty} \mathsf{P}(F_k) = \infty \quad \Rightarrow \quad \mathsf{P}[F_n, \infty \times n] = 1.$$

Buďte $m > n$ přirozená čísla. Podle tvrzení (b) věty 2.3 je

$$\begin{aligned} \mathsf{P}\left(\bigcup_{k=n}^{m} F_k\right) &= 1 - \prod_{k=n}^{m} (1 - \mathsf{P}(F_k)) \\ &\geq 1 - \prod_{k=n}^{m} \exp\left(-\mathsf{P}(F_k)\right) \\ &= 1 - \exp\left(-\sum_{k=n}^{m} \mathsf{P}(F_k)\right), \end{aligned}$$

protože $1 - x \le \mathrm{e}^{-x}$ všude na $(-\infty, \infty)$. Odtud (podle (1.36) ve větě 1.5) vypočteme pro $n \ge 1$:

$$\mathsf{P}\left(\bigcup_{k=n}^{\infty} F_k\right) = \lim_{m\to\infty} \mathsf{P}\left(\bigcup_{k=n}^{m} F_k\right) \ge \lim_{m\to\infty} \left(1 - \exp\left(-\sum_{k=n}^{m} \mathsf{P}(F_k)\right)\right) = 1,$$

protože je $\sum_{k=n}^{\infty} \mathsf{P}(F_k) = +\infty$. Závěrem, opět podle (1.37), dostaneme

$$\mathsf{P}[F_n, \infty \times n] = \lim_{n\to\infty} \mathsf{P}\left(\bigcup_{k=n}^{\infty} F_k\right) = 1. \qquad \square$$

Borelův-Cantelliův nula-jedničkový zákon má v teorii pravděpodobnosti význam sám o sobě, jak uvidíme později v příkladech. Nyní nám poslouží při důkazu zákona velkých čísel.

Věta 9.5. (Borelův silný zákon velkých čísel) Buďte $X_1, X_2, \ldots$ nezávislé stejně rozdělené náhodné veličiny s konečnou střední hodnotou $\mathsf{E}X_1 = \mu$. Potom platí

$$\mathsf{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} X_j(\omega) = \mu\right\}\right) = 1.$$

D ů k a z: Důkaz provedeme s jistou újmou na obecnosti, budeme předpokládat, že pro všechna $n \ge 1$ platí $|X_n| \le c < \infty$, tj. že veličiny X_n jsou stejnoměrně omezené. Označme $\sigma^2 = \mathsf{var}X_n$. Buďtež $\varepsilon > 0$ a $F_n^{\varepsilon} = \left[\left|\frac{1}{n^2}\sum_{j=1}^{n^2}(X_j - \mu)\right| \ge \varepsilon\right]$. Pomocí Čebyševovy nerovnosti dostáváme

$$\mathsf{P}(F_n^{\varepsilon}) \le \varepsilon^{-2}\frac{n^2\sigma^2}{n^4} \quad \Rightarrow \quad \sum_{n=1}^{\infty} \mathsf{P}(F_n^{\varepsilon}) < \infty,$$

což s použitím věty 9.4 vede k $\mathsf{P}[F_n^{\varepsilon}, \infty \times n] = 0$. Nyní (pozor – obtížné místo důkazu)

$$\begin{aligned} N &= \left\{\omega \in \Omega : \lim_{n\to\infty} n^{-2} \sum_{j=1}^{n^2} (X_j - \mu) \ne 0\right\} \\ &= \left\{\omega \in \Omega : \exists_{\varepsilon(\omega)} > 0 : \left|n^{-2}\sum_{j=1}^{n^2}(X_j - \mu)\right| \ge \varepsilon(\omega)\right. \\ &\qquad\qquad \left.\text{pro nekonečně mnoho } n \in \mathbb{N}\right\}. \end{aligned}$$

Jinými slovy:

$$N = \bigcup_{\varepsilon>0} [F_n^{\varepsilon}, \infty \times n] = [F_n^1, \infty \times n] \cup [F_n^{1/2}, \infty \times n] \cup [F_n^{1/3}, \infty \times n] \cup \ldots$$

a

$$\mathsf{P}(N) \quad = \quad \mathsf{P}[F_n^1, \infty \times n] + \mathsf{P}[F_n^{1/2}, \infty \times n] + \mathsf{P}[F_n^{1/3}, \infty \times n] + \ldots = \quad 0,$$

a podle (1.38) ve větě 1.5. Je proto $\lim_{n\to\infty} n^{-2}\sum_{j=1}^{n^2}(X_j(\omega) - \mu) = 0$ pro ω z množiny $\Omega - N$, která má pravděpodobnost 1, což ovšem znamená, že také pro tyto elementární jevy platí $\lim_{n\to\infty} n^{-2}\sum_{j=1}^{n^2} X_j(\omega) = \mu$. Dokázali jsme tvrzení věty pro vybranou posloupnost aritmetických průměrů $n^{-2}\sum_{j=1}^{n^2} X_j$. Při stejné omezenosti sčítanců X_j to však již znamená, že jsme s důkazem hotovi, protože platí implikace

$$n^2 \le k < (n+1)^2 \quad \Rightarrow \quad 0 \le k - n^2 \le 2n,$$

takže

$$\begin{aligned} &\left|k^{-1}\sum_{j=1}^{k} X_j - n^{-2}\sum_{j=1}^{n^2} X_j\right| \\ &\le \left|k^{-1}\sum_{j=1}^{k} X_j - n^{-2}\sum_{j=1}^{k} X_j\right| + \left|n^{-2}\sum_{j=1}^{k} X_j - n^{-2}\sum_{j=1}^{n^2} X_j\right| \\ &\le \frac{(k-n^2)c}{n^2} + \frac{(k-n^2)c}{n^2} \le \frac{4c}{n}, \end{aligned}$$

odkud plyne

$$\lim_{n\to\infty} n^{-1}\sum_{j=1}^{n} X_j = \lim_{n\to\infty} n^{-2}\sum_{j=1}^{n^2} X_j,$$

pokud aspoň jedna z limit existuje. $\square$

Všimněme si, že dodatečný předpoklad $|X_n| \le c$ byl skutečně na újmu obecnosti důkazu. Vlastně jsme nepotřebovali nic více, než platnost Čebyševovy nerovnosti a tedy jako při důkazu věty 9.3 nic více, než nezávislost po dvou (dokonce nekorelovanost) náhodných veličin X_n. Pro platnost zákona velkých čísel ve formě věty 9.5 je ovšem stochastická nezávislost veličin velmi podstatná (viz [18, str. 253]).

Vyzkoušíme účinnost právě dokázaných hlubokých tvrzení klasické teorie pravděpodobnosti na některých příkladech.

Příklad 9.1. Při opakovaných hodech symetrickou mincí bychom sérii jednoho milionu po sobě jdoucích hodů, při nichž by pokaždé padl líc, považovali za něco podobného zázraku. Pravděpodobnost $p = 2^{-10^6}$ takové události je opravdu malá, nicméně je $p > 0$, takže podle nula-jedničkového zákona při nekonečně (!) mnoha hodech symetrickou mincí bude s pravděpodobností 1 registrováno nekonečně mnoho takových sérií. Formálně: Nechť F_n označuje náhodný jev, který spočívá v tom, že v hodech s pořadovými čísly $n, n+1, \ldots, n+10^6-1$ padl líc. Náhodný jev $[F_n, \infty \times n]$ přesně popisuje situaci, která nás zajímá. Samozřejmě, že je $\mathsf{P}(F_n) = \sum_1^\infty p = +\infty$, ale náhodné jevy F_n nejsou nezávislé (časově se překrývají). S tím si ovšem umíme poradit tak, že uvážíme nepřekrývající se série $G_1 = F_1$, $G_2 = F_{10^6+1}$, $G_3 = F_{2\cdot 10^6+1}, \ldots$. Náhodné jevy $G_1, G_2, \ldots$ již nezávislé jsou, $\sum_{n=1}^\infty \mathsf{P}(G_n) = +\infty$ a proto podle tvrzení věty 9.5 dostaneme $\mathsf{P}[F_n, \infty \times n] \geq \mathsf{P}[G_n, \infty \times n] = 1$. ○

Příklad 9.2. Připomeňme model náhodné procházky částice po celočíselné přímce (oddíl 3.7). Předpokládáme-li nekonečný život částice, modelujeme její trajektorii $S_1, S_2, \ldots, S_n, \ldots$ ($S_n \in \mathbb{Z}$ je poloha částice v čase n) posloupností náhodných veličin $S_n = \sum_{j=1}^n X_j$, $n = 1, 2, \ldots$, kde $X_1, X_2, \ldots$, jsou nezávislé náhodné veličiny splňující $\mathsf{P}[X_n = 1] = \mathsf{P}[X_n = -1] = 1/2$, tj. $\mathsf{E}X_n = 0$. Borelův zákon velkých čísel (věta 9.5) tedy říká, že s pravděpodobností 1 platí limitní přechod $\lim_{n\to\infty} n^{-1}S_n = 0$, tj. $S_n = o(n)$ při $n \to \infty$. (Lze ukázat, že $S_n = o(n^{1/2+\delta})$ pro každé $\delta > 0$ a že $S_n \neq o(n^{1/2})$ s pravděpodobností 1.) ○

Zákon velkých čísel má pro matematickou statistiku význam spíše principiální, než technický. Ubezpečuje nás, že při opakovaných nezávislých pozorováních hodnoty náhodné veličiny X dokážeme pomocí aritmetického průměru odhadnout neznámou střední hodnotu $\mathsf{E}X$ s libovolnou přesností, budeme-li ovšem schopni pozorovat dosti dlouho. To však není případ experimentálních věd, kde jsou statistické metody aplikovány. Zákon velkých čísel poskytuje pouze základní orientaci pro konstrukci odhadu.

Příklad 9.3. Buďte $X_1, X_2, \ldots$, nezávislé náhodné veličiny s $X_n \sim \mathsf{N}(0, \sigma^2)$. Podle věty 9.5 je s pravděpodobností 1 platný limitní přechod $\frac{1}{n}\sum_{j=1}^n X_j^2 \to \sigma^2$ při $n \to \infty$, protože $\mathsf{E}X_j^2 = \sigma^2$. Aritmetický průměr $n^{-1}\sum_{j=1}^n X_j^2$ se tedy nabízí jako odhad pro neznámý rozptyl σ^2, máme-li k disposici nezávislá pozorování $X_1, \ldots, X_n$ náhodné veličiny s rozdělením $\mathsf{N}(0, \sigma^2)$. Statistik ovšem učiní důkladnější indukci o velikosti σ^2, když využije skutečnost, že platí $\sum_{j=1}^n \left(\frac{X_j}{\sigma}\right)^2 \sim \chi^2(n)$. ○

Současná matematika využívá ve stále větší míře pravděpodobnostní metody jako velmi účinnou důkazovou techniku. Dokážeme takto známou Weierstrassovu větu.

Příklad 9.4. (Weierstrassova věta) Buď $f(p)$ spojitá funkce na intervalu $\langle 0, 1\rangle$. Pak

$$B_n(p) = \sum_{k=0}^n f\left(\frac{k}{n}\right)\binom{n}{k}p^k(1-p)^{n-k}$$

(Bernsteinovy polynomy) je posloupnost funkcí, která konverguje stejnoměrně k funkci $f(x)$ na intervalu $\langle 0, 1\rangle$.

Pravděpodobnostní důkaz: Funkce f je stejnoměrně spojitá a omezená na intervalu $\langle 0, 1\rangle$, tj. platí $|f(p)| \leq M < \infty$ pro $p \in \langle 0, 1\rangle$ a ke každému $\varepsilon > 0$ existuje $\delta > 0$ tak, že $|f(p) - f(p')| < \varepsilon$ pro $|p - p'| < \delta$. Zvolme $\varepsilon > 0$ libovolné a $p \in \langle 0, 1\rangle$. Pak

$$\begin{aligned}
|f(p) - B_n(p)| &= \left|\sum_{k=0}^n \left(f(p) - f\left(\frac{k}{n}\right)\right)\binom{n}{k}p^k(1-p)^{n-k}\right| \\
&\leq \sum_{\{k:|\frac{k}{n}-p|<\delta\}} \left|f(p) - f\left(\frac{k}{n}\right)\right|\binom{n}{k}p^k(1-p)^{n-k} \\
&\quad + \sum_{\{k:|\frac{k}{n}-p|\geq\delta\}} \left|f(p) - f\left(\frac{k}{n}\right)\right|\binom{n}{k}p^k(1-p)^{n-k} \\
&\leq \varepsilon\cdot 1 + 2M \sum_{\{k:|k-np|\geq n\delta\}} \binom{n}{k}p^k(1-p)^{n-k} \\
&= \varepsilon + 2M\mathsf{P}[|X - \mathsf{E}X| \geq \delta] \quad (\text{ kde} X \sim \mathsf{bi}(n,p)).
\end{aligned}$$

Použijeme-li Čebyševovu nerovnost, pak

$$|f(p) - B_n(p)| \leq \varepsilon + 2M\frac{np(1-p)}{n^2\delta^2} \leq \varepsilon + \frac{2M}{n\delta^2},$$

tj.

$$\lim_{n\to\infty} \max_{p\in\langle 0,1\rangle} |f(p) - B_n(p)| \leq \varepsilon.$$

Jelikož $\varepsilon > 0$ bylo zvoleno libovolně, důkaz je ukončen. ○

Vlastně jsme se nikdy nezabývali otázkou, lze-li konstruovat posloupnost nezávislých náhodných veličin s předepsaným rozdělením pravděpodobností, můžeme-li tedy například konstruovat model náhodné procházky ve smyslu příkladu 9.2.

Příklad 9.5. Existuje Kolmogorovův pravděpodobnostní prostor uvnitř kterého žijí nezávislé náhodné jevy $F_1, F_2, \ldots$ s $\mathsf{P}(F_n) = 1/2$, tj. takový pravděpodobnostní prostor, který modeluje nekonečnou sérii Bernoulliovských pokusů s pravděpodobností zdaru $p = 1/2$. Položíme $\Omega = \langle 0, 1\rangle$, $\mathcal{A} =$ [Borelovské podmnožiny Ω] (jde o nejmenší σ-algebru obsahující všechny intervaly v Ω) a P nechť je pravděpodobnost na $\mathcal{A}$ *jednoznačně* definovaná tím, že pravděpodobnost intervalu $\langle a, b\rangle \subset \Omega$ je dána jeho délkou $b - a$ (v matematice se nazývá Lebesgueova míra). Položíme-li

$$F_1 = \left\langle \frac{1}{2}, 1 \right\rangle, F_2 = \left\langle \frac{1}{4}, \frac{1}{2} \right\rangle \cup \left\langle \frac{3}{4}, 1 \right\rangle, \ldots,$$

$$F_k = \bigcup_{\{1 \le i \le 2^n, i \text{ sudé}\}} \left\langle \frac{i-1}{2^n}, \frac{i}{2^n} \right\rangle, \ldots,$$

pak $\mathsf{P}(F_k) = 1/2$ a $\mathsf{P}\left(\cap_{j=1}^{l} F_j\right) = \left(\frac{1}{2}\right)^l = \prod_{j=1}^{l} \mathsf{P}(F_j)$ pro $l = 1, 2, \ldots$. Odtud (indukcí) plyne nezávislost náhodných jevů F_n.

Zapišme nyní číslo $x \in \langle 0, 1\rangle$ v jeho dvojkovém rozvoji jako

$$x = \sum_{n=1}^{\infty} \frac{X_n(x)}{2^n},$$

kde $X_n(x)$ je 0 nebo 1; pro určitost zvolíme ten dvojkový rozvoj, který má nekonečně mnoho jedniček. Pak $X_1, X_2, \ldots$ jsou nezávislé náhodné veličiny na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, \mathsf{P})$ s $X_n \sim \mathsf{bi}(1, 1/2)$. Argument je zřejmý: $[X_1 = 1] = F_1$, $[X_2 = 1] = F_2, \ldots$, $[X_k = 1] = F_k, \ldots$. Z vět 9.4 a 9.5 potom plynou dvě dosti hluboká tvrzení:

Vybereme-li náhodně bod $x \in \langle 0, 1\rangle$ (náhodně ve smyslu Lebesgueovy pravděpodobnosti P), pak:

(a) Ve dvojkovém rozvoji čísla x je nekonečně mnoho nul a nekonečně mnoho jedniček.

(b) Je-li $f_n(x)$ relativní četnost jedniček mezi prvními n členy dvojkového rozvoje x, pak platí

$$\lim_{n\to\infty} f_n(x) = 1/2.$$

Skutečně: Tvrzení (a) platí pro x, která náleží náhodnému jevu

$$A = [F_n, \infty \times n] \cap [F_n^c, \infty \times n],$$

tvrzení (b) pro x, která náleží náhodnému jevu

$$B = \left\{ x \in \Omega : n^{-1} \sum_{j=1}^{n} X_j(x) \to 1/2 \right\}.$$

Věta 9.4 implikuje $\mathsf{P}(A) = 1$, věta 9.5 říká, že $\mathsf{P}(B) = 1$, což je předmětem tvrzení (a), resp. tvrzení (b). ○

9.2 Centrální limitní věta

Uvažujme nezávislé náhodné veličiny $U_1, U_2, \ldots$, které mají všechny stejné rozdělení s nulovou střední hodnotou, jednotkovým rozptylem a konstantou M omezeným třetím momentem. Nemusí jít o normální rozdělení. Posloupnost průměrů, jak již víme, konverguje podle pravděpodobnosti k nule. Pokud nás zajímá asymptotické rozdělení, nesmíme součet dělit tak velkou hodnotou, jako je n. Zkusme tedy dát do jmenovatele pouze $\sqrt{n}$. Označme

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i.$$

Označíme-li jako $M_U(t)$ momentovou vytvořující funkci náhodných veličin U_i, lze momentovou vytvořující funkci náhodné veličiny S_n zapsat jako

$$M_{S_n}(t) = \mathsf{E}e^{(t/\sqrt{n})\sum U_i} = \prod_{i=1}^{n} \mathsf{E}e^{(t/\sqrt{n})U_i} = \left(M_U(t/\sqrt{n})\right)^n$$

pro $t \in (-b, b)$. Použijme nyní tvrzení (7.24) věty 7.7. Protože předpokládáme omezený třetí moment, můžeme pro libovolné $t \in (-b, b)$ psát

$$M_U(t/\sqrt{n}) = 1 + 0 \cdot \frac{t}{\sqrt{n}} + 1 \cdot \frac{t^2}{2n} + \mathrm{o}\left(\frac{t^2}{n}\right).$$

V limitě má tedy vytvořující funkce S_n tvar

$$\lim_{n\to\infty} M_{S_n}(t) = \lim_{n\to\infty} \left(1 + \frac{t^2}{2n} + \mathrm{o}\left(\frac{1}{n}\right)\right)^n = e^{\frac{t^2}{2}},$$

což je, jak známo z příkladu 7.6, momentová vytvořující funkce rozdělení $\mathsf{N}(0, 1)$.

Asymptoticky má tedy náhodná veličina S_n normované normální rozdělení. Úvahu můžeme poněkud rozšířit pomocí normování (viz (7.4)).

Věta 9.6. Nechť $Y_1, Y_2, \ldots$ je posloupnost nezávislých náhodných veličin, pro které platí

$$\begin{aligned} \mathsf{E}Y_i &= \mu, \\ \mathsf{var}Y_i &= \sigma^2 > 0, \\ \mathsf{E}|Y_i|^3 &< \infty. \end{aligned}$$

Pro distribuční funkci náhodné veličiny

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma} \right)$$

platí

$$\lim_{n\to\infty} \mathsf{P}[S_n < x] = \Phi(x). \tag{9.7}$$

Pro snazší zapamatování výrazu S_n si uvědomme, že za uvedených předpokladů má součet $X_n = \sum_{i=1}^{n} Y_i$ střední hodnotu $n\mu$ a rozptyl $n\sigma^2$. Výraz S_n tedy dostaneme normováním X_n. Poznamenejme, že tuto větu již známe pro nezávislé náhodné veličiny $Y_1, Y_2, \ldots$ s $Y_n \sim \mathsf{bi}(1,p)$ (Moivreova-Laplaceova integrální věta 4.7). Zopakujme podstatu tohoto tvrzení:

Příklad 9.6. Mějme náhodnou veličinu X s binomickým rozdělením s parametry n, p, $0 < p < 1$. Jak víme z příkladu 6.4, můžeme ji vyjádřit jako součet n nezávislých náhodných veličin Y_i s alternativním rozdělením s parametrem p. Zřejmě je $\mathsf{E}Y_i^3 = p < M = 1$. Součet S_n má pak tvar

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{Y_i - p}{\sqrt{p(1-p)}} \right) = \frac{X - np}{\sqrt{np(1-p)}},$$

jde tedy o normovanou veličinu k binomickému rozdělení. Pro velká n má tato normovaná veličina přibližně rozdělení $\mathsf{N}(0,1)$. Jinak můžeme tuto aproximaci zapsat také jako

$$X \dot{\sim} \mathsf{N}(np, np(1-p)). \tag{9.8}$$

Aproximaci binomického rozdělení normálním považujeme zpravidla za vyhovující, platí-li $np(1-p) > 9$. ○

Příklad 9.7. Zajímá nás neznámý podíl p osob s krevní skupinou A v dané velké populaci. U kolika osob musíme zjistit, zda má či nemá skupinu A, abychom s pravděpodobností aspoň 0,9 odhadli neznámou pravděpodobnost s chybou nejvýše 0,05?

Vyberme z dané populace náhodně n osob. Náhodný počet osob s krevní skupinou A je zřejmě náhodná veličina $X \sim \mathsf{bi}(n,p)$. Neznámý podíl p budeme odhadovat pomocí X/n. Počet oslovených osob n zvolíme tak, aby platilo

$$\mathsf{P}\left[\left|\frac{1}{n}X - p\right| < 0{,}05\right] \doteq 0{,}9.$$

Použijeme-li (9.8), pak postupnými úpravami dostaneme

$$\begin{aligned} 0{,}9 &\doteq \mathsf{P}\left[\left|\frac{1}{n}X - p\right| < 0{,}05\right] \\ &= \mathsf{P}\left[-\frac{0{,}05n}{\sqrt{np(1-p)}} < \frac{X-np}{\sqrt{np(1-p)}} < \frac{0{,}05n}{\sqrt{np(1-p)}}\right] \\ &\doteq \Phi\left(\frac{0{,}05n}{\sqrt{np(1-p)}}\right) - \Phi\left(-\frac{0{,}05n}{\sqrt{np(1-p)}}\right) \\ &= 2\Phi\left(\frac{0{,}05n}{\sqrt{np(1-p)}}\right) - 1. \end{aligned}$$

Má tedy být

$$\Phi\left(\frac{0{,}05n}{\sqrt{np(1-p)}}\right) \doteq (1 + 0{,}9)/2 = 0{,}95,$$

což zaručíme volbou

$$\frac{0{,}05n}{\sqrt{np(1-p)}} \doteq z(0{,}05) = 1{,}644854.$$

Použijeme-li ještě nerovnost $4p(1-p) \le 1$, dostaneme nakonec požadavek $n > 270$. ○

Tvrzení centrální limitní věty 9.6 pro nezávislé stejně rozdělené náhodné veličiny je přesněji vyjádřeno Berry-Essénovou nerovností (viz též věta 4.8):

Věta 9.7. Buďte $Y_1, Y_2, \ldots, Y_n$ nezávislé stejně rozdělené náhodné veličiny s $\mathsf{E}Y_i = \mu$, $\mathsf{var}Y_i = \sigma^2 \in (0, \infty)$. Označíme-li

$$F_n(x) = \mathsf{P}\left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma}\right) < x\right],$$

pak platí

$$|F_n(x) - \Phi(x)| \le 0{,}7975 \cdot n^{-\frac{1}{2}} \frac{\mathsf{E}|Y_i - \mu|^3}{\sigma^3} \quad \text{pro } -\infty < x < +\infty.$$

Důkaz je možno nalézt v [18, str. 282]. Poznamenejme, že tato nerovnost určuje rychlost konvergence (9.7) v centrální limitní větě, jsou-li veličiny $Y_1, Y_2, \ldots$ stejně rozdělené. Řád přiblížení k distribuční funkci $\Phi(x)$

je $O(n^{-1/2})$, při pevném n je normální aproximace tím lepší, čím menší je charakteristika $\sigma^{-3}\mathsf{E}|Y_i - \mu|^3$ společného rozdělení veličin $Y_1, Y_2, \ldots$.

Příklad 9.8. Buď Z_n^2 náhodná veličina s rozdělením $\chi^2(n)$. Bez újmy na obecnosti můžeme předpokládat, že $Z_n^2 = \sum_{i=1}^n X_i^2$, kde $X_1, X_2, \ldots, X_n$ jsou nezávislé náhodné veličiny s $X_j \sim \mathsf{N}(0,1)$. Protože $\mathsf{E}X_i^2 = 1$ a $\mathsf{var}X_i^2 = \mathsf{E}X_i^4 - (\mathsf{E}X_i^2)^2 = 2$, plyne z centrální limitní věty 9.6, že

$$\mathsf{P}\left[\frac{1}{\sqrt{2n}}(Z_n^2 - n) < x\right] \to \Phi(x) \quad \text{při } n \to \infty.$$

Berry-Essénova nerovnost poskytuje odhad:

$$\left|\mathsf{P}\left[\frac{1}{\sqrt{2n}}(Z_n^2 - n) < x\right] - \Phi(x)\right| \le 0{,}7995 \cdot n^{-1/2} \frac{\mathsf{E}|X_1^2 - 1|^3}{(\sqrt{2})^3} \doteq 2{,}457 \cdot n^{-1/2}.$$

○

9.3 Cvičení

9.1. Odhadněte pravděpodobnost, s jakou bude počet šestek, které padnou v 1000 nezávislých hodech ideální kostkou, ležet v mezích od 147 do 186.

9.2. V 10 000 nezávislých hodech mincí padlo celkem 5 087 líců. Rozhodněte, zda lze tuto minci považovat za symetrickou. Návod: Spočítejte pravděpodobnost, s jakou u symetrické mince padne v 10 000 nezávislých hodech padne *aspoň* 5 087 líců.

V následujících cvičeních značí S_n polohu v čase n částice, která koná náhodnou procházku po celočíselné přímce (viz odstavec 3.7).

9.3. Ověřte platnost limitního přechodu

$$\lim_{n\to\infty} \mathsf{P}[S_n < \sqrt{n}x] = \Phi(x), \qquad x \in \mathbb{R}$$

(viz též příklad 4.15), tj. že přibližně platí

$$\frac{S_n}{\sqrt{n}} \sim \mathsf{N}(0,1) \qquad \text{pro } n \to \infty.$$

Návod: Uvědomte si, že $S_n = \sum_{j=1}^n X_j$, kde $X_1, X_2, \ldots$ jsou nezávislé náhodné veličiny takové, že $\mathsf{P}[X_j = 1] = \mathsf{P}[X_j = -1] = 1/2$.

9.4. Označte $M_n = \max_{1\le j\le n} S_j$ (nejvzdálenější bod celočíselné osy navštívený částicí v časovém intervalu $\langle 0,, \rangle$). Ověřte platnost limitního přechodu

$$\lim_{n\to\infty} \mathsf{P}[M_n \ge \sqrt{n}x] = 2(1 - \Phi(x)), \qquad x \ge 0,$$

tj. že náhodná veličina $M_n/\sqrt{n}$ má při velkých hodnotách časového parametru n přibližně takové rozdělení jako náhodná veličina $|X|$, kde $X \sim \mathsf{N}(0,1)$. Návod: Použijte cvičení 9.3, druhou rovnost ze vztahu (3.8) a uvažte, že

$$\lim_{n\to\infty} \frac{[\sqrt{n}x]}{\sqrt{n}} = x.$$

9.5. Ověřte platnost limitního přechodu

$$\lim_{n\to\infty} \sum_{0\le k<x\sqrt{n}+n} \frac{\mathrm{e}^{-n}n^k}{k!} = \Phi(x), \qquad x \in \mathbb{R}.$$

Návod: Uvažte, že

$$\sum_{0\le k<x\sqrt{n}+n} \frac{\mathrm{e}^{-n}n^k}{k!} = \mathsf{P}\left[\frac{X_n - n}{\sqrt{n}} < x\right],$$

kde $X_n \sim \mathsf{Po}(n)$.

10. Popisná statistika

Statistika zkoumá jevy na rozsáhlém souboru případů a hledá ty vlastnosti jevů, které se projevují teprve ve velkém souboru případů, nikoliv v případech jednotlivých. Výchozím pojmem je **statistický soubor**, což je nějaká dobře definovaná množina statistických jednotek. Statistický soubor může být určen seznamem svých prvků (jednotek) nebo může být definován pomocí nějakého pravidla. V pochybnostech musí být možnost ověřit, zda daná jednotka patří do statistického souboru.

Pracuje tedy se **statistickými jednotkami**, na každé se měří jeden nebo několik **statistických znaků**. Měření ovšem chápe poněkud šíře, než v běžném životě. Tomu odpovídá řada *měřítek*, která je vhodné rozlišovat. Uvedeme čtyři typy měřítek, od nejjednoduššího k nejsložitějšímu.

Nominální měřítko předpokládá disjunktní kategorie, které obsáhnou všechny možné hodnoty měření. Jde o obdobu rozkladu jistého jevu v teorii pravděpodobnosti. Mezi jednotlivými hodnotami není žádný vztah, žádné uspořádání. Příkladem může být barva očí ve studiích dědičnosti nebo politická strana při zkoumání volebních preferencí.

Ordinální měřítko je vlastně měřítkem nominálním, v němž přibylo *uspořádání* jednotlivých hodnot. Možné hodnoty tedy můžeme opatřit pořadovými čísly (indexy) s tím, že hodnota s menším indexem předchází každou hodnotu s větším indexem. Příkladem je nejvyšší dosažené vzdělání nebo počet hvězdiček u hotelových kategorií. Zmíněné indexy se někdy používají k zápisu jednotlivých hodnot. Udávají však pouze pořadí těchto hodnot, nikoliv nějakou jejich vzdálenost. Nemá smysl porovnávat rozdíl mezi absolvováním střední a základní školy s rozdílem mezi absolvováním školy vysoké a střední.

Intervalové měřítko už nutně předpokládá číselné označení jednotlivých možných hodnot, což určuje jejich uspořádání, ale předpokládá se, že vzdálenosti mezi sousedními hodnotami jsou konstantní. Podstatné ovšem je, že umístění nuly je pouze dohodnuté, jako je to třeba u teplotních stupnic používaných v běžném životě.

Poměrové měřítko vztahuje velikost měřené veličiny k nějaké dohodnuté jednotce, udává její násobek. Nula tedy znamená neexistenci měřené vlastnosti. Patří sem většina fyzikálních veličin.

Statistické znaky měřené v nominálním či ordinálním měřítku se někdy nazývají **kvalitativní**, znaky měřené v intervalovém či poměrovém měřítku se někdy nazývají **spojité** nebo **kvantitativní**.

Často záleží na účelu, proč provádíme měření. I když můžeme hodnotit například chemickou reakci v podílovém měřítku, mnohdy stačí (a je přiměřeně levnější) použít jednodušší měřítko, například ordinální.

Předpokládejme nyní, že jsme na n statistických jednotkách naměřili **soubor hodnot**

$$x_1, x_2, \ldots, x_n$$

daného znaku. Celkovému počtu prvků souboru říkáme **rozsah** souboru. Jak můžeme hodnoty souboru někomu sdělit, zpracovat je, shrnout?

Pokud jednotlivé hodnoty (v měřítku alespoň ordinálním) srovnáme do neklesající posloupnosti

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}, \tag{10.1}$$

dostaneme **uspořádaný soubor hodnot**. Indexy v závorkách udávají pořadí jednotlivých zjištěných hodnot. Nejmenší z nich jsme tedy označili jako $x_{(1)}$, největší jako $x_{(n)}$.

Pokud je soubor veliký a hodnoty se často opakují, můžeme jej učinit přehlednějším tak, že jej přepíšeme do **četnostní tabulky**, v níž $a_1 < a_2 < \ldots < a_m$ jsou navzájem *různé* uspořádané hodnoty souboru (v případě nominálního měřítka pouze různé hodnoty) a $n_1, n_2, \ldots, n_m$ jsou zjištěné (absolutní) četnosti těchto hodnot. Zřejmě musí platit $n = \sum_{j=1}^{m} n_j$. Typické je takové zpracování u znaků, které nabývají pouze celočíselných hodnot a u znaků kvalitativních. Četnostní tabulku bychom mohli vytvořit pro každý znak, bez ohledu na typ měřítka. Ovšem uspořádání možných hodnot připadá v úvahu pro alespoň ordinální měřítko.

Další výklad se týká pouze znaků kvantitativních. Pokud měřený znak nabývá příliš mnoha různých číselných hodnot, uměle zmenšujeme počet rozlišovaných hodnot tak, že obor hodnot rozdělíme na disjunktní intervaly. Zvolíme například hraniční body $(-\infty \le)\ t_0 < t_1 < \cdots < t_m\ (\le \infty)$ a všechny body z j-tého intervalu (t_{j-1}, t_j) (někdy je vhodnější $\langle t_{j-1}, t_j)$) ztotožníme se středem intervalu $a_j = (t_{j-1} + t_j)/2$. Pokud je $t_0 = -\infty$, zvolíme zpravidla $a_1 = t_1 - (t_2 - t_1)/2$, takže bod t_1 je ve středu intervalu (a_1, a_2). Podobně pro $t_m = \infty$ zpravidla volíme $a_m = t_{m-1} + (t_{m-1} - t_{m-2})/2$. Nejčastěji se volí dělicí body tak, aby intervaly měly (případně až na krajní intervaly) stejnou délku, tedy $h = t_j - t_{j-1}, j = 2, \ldots, m-1$. Když pak určíme počty n_j (třídní četnosti) hodnot x_i, které patří do jednotlivých intervalů (tříd), přejdeme vlastně k četnostní tabulce.

Četnostní tabulku můžeme znázornit pomocí četnostního polygonu, kdy lomenou čarou spojujeme body o souřadnicích a_j, n_j. Častěji znázorňujeme četnostní tabulku do **histogramu**, v němž nad označení hodnoty a_j kreslíme obdélník, jehož výška je úměrná zjištěné četnosti n_j. Pokud kreslíme histogram podle třídních četností a třídní intervaly nemají stejné šířky, určuje

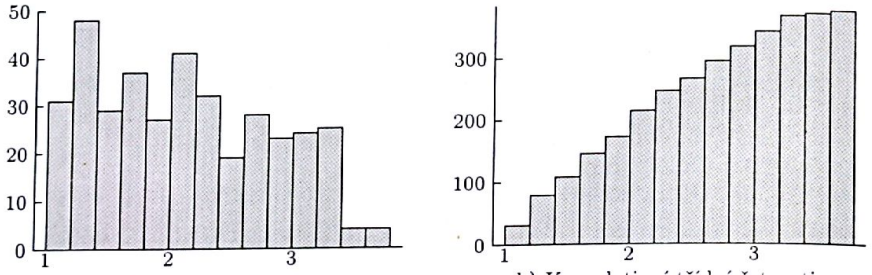
četnost n_j *plochu* obdélníku nad odpovídajícím intervalem. Do uvedených grafů lze místo absolutních četností n_j znázorňovat také **relativní** četnosti $f_j = n_j/n$, případně absolutní nebo relativní četnosti sčítat (kumulovat) a použít buď $\sum_{i=1}^{j} n_i$ nebo $\sum_{i=1}^{j} f_i$ (kumulativní diagramy).

interval $\langle t_{j-1}, t_j)$	střed a_j	četnost n_j	kum. čet. N_j
⟨1,0, 1,2)	1,1	31	31
⟨1,2, 1,4)	1,3	48	79
⟨1,4, 1,6)	1,5	29	108
⟨1,6, 1,8)	1,7	37	145
⟨1,8, 2,0)	1,9	27	172
⟨2,0, 2,2)	2,1	41	213
⟨2,2, 2,4)	2,3	32	245
⟨2,4, 2,6)	2,5	19	264
⟨2,6, 2,8)	2,7	28	292
⟨2,8, 3,0)	2,9	23	315
⟨3,0, 3,2)	3,1	24	339
⟨3,2, 3,4)	3,3	25	364
⟨3,4, 3,6)	3,5	4	368
⟨3,6, 3,8)	3,7	4	372

Tabulka 10.1: Třídní četnosti a kumulativní třídní četnosti průměrného prospěchu

Příklad 10.1. V tabulce 10.1 jsou uvedeny zjištěné třídní četnosti průměrných známek na výročním vysvědčení celkem 372 dětí. Odpovídající histogramy jsou uvedeny na obrázku 10.1. ○

Pokud třídíme data do intervalů, používáme zpravidla intervaly konstantní šířky. Při volbě počtu intervalů můžeme vycházet například ze Sturgesova pravidla, podle kterého volíme $m \doteq 1 + 3{,}3 \log_{10}(n) \doteq 1 + 1{,}43 \log_e(n)$. Této hodnoty se přidržujeme jen přibližně, dbáme především na to, aby délky intervalů byly dostatečně okrouhlé, stejně jako středy či krajní body těchto intervalů.



a) Třídní četnosti b) Kumulativní třídní četnosti

Obrázek 10.1: Histogram třídních četností a kumulativních třídních četností průměrného prospěchu

10.1 Míry polohy

Míry polohy udávají hodnotu, kolem které se jednotlivá pozorování shromažďují.

Průměr (též výběrový průměr)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{10.2}$$

$$= \frac{1}{n}\sum_{j=1}^{m} n_j a_j \tag{10.3}$$

vyžaduje aspoň intervalové měřítko a závisí stejně na všech hodnotách statistického znaku. Zřejmě platí (pro libovolná a, b)

$$\overline{(a + bx)} = a + b\bar{x}, \tag{10.4}$$

takže se přirozeně mění se změnou měřítka.

Geometrický průměr

$$\bar{x}_G = \sqrt[n]{x_1 x_2 \cdots x_n}$$

má smysl, jen když jsou všechny hodnoty znaku kladné. Je obdobou průměru aritmetického, neboť jeho logaritmus je aritmetickým průměrem logaritmů. Geometrický průměr není invariantní vůči lineární transformaci. Používá se tam, kde jde o násobení. Je-li například inflace v pěti po sobě jdoucí letech postupně 20%, 50%, 30%, 20% a 5%, je to totéž, jako kdyby v každém z oněch pěti roků byla inflace přibližně 24%. Je totiž

$$\sqrt[5]{1{,}20 \cdot 1{,}50 \cdot 1{,}30 \cdot 1{,}20 \cdot 1{,}05} \doteq 1{,}24.$$

Harmonický průměr má smysl jen pro kladné hodnoty znaku. Je definován vztahem

$$\bar{x}_H = \left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)\right)^{-1},$$

takže opět není invariantní vůči lineární transformaci.

Lze dokázat (viz např. [2]), že pokud jsou všechny hodnoty znaku kladné, platí nerovnosti

$$\bar{x}_H \le \bar{x}_G \le \bar{x}.$$

Medián (též výběrový medián) je definován pomocí uspořádaného souboru hodnot (10.1) jako

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & n \text{ liché}, \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & n \text{ sudé}. \end{cases} \tag{10.5}$$

Jde o hodnotu, která dělí uspořádané hodnoty (10.1) na dva stejně početné díly. Nezáleží tedy na konkrétních hodnotách prvních ani posledních členů uspořádaného souboru hodnot. Medián má stejnou vlastnost, jako průměr (viz (10.4)), pro libovolná a, b platí

$$\widetilde{(a+bx)} = a + b\tilde{x}.$$

Je-li g monotonní funkce, pak platí analogická vlastnost pro transformované hodnoty $g(x_i)$. Při lichém n platí tato vlastnost přesně, při sudém n téměř přesně ($g(\tilde{x})$ není obecně průměrem hodnot $g(x_{(n/2)}), g(x_{(n/2+1)})$. Medián má při lichém n smysl už pro ordinální měřítko, k výpočtu mediánu při sudém n potřebujeme měřítko aspoň intervalové. Pokud bychom při sudém n definovali jako medián jakoukoliv hodnotu znaku, která splňuje nerovnost $x_{(\frac{n}{2})} \le \tilde{x} \le x_{(\frac{n}{2}+1)}$, ztratili bychom sice jednoznačnost definice z (10.5), ale byla by použitelná i pro ordinální měřítko.

Medián můžeme zobecnit. Místo, aby odděloval polovinu nejmenších dat od ostatních, může oddělovat p-tý díl dat. Zvolme p, $0 < p < 1$. Definujeme (výběrový) p-tý **kvantil (percentil)** vztahem

$$x_p = \begin{cases} x_{([np]+1)} & np \neq [np] \\ \frac{1}{2}\left(x_{(np)} + x_{(np+1)}\right) & np = [np] \end{cases} \tag{10.6}$$

Pokud není výraz np celočíselný, použije se hodnota z uspořádaného seznamu hodnot s nejbližším *větším* indexem. Pokud je výraz np celočíselný, použije se průměr ze dvou hodnot (u ordinálního měřítka můžeme použít libovolnou hodnotu, která leží mezi $x_{([np])}$ a $x_{([np]+1)}$). Medián je speciálním případem výběrového kvantilu: $\tilde{x} = x_{0,5}$. Především v grafických pomůckách se používají **dolní** a **horní kvartil**

$$Q_1 = x_{0,25},\ Q_3 = x_{0,75}.$$

Výběrové kvantily mají stejný obor použití jako medián.

Modus je nejčetnější hodnotou. Má smysl zejména tehdy, kdy je počet m skutečně se vyskytujících hodnot podstatně menší než rozsah souboru n. Modus je použitelný při každém měřítku (i když u nominálního měřítka je těžké hovořit o míře polohy) a nemusí být určen jednoznačně.

10.2 Míry variability

Míry variability charakterizují velikost variability hodnot kolem nějaké její míry polohy nebo velikost jejich vzájemné rozdílnosti. Základním požadavkem by měla být invariance vůči posunutí, neboť přičtením stejné konstanty ke všem hodnotám se tato variabilita nezmění.

Rozptyl je definován jako

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{10.7}$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2$$

$$= \frac{1}{n}\sum_{j=1}^{m} n_j(a_j - \bar{x})^2 \tag{10.8}$$

Někdy se ve jmenovateli místo n používá $n-1$. Pokud se vzorec (10.8) použije na třídní četnosti, doporučuje se *Sheppardova korekce*, totiž zmenšit výraz z (10.8) o hodnotu $h^2/12$, kde h je šířka stejně širokých intervalů.

Směrodatná odchylka s_x je definována jako odmocnina z výběrového rozptylu. Fyzikálně je tedy vyjádřena ve stejných jednotkách jako samotná měření. Stejně jako rozptyl a průměr záleží směrodatná odchylka na všech pozorováních.

Rozpětí je rovno rozdílu maximální a minimální hodnoty, tedy

$$R = x_{(n)} - x_{(1)}.$$

Na rozdíl od směrodatné odchylky či rozptylu závisí rozpětí pouze na velikosti nejmenší a největší hodnoty.

Kvartilové rozpětí je definováno jako rozdíl obou kvartilů

$$R_Q = Q_3 - Q_1,$$

polovina z této hodnoty se někdy nazývá **kvartilová odchylka**.

Průměrná odchylka je dána průměrnou vzdáleností jednotlivých hodnot x_i na reálné přímce od mediánu:

$$d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \tilde{x}|.$$

Někdy se v definici průměrné odchylky používá průměr místo mediánu.

Všechny uvedené míry variability vyžadují aspoň intervalové měřítko.

Dvě z uvedených měr polohy mají v souvislosti s mírami variability zajímavé extremální vlastnosti.

Věta 10.1. Funkce

$$S(t) = \frac{1}{n}\sum_{i=1}^{n}(x_i - t)^2$$

nabývá svého minima, právě když je $t = \bar{x}$.

D ů k a z: Snadno se ověří

$$\begin{aligned}\sum_{i=1}^{n}(x_i - t)^2 &= \sum_{i=1}^{n}\left((x_i - \bar{x}) + (\bar{x} - t)\right)^2 \\ &= \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - t)^2,\end{aligned}$$

odkud již dokazované tvrzení ihned plyne. □

Věta 10.2. Funkce

$$T(t) = \frac{1}{n}\sum_{i=1}^{n}|x_i - t|^2$$

nabývá svého minima pro $t = \tilde{x}$.

D ů k a z: Vyjdeme z uspořádaného souboru hodnot (10.1). Součet

$$\sum_{i=1}^{n}|x_i - t| = \sum_{i=1}^{n}|x_{(i)} - t| \tag{10.9}$$

můžeme vyjádřit jako součet $[(n+1)/2]$ sčítanců jako

$$\left(|x_{(1)} - t| + |x_{(n)} - t|\right) + \left(|x_{(2)} - t| + |x_{(n-1)} - t|\right) + \cdots. \tag{10.10}$$

Pokud zvolíme $t \in \langle x_{(1)}, x_{(n)}\rangle$, bude první sčítanec v (10.10) roven vzdálenosti těchto bodů na reálné přímce, jinak bude tento sčítanec větší. Minimalizace druhého sčítance podobně znamená zvolit t z intervalu $\langle x_{(2)}, x_{(n-1)}\rangle$, což není ve sporu s požadavkem minimalizovat první sčítanec, protože je tento interval částí intervalu $\langle x_{(1)}, x_{(n)}\rangle$. Takto lze pokračovat, dokud nedojdeme k poslednímu sčítanci. Je-li n sudé, můžeme udělat stejnou úvahu i pro poslední sčítanec, takže stejné minimální hodnoty nabude součet (10.9) pro každé $t \in \langle x_{(n/2)}, x_{(n/2+1)}\rangle$. Ovšem medián $\tilde{x}$ z definice (10.5) je právě ve středu tohoto intervalu. Je-li n liché, je poslední sčítanec v (10.10) roven právě hodnotě $|\tilde{x} - t|$, takže součet (10.9) je minimální právě pro $t = \tilde{x}$. □

Pro veličiny měřené v nominálním měřítku se nehodí žádná z dosud uvedených charakteristik. V tomto případě lze variabilitu výsledků charakterizovat pomocí **entropie** definované v případě m možných hodnot jako

$$H = -\sum_{j=1}^{m}\frac{n_j}{n}\log\frac{n_j}{n}, \tag{10.11}$$

kde $n_1, n_2, \ldots, n_m$ jsou zjištěné nenulové četnosti jednotlivých hodnot měřené veličiny.

10.3 Míry šikmosti a špičatosti

Čtenář si jistě uvědomil souvislost s momentovými charakteristikami náhodné veličiny. Pokud bychom zavedli náhodnou veličinu X tak, že bychom každé z hodnot x_i přidělili stejnou pravděpodobnost $1/n$ (každé z hodnot a_j pravděpodobnost n_j/n), pak by střední hodnota $\mathsf{E}X$ byla totožná s aritmetickým průměrem a rozptyl $\mathsf{var}X$ by byl totožný s rozptylem s_x^2 z (10.7) resp. (10.8). Podobně jako koeficient šikmosti v (7.19) zavedeme **výběrový koeficient šikmosti** pomocí

$$g_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns^3}.$$

V literatuře se tento koeficient značí někdy také jako $\sqrt{b_1}$, i když g_1 může být zřejmě i záporné.

Pomocí výběrových kvantilů se pro $0 < p < 0{,}5$ zavádí **kvantilový koeficient šikmosti** jako

$$\frac{(x_{1-p} - \tilde{x}) - (\tilde{x} - x_p)}{x_{1-p} - x_p}.$$

Speciálně pro $p = 0{,}25$ dostaneme **kvartilový koeficient šikmosti**

$$\frac{(Q_3 - \tilde{x}) - (\tilde{x} - Q_1)}{Q_3 - Q_1}.$$

Podobně **výběrový koeficient špičatosti** je definován vztahem

$$g_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4} - 3,$$

přičemž se někdy tento koeficient definuje jako $b_2 = g_2 + 3$, zejména v anglické literatuře.

Kvantilový koeficient špičatosti se pro $0 < p < 1/2$ definuje jako

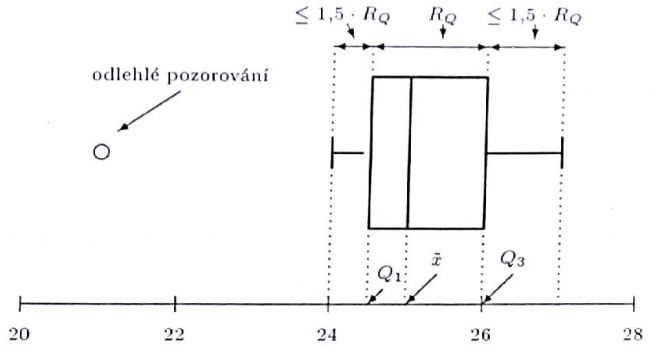
$$\frac{x_{(n)} - x_{(1)}}{x_{1-p} - x_p}.$$

10.4 Diagramy

Vedle klasického histogramu, který jsme již popsali, se používá řada dalších grafických pomůcek, které umožňují znázornit různé vlastnosti dat. Tyto

pomůcky jsou řazeny k **exploračním** statistickým metodám (také **EDA** – Exploratory Data Analysis).

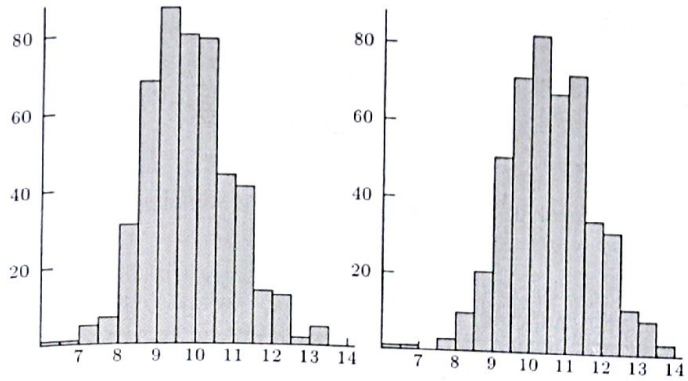
Asi nejznámější je **krabicový diagram** (box plot, box and whisker plot, vousatá krabička), který se vyskytuje v mnoha modifikacích. Na obrázku 10.2 jsou kromě mediánu znázorněny oba kvartily. Tykadla sahají k takovému nejvzdálenějšímu od odpovídajícího kvartilu pozorování, které není od něho dále než činí jeden a půl násobek kvartilového rozpětí (trojnásobek kvartilové odchylky). Jednotlivě jsou vyznačena pozorování, která jsou ve větší vzdálenosti (např. program STATGRAPHICS, někdy také STATISTICA). U některých programů sahají tykadla k nejmenšímu resp. největšímu pozorování (standardní postup programu STATISTICA), jindy k výběrovému 10% a 90% výběrovému kvantilu (SOLO).



Obrázek 10.2: Krabicový diagram

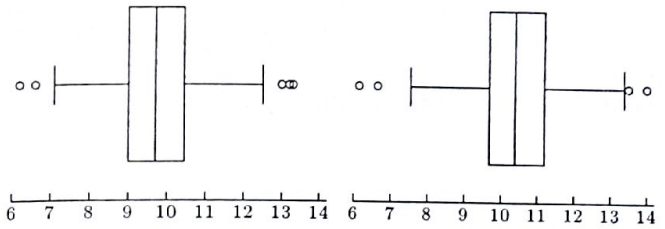
Příklad 10.2. Ukažme si krabicový diagram na jednoduchých datech, která umožní snadný výpočet jednotlivých statistik. Při zjišťování postojů matek ke kojení byl mimo jiné zjišťován věk matek v okamžiku porodu. Mezi asi dvěma stovkami matek zahrnutých do výzkumu jich bylo celkem n=12 takových, že své vzdělání ukončily maturitou, šlo o plánované těhotenství a první porod. Uspořádaný soubor hodnot obsahuje čísla: 21, 24, 24, 25, 25, 25, 25, 25, 26, 26, 27, 27. Snadno dostaneme $\tilde{x}$=25, Q_1=24,5, Q_3=26. Odtud je kvartilové rozpětí rovno $R_Q = 26 - 24{,}5 = 1{,}5$. Nejmenší pozorování je označeno jako odlehlé, protože je $21 < 24{,}5 - 1{,}5 \cdot 1{,}5 = 22{,}25$. Krabicový diagram je uveden na obrázku 10.2 ○

Příklad 10.3. Součástí rozsáhlého výzkumu bylo mimo jiné zjišťování



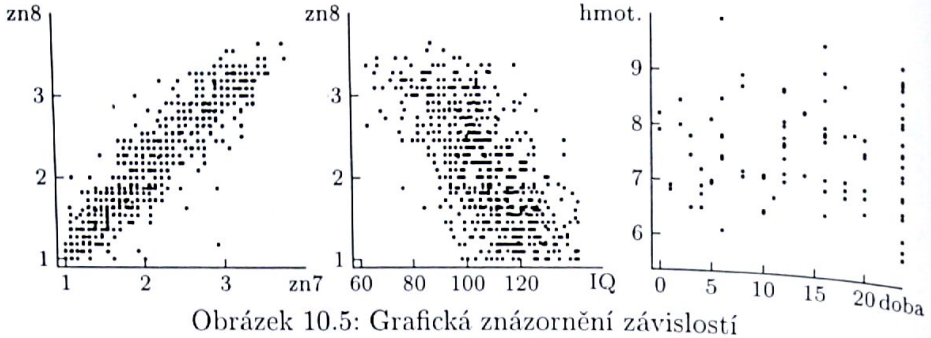
a) Hmotnosti dívek b) Hmotnosti chlapců

Obrázek 10.3: Histogram hmotnosti dvanáctiměsíčních dětí



a) Hmotnosti dívek b) Hmotnosti chlapců

Obrázek 10.4: Krabicový diagram hmotnosti dvanáctiměsíčních dětí

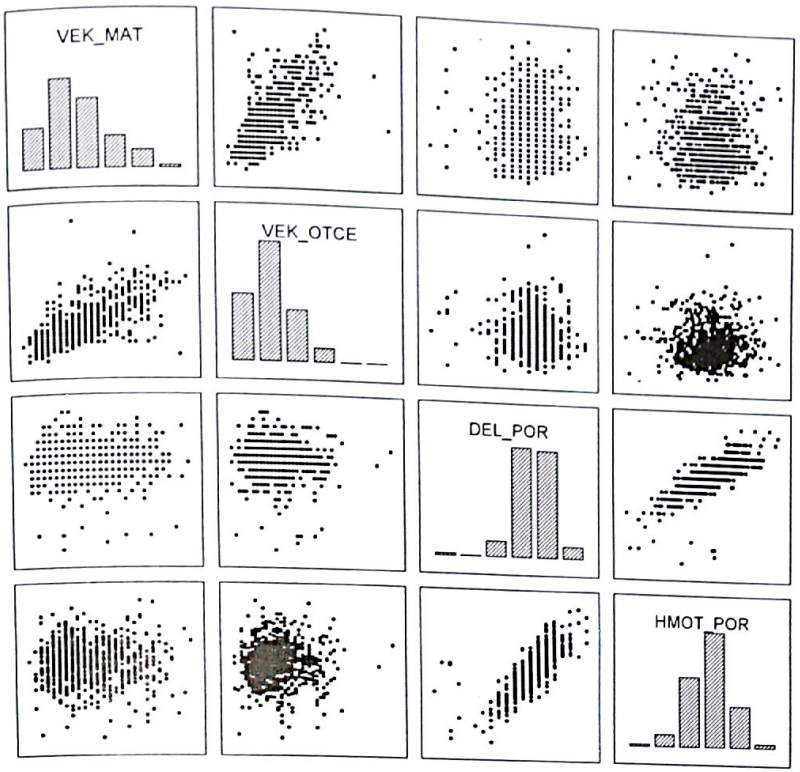


Obrázek 10.5: Grafická znázornění závislostí

hmotnosti dětí ve dvanáctém měsíci jejich věku. Na obrázku 10.3 jsou uvedeny histogramy hmotnosti děvčat a chlapců. Oba histogramy ukazují na kladnou šikmost (jsou poněkud protaženy na pravou stranu). Je z nich zřejmé, že hmotnosti chlapců jsou v průměru větší než hmotnosti dívek. Na obrázku 10.4 jsou uvedeny odpovídající krabicové diagramy. ○

Chceme-li vyjádřit závislost dvou spojitých veličin s hodnotami (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), použijeme **rozptylový diagram**, v němž znázorníme body $[x_i, y_i]$. Na obrázku 10.5 jsou znázorněny tři různé závislosti. Nejprve silná závislost průměrné známky v 8. třídě na průměrné známce v 7. třídě, kdy s rostoucí hodnotou jedné veličiny očekáváme rostoucí hodnoty druhé veličiny. Prostřední obrázek ukazuje poněkud slabší závislost stejných známek z 8. třídy na hodnotě IQ, odpovídající nepřímé úměrnosti. Třetí obrázek ukazuje prakticky nezávislé veličiny doba kojení (v týdnech) a hmotnost dítěte ve 24. týdnu.

Sledujeme-li současně chování několika veličin, je užitečný **maticový diagram**, znázorňující histogramy jednotlivých veličin a současně jejich rozptylové diagramy. Na obrázku 10.6 jsou takto znázorněny veličiny věk matky, věk otce, porodní délka a porodní hmotnost jejich novorozeného syna.



Obrázek 10.6: Grafická znázornění závislostí (maticový diagram)

11. Výběr

O statistice se hovoří jako o metodologické nauce, která objektivizuje proces poznávání. Zkusme popsat, jak toho dosahuje.

11.1 Výběr bez vracení z konečné populace

Uvažujme veliký statistický soubor, který budeme nazývat **populace** (též **základní soubor**) a na každé z jeho N jednotek principiálně můžeme změřit hodnotu zvoleného číselného znaku X, čímž bychom získali hodnoty $x_1, x_2, \ldots, x_N$. Průměr $\bar{x}$ podnot znaku na celé populaci spočítaný podle (10.3) označíme symbolem μ, populační rozptyl podle (10.7) označíme symbolem σ^2:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2.$$

Uvedený soubor je však tak veliký, že není možné nebo aspoň hospodárné zjistit hodnotu zkoumaného znaku u každé statistické jednotky. Proto vybereme skupinu n statistických jednotek a zjistíme hodnoty sledovaného znaku pouze pro tyto vybrané jednotky. Tento **výběr** (též **výběrový soubor**) musí být takový, aby dobře reprezentoval celou populaci (celý základní soubor). Všude dále budeme předpokládat $n < N$.

Výběrový soubor lze ze základního souboru vybrat celkem $\binom{N}{n}$ způsoby. Když budeme prvky výběrového souboru vybírat náhodně, pak požadované reprezentativnosti dosáhneme, když každá n-tice bude mít stejnou pravděpodobnost, že bude vybrána. Nejjednodušším způsobem, jak toho dosáhnout, je použít **náhodný výběr bez vracení** (prostý náhodný výběr), při němž nejprve náhodně vybereme jeden z N prvků základního souboru, potom opět náhodně jeden ze zbývajících $N-1$ prvků základního souboru atd. až náhodně vybereme n-tý prvek výběrového souboru ze zbývajících $N-n+1$ prvků základního souboru. Konkrétní uspořádanou n-tici indexů $(i_1, i_2, \ldots, i_n)$, vlastně elementární jev ω, dostaneme s pravděpodobností

$$\frac{1}{N}\frac{1}{N-1}\cdots\frac{1}{N-n+1} = \frac{(N-n)!}{N!}.$$

K tomuto elementárnímu ω jevu přiřadíme n-tici

$$(X_1(\omega), X_2(\omega), \ldots, X_n(\omega)) = (x_{i_1}, x_{i_2}, \ldots, x_{i_n}),$$

čímž jsme zavedli náhodný vektor $(X_1, X_2, \ldots, X_n)$.

V našich dalších úvahách velmi často budeme pracovat se symetrickými funkcemi složek vektoru $(X_1, X_2, \ldots, X_n)$, u nichž nezávisí na pořadí složek. Bude záležet pouze na tom, které prvky základního souboru se dostaly do výběru, takže bude stačit pracovat s n člennou podmnožinou prvků základního souboru. Každý takový výběr s je sjednocením celkem $n!$ elementárních jevů, má tedy pravděpodobnost $1/\binom{N}{n}$.

Nyní si náhodný výběr vyjádříme ještě jinak. K danému výběru s zavedeme (srovnej s příkladem 1.2) náhodné veličiny $W_1, W_2, \ldots, W_N$ (náhodný vektor $\boldsymbol{W}=(W_1, \ldots, W_N)^T$) předpisem

$$W_i = W_i(s) = \begin{cases} 1, & \text{pokud } i \in s, \\ 0, & \text{pokud } i \notin s. \end{cases}$$

Veličiny W_i se někdy nazývají indikátory zahrnutí, označují jedničkou ty prvky základního souboru, které se dostaly do výběru s.

Věta 11.1. Pro veličiny $W_1, \ldots, W_N$ platí

$$\mathsf{P}[W_i = 1] = \frac{n}{N}, \tag{11.1}$$

$$\mathsf{E}W_i = \frac{n}{N}, \tag{11.2}$$

$$\mathrm{var}W_i = \frac{n}{N}\left(1 - \frac{n}{N}\right), \tag{11.3}$$

$$\mathsf{E}W_iW_j = \frac{n(n-1)}{N(N-1)} \quad \text{pro } i \neq j, \tag{11.4}$$

$$\mathrm{cov}\,(W_i, W_j) = -\frac{n(N-n)}{N^2(N-1)} \quad \text{pro } i \neq j. \tag{11.5}$$

D ů k a z: Zřejmě je $W_i = 1$ právě když $i \in s$. Statistická jednotka i je prvkem celkem $\binom{N-1}{n-1}$ stejně pravděpodobných n-tic. Proto je pro každé pevné i

$$\mathsf{P}[W_i = 1] = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N},$$

což znamená, že jsme dokázali vztah (11.1). Pro každé pevné i má náhodná veličina W_i alternativní rozdělení s pravděpodobností jedničky podle (11.1). Odtud plynou vztahy (11.2) a (11.3). Zvolme nyní libovolně, ale pevně $i \neq j$. Podobně jako na začátku důkazu dostaneme (11.4):

$$\begin{aligned} \mathsf{E}W_iW_j &= 1 \cdot \mathsf{P}\,[W_i = 1, W_j = 1] + 0 \\ &= \frac{\binom{N-2}{n-2}}{\binom{N}{n}} \end{aligned}$$

$$= \frac{n(n-1)}{N(N-1)}.$$

Zbývající tvrzení (11.5) odtud dostaneme snadno:

$$\begin{aligned} \mathsf{cov}\,(W_i, W_j) &= \mathsf{E}W_iW_j - \mathsf{E}W_i\mathsf{E}W_j \\ &= -\frac{n(N-n)}{N^2(N-1)}. \end{aligned}$$ □

I když jsme v důkazu zjistili, že pro každé i má náhodná veličina W_i alternativní rozdělení, přesto odtud neplyne, že by součet $\sum_{i=1}^{N} W_i$ měl binomické rozdělení. Naopak, tento součet je *vždy* roven nenáhodné hodnotě n. V čem je rozdíl v porovnání s příkladem 7.6? (Jsou náhodné veličiny $W_1, W_2, \ldots, W_N$ nezávislé? Všimněte si kovariance $\mathsf{cov}(W_i, W_j)$ v (11.5).)

Jak můžeme pomocí dříve zavedených náhodných veličin $X_1, X_2, \ldots, X_n$ odhadnout populační průměr μ? Zkusme použít obdobu populačního průměru založenou pouze na dostupných hodnotách, tedy na výběru.

Populační průměr μ odhadneme pomocí průměru spočítaného z hodnot sledovaného znaku na prvcích výběrového souboru

$$\bar{X} = \bar{X}(s) = \frac{1}{n}\sum_{j=1}^{n} X_j. \tag{11.6}$$

Velké písmeno na levé straně (11.6) vyjadřuje skutečnost, že jde o náhodnou veličinu, totiž o funkci výběru (a tedy elementárních jevů ω). **Výběrový průměr** $\bar{X}$ můžeme vyjádřit také jako

$$\begin{aligned} \bar{X} &= \frac{1}{n}\sum_{i\in s} x_i \\ &= \frac{1}{n}\sum_{i=1}^{N} x_i W_i. \end{aligned} \tag{11.7}$$

Nyní využijeme tvrzení věty 11.1 a dokážeme základní statistické vlastnosti výběrového průměru $\bar{X}$.

Věta 11.2. Pro výběrový průměr spočítaný z prvků náhodného výběru z konečné polulace platí

$$\mathsf{E}\bar{X} = \mu, \tag{11.8}$$

$$\mathsf{var}\bar{X} = \frac{N-n}{N-1}\frac{\sigma^2}{n}. \tag{11.9}$$

D ů k a z: Střední hodnotu spočítáme s využitím (11.7) jako

$$\begin{aligned} \mathsf{E}\bar{X} &= \mathsf{E}\frac{1}{n}\sum_{i=1}^{N} x_i W_i \\ &= \frac{1}{n}\sum_{i=1}^{N} x_i \frac{n}{N} \\ &= \frac{1}{N}\sum_{i=1}^{N} x_i \\ &= \mu \end{aligned}$$

Využijeme-li skutečnosti, že je $\sum_{i=1}^{N} W_i = n$, postupně dostaneme také

$$\begin{aligned} \mathsf{var}(\bar{X}) &= \mathsf{E}\left(\bar{X}-\mu\right)^2 \\ &= \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{N}(x_i-\mu)W_i\right)^2 \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{N}(x_i-\mu)^2\mathsf{E}W_i^2 + \sum\sum_{i\neq j}(x_i-\mu)(x_j-\mu)\mathsf{E}W_iW_j\right) \\ &= \frac{1}{n^2}\left(\frac{n}{N}\sum_{i=1}^{N}(x_i-\mu)^2 + \frac{n(n-1)}{N(N-1)}\sum\sum_{i\neq j}(x_i-\mu)(x_j-\mu)\right) \\ &= \frac{1}{n^2}\left(\frac{n}{N}\sum_{i=1}^{N}(x_i-\mu)^2 + \frac{n(n-1)}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i-\mu)(x_j-\mu)\right. \\ &\qquad \left. - \frac{n(n-1)}{N(N-1)}\sum_{i=1}^{N}(x_i-\mu)^2\right) \\ &= \frac{1}{n^2}\left(\left(\frac{n}{N}-\frac{n(n-1)}{N(N-1)}\right)\sum_{i=1}^{N}(x_i-\mu)^2\right. \\ &\qquad \left. + \frac{n(n-1)}{N(N-1)}\left(\sum_{i=1}^{N}(x_i-\mu)\right)^2\right) \\ &= \frac{1}{n^2}\left(\frac{n}{N}-\frac{n(n-1)}{N(N-1)}\right)N\sigma^2 + 0 \quad = \quad \frac{N-n}{N-1}\frac{\sigma^2}{n}. \end{aligned}$$ □

V případě náhodného výběru z konečné populace je střední hodnota výběrového průměru $\bar{X}$ rovna populačnímu průměru μ, tedy odhadovanému parametru. Uvedenou vlastnost stručně formulujeme do prohlášení, že $\bar{X}$ je **nestranným odhadem** parametru μ.

Výrazu $\frac{N-n}{N-1}$ se říká **konečnostní násobitel** . Pro $n << N$ je vliv konečnostního násobitele zanedbatelný.

Zvolíme-li $n = 1$, pak poslední věta udává vlastnosti jediného pozorování, provedeného na jediné náhodně vybrané statistické jednotce. Označíme-li takto získanou hodnotu jako X, pak zřejmě platí

$$\mathsf{E}X = \mu, \qquad \mathsf{var}X = \sigma^2.$$

Podobně pomocí rozptylu spočítaného z výběrového souboru zkusme odhadnout rozptyl. Jako odhad rozptylu použijme **výběrový rozptyl** daný vztahem

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2 \tag{11.10}$$

Věta 11.3. Pro výběrový rozptyl spočítaný z náhodného výběru z konečné populace platí

$$\mathsf{E}S^2 = \frac{N-1}{N}\sigma^2. \tag{11.11}$$

D ů k a z: Jednoduchou úpravou lze dokázat

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 + n\left(\bar{X} - \mu\right)^2 \tag{11.12}$$

Odtud je

$$\begin{aligned}
\mathsf{E}S^2 &= \frac{1}{n-1}\mathsf{E}\sum_{i\in s}(x_i - \mu)^2 - \frac{n}{n-1}\mathsf{E}(\bar{X} - \mu)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{N}(x_i - \mu)^2\mathsf{E}W_i - \frac{n}{n-1}\mathsf{var}\bar{X} \\
&= \frac{1}{n-1}N\sigma^2\frac{n}{N} - \frac{n}{n-1}\frac{N-n}{N-1}\frac{\sigma^2}{n} \\
&= \frac{N}{N-1}\sigma^2. \qquad \square
\end{aligned}$$

Odhad S^2 tedy není nestranným odhadem populačního rozptylu σ^2, jako nestranný odhad bychom mohli použít například výraz $\frac{N-1}{N}S^2$.

11.2 Náhodný výběr

Nyní provede podobný **výběr** z konečné populace, ale **s vracením**. Jakmile vybereme z populace nějaký prvek, zjistíme hodnotu měřeného znaku, prvek do populace opět vrátíme a teprve pak znovu vybíráme. Vybírání provádíme tak, aby každý prvek populace měl (v každém kroku) stejnou pravděpodobnost být vybrán, totiž $1/N$. Označme hodnoty naměřené na takto náhodně vybraných prvcích jako $X_1, \ldots, X_n$. Jsou to zřejmě *nezávislé* náhodné veličiny, pokaždé vybírané z téže populace. Protože navíc každý krok lze považovat za speciální případ výběru rozsahu 1 s vracením, znamená to, že podle (11.8) a (11.9) platí

$$\mathsf{E}X_j = \mu, \quad \mathsf{var}X_j = \sigma^2, \qquad j = 1, \ldots, n.$$

Pokusme se o nový pohled na výběr s vracením. Není třeba se zabývat hodnotami sledovaného statistického znaku u jednotlivých statistických jednotek, ale pouze podílem jednotek, u nichž sledovaný znak má hodnotu menší než je zvolená konstanta x. Uvedený podíl označme symbolem $p(x)$. Vyberme *náhodně* jednu statistickou jednotku a označme hodnotu sledovaného znaku symbolem X. Pravděpodobnost toho, že hodnota náhodné veličiny X je menší než zvolené x, je rovna podílu $p(x)$ statistických jednotek s hodnotou znaku menší než x. Zjištěná hodnota je tedy náhodnou veličinou s distribuční funkcí $F_X(x) \equiv p(x)$. Budeme předpokládat, že u rozdělení daného distribuční funkcí $F_X(x)$ existuje konečný rozptyl σ^2, jeho střední hodnotu označíme μ. Všimněme si, že nás už nezajímá, zda je výchozí populace konečná či nekonečná.

Mnohdy si pod základním souborem představujeme tak veliký soubor, že jej lze považovat za nekonečný. Takový soubor nemusí existovat aktuálně, ale jen potenciálně. I když v tomto souboru provedeme výběr bez vracení, jak jsme jej popsali v předchozím oddílu, bude možné zjištěné hodnoty přesto považovat za nezávislé. Intuitivně to plyne z toho, že při nekonečně velkém základním souboru odebrání několika statistických jednotek prakticky nezmění jeho složení, tedy funkci $p(x)$.

V obou případech, tj. u náhodného výběru bez vracení z nekonečně velké populace i u náhodného výběru s vracením z konečné či nekonečné populace je výsledkem pokusu n-tice *nezávislých* náhodných veličin $X_1, X_2, \ldots, X_n$ se *stejným rozdělením* daným distribuční funkcí $F_X(x)$. Taková n tice náhodných veličin se nazývá **náhodný výběr rozsahu** n. Někdy se hovoří o n nezávislých *kopiích* náhodné veličiny X. Připomeňme označení

$$\mathsf{E}X = \mu, \qquad \mathsf{var}X = \sigma^2.$$

Stejně jako v předchozím oddílu vyšetříme vlastnosti **výběrového průměru**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{11.13}$$

Věta 11.4. Pro výběrový průměr spočítaný z náhodného výběru rozsahu n z rozdělení s konečnou střední hodnotou a konečným rozptylem platí

$$\mathsf{E}\bar{X} = \mu \tag{11.14}$$

$$\mathsf{var}\bar{X} = \frac{\sigma^2}{n} \tag{11.15}$$

D ů k a z: K důkazu (11.14) použijeme známou vlastnost střední hodnoty součtu náhodných veličin (6.9):

$$\mathsf{E}\bar{X} = \mathsf{E}\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}X_i = \mu.$$

Při důkazu druhého tvrzení musíme použít předpokládanou nezávislost náhodných veličin $X_1, X_2, \ldots, X_n$, což s použitím (7.2) a (7.9) vede k výpočtu

$$\mathsf{var}\bar{X} = \mathsf{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathsf{var}X_i = \frac{\sigma^2}{n}. \qquad \square$$

Porovnejme vlastnosti výběrového průměru v případech konečného a nekonečného základního souboru. V obou případech je nestranným odhadem střední hodnoty μ. Rozptyl výběrového průměru je pro $n > 1$ v případě konečného základního souboru rozsahu menší než v případě stejně velkého výběru z nekonečného základního souboru zásluhou konečnostního násobitele $\frac{N-n}{N-1}$. Pokud je konečný základní soubor podstatně větší než výběrový soubor, zpravidla jej považujeme za soubor nekonečný.

Jako odhad rozptylu σ^2 se používá výběrový rozptyl

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{N}(X_i - \bar{X})^2. \tag{11.16}$$

Věta 11.5. Pro náhodný výběr rozsahu n z rozdělení s konečným rozptylem σ^2 platí

$$\mathsf{E}S^2 = \sigma^2. \tag{11.17}$$

D ů k a z: Důkaz je analogický důkazu věty 11.3. Použijeme-li identitu (11.12), můžeme psát

$$\begin{aligned}
\mathsf{E}S^2 &= \mathsf{E}\frac{1}{n-1}\sum_{j=1}^{n}(X_i-\mu)^2 - \mathsf{E}\frac{n}{n-1}(\bar{X}-\mu)^2 \\
&= \frac{1}{n-1}\sum_{j=1}^{n}\mathsf{var}X_j - \mathsf{E}\frac{n}{n-1}\mathsf{var}\bar{X} \\
&= \frac{n}{n-1}\sigma^2 - \frac{1}{n-1}\sigma^2 \\
&= \sigma^2. \qquad \square
\end{aligned}$$

11.3 Náhodný výběr z normálního rozdělení

Uvažujme nyní speciální případ, kdy máme náhodný výběr z normálního rozdělení $\mathsf{N}(\mu, \sigma^2)$.

Věta 11.6. Pro náhodný výběr z normálního rozdělení $\mathsf{N}(\mu, \sigma^2)$ platí

$$\bar{X} \sim \mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right). \tag{11.18}$$

D ů k a z: Zaveďme

$$Z_i = \frac{X_i - \mu}{\sigma}, \qquad i = 1, \ldots, n,$$

což jsou zřejmě nezávislé náhodné veličiny s normovaným normálním rozdělením. Označíme-li $\mathbf{a} = \mu\mathbf{1}$, $\mathbf{B} = \sigma\mathbf{I}_n$ a $\boldsymbol{X} = (X_1, \ldots, X_n)^T$, můžeme psát $\boldsymbol{X} = \mathbf{a} + \mathbf{B}Z$, takže náhodný vektor $\boldsymbol{X}$ má zřejmě (viz definice 8.1) mnohorozměrné normální rozdělení. Zvolíme-li $\mathbf{d} = (1/n)\mathbf{1}$, z věty 8.4 plyne, že náhodná veličina $\bar{X} = \mathbf{d}^T\boldsymbol{X}$ má (jednorozměrné) normální rozdělení se střední hodnotou $\mathbf{d}^T\mathbf{a} = \mu$ a rozptylem

$$\mathbf{d}^T\sigma^2\mathbf{I}\mathbf{d} = \frac{\sigma^2}{n^2}\mathbf{1}^T\mathbf{1} = \frac{\sigma^2}{n}. \qquad \square$$

Všimněme si ještě odhadu rozptylu podle (11.16).

Věta 11.7. Je-li $X_1, \ldots, X_n$ náhodný výběr z rozdělení $\mathsf{N}(\mu, \sigma^2)$, potom $\bar{X}$ a S^2 jsou nezávislé náhodné veličiny a platí

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2(n-1).$$

D ů k a z: Náhodné veličiny $Z_1, \ldots, Z_n$, zavedené v (11.3) jsou nezávislé a mají normované normální rozdělení. Podle věty 8.3 mají stejnou vlastnost také složky $U_1, \ldots, U_n$ náhodného vektoru definovaného vztahem $\boldsymbol{U} = \mathbf{Q}^T\boldsymbol{Z}$, kde jako $\mathbf{Q}$ zvolíme matici, jejíž první sloupec je dán předpisem

$$\mathbf{q}_1 = \frac{1}{\sqrt{n}}\mathbf{1} \tag{11.19}$$

Platí ovšem

$$\begin{aligned} U_1 &= \sqrt{n}\bar{Z} \\ &= \sqrt{n}\frac{\bar{X}-\mu}{\sigma}. \end{aligned}$$

Kromě toho platí (viz (11.12))

$$\sum_{i=1}^{n} U_i^2 = \sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n}(Z_i - \bar{Z})^2 + n\bar{Z}^2.$$

Uvážíme-li, že je $U_1^2 = n\bar{Z}^2$, platí zřejmě

$$\begin{aligned} \sum_{i=2}^{n} U_i^2 &= \sum_{i=1}^{n}(Z_i - \bar{Z})^2 \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2, \end{aligned}$$

takže náhodná veličina

$$\frac{n-1}{\sigma^2}S^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{11.20}$$

má rozdělení $\chi^2(n-1)$ a je nezávislá s náhodnou veličinou $U_1 = (\bar{X} - \mu)/\sigma$. □

Poznámka. Když si připomeneme definici Studentova t rozdělení z příkladu 8.8, snadno nahlédneme, že náhodná veličina

$$T = \frac{\bar{X} - \mu}{S}\sqrt{n} \tag{11.21}$$

má rozdělení $\mathrm{t}(n-1)$.

11.4 Cvičení

11.1. Uvažujte náhodný výběr z rozdělení s konečným rozptylem σ^2. Určete konstantu c tak, aby statistika $c\sum_{i=2}^{n}(X_i - X_{i-1})^2$ byla nestranným odhadem rozptylu σ^2.

11.2. Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozdělení $\mathsf{N}(\mu, \sigma^2)$. Dokažte, že platí $\mathrm{var}S^2 = 2\sigma^2/(n-1)$. Návod: Jaký rozptyl má rozdělení $\chi^2(n-1)$?

12. Základy statistické indukce

Tato kapitola je věnována základním principům statistického uvažování. Prvně zde narazíme na problém s rozlišením náhodné veličiny a její hodnoty. Budeme se snažit dodržovat dosud užívaný způsob, který například znamená, že výběrový průměr spočítaný z náhodného výběru $X_1, \ldots, X_n$ označíme symbolem $\bar{X}$. Když se ale náhodný výběr realizuje konkrétními hodnotami $x_1, \ldots, x_n$ (např. 27, 15, ...), pak dostaneme konkrétní realizaci $\bar{x}$ náhodné veličiny $\bar{X}$ (např. $\bar{x} = 18{,}25$). Ne vždy se nám podaří tuto konvenci dodržet, například při označování odhadů získaných metodou maximální věrohodnosti použijeme stříšku nad příslušný symbol řecké abecedy.

12.1 Výběr z normálního rozdělení se známou střední hodnotou

Předpokládejme, že náhodné veličiny $X_1, X_2, \ldots, X_n$ tvoří náhodný výběr rozsahu n z normálního rozdělení $\mathsf{N}(\mu, \sigma^2)$. Předpokládejme navíc, že *známe* rozptyl σ^2. Jakou informaci o hodnotě μ můžeme získat z tohoto náhodného výběru?

Podle věty 11.6 víme, že v tomto případě výběrový průměr má normální rozdělení $\mathsf{N}(\mu, \sigma^2/n)$. Je tedy nestranným odhadem parametru μ. Provedeme-li normování této náhodné veličiny (viz (7.4)), dostaneme náhodnou veličinu s normovaným normálním rozdělením, takže pro libovolné $\alpha \in (0,1)$ bude platit

$$1-\alpha = \mathsf{P}[|Z| < z(\alpha/2)] \tag{12.1}$$

$$= \mathsf{P}\left[\left|\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}\right| < z(\alpha/2)\right] \tag{12.2}$$

$$= \mathsf{P}\left[\bar{X} - \frac{\sigma}{\sqrt{n}} z(\alpha/2) < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} z(\alpha/2)\right]. \tag{12.3}$$

Našli jsme interval s náhodnými konci, který s *předem danou* pravděpodobností pokrývá neznámý parametr μ. Říkáme, že

$$\left(\bar{X} - \frac{\sigma}{\sqrt{n}} z(\alpha/2), \bar{X} + \frac{\sigma}{\sqrt{n}} z(\alpha/2)\right) \tag{12.4}$$

je **interval spolehlivosti** pro parametr μ s **koeficientem spolehlivosti** $1-\alpha$. Zpravidla se pracuje s 95% nebo s 99% intervalem spolehlivosti. Výraz $\sigma/\sqrt{n}$, který je směrodatnou odchylkou výběrového průměru $\bar{X}$, se v této souvislosti nazývá **standardní (střední) chyba** průměru.

Příklad 12.1. Zabývejme se výškou desetiletých chlapců. V roce 1961 byla u 15 náhodně vybraných chlapců z populace všech desetiletých chlapců žijících v Československu zjištěna výška (tabulka 12.1). Základní soubor

130	140	136	141	139	133	149	151
139	136	138	142	127	139	147	

Tabulka 12.1: Výšky 15 desetiletých chlapců

všech desetiletých chlapců je sice konečný, ale v porovnání s pořízeným výběrem je tak veliký, že zjištěné výšky můžeme považovat za nezávislé náhodné veličiny. Ptáme se, zda se v porovnání s rokem 1951 změnila střední výška všech desetiletých chlapců, když víme, že v roce 1951 byla tato střední výška (zjištěná pomocí velmi rozsáhlého výběru) rovna hodnotě $\mu_0 = 136{,}1$ cm, směrodatná odchylka byla $\sigma = 6{,}4$ cm. Je známo, že variabilita výšek postavy se v různých generacích příliš nemění, kdežto samotný populační průměr se možná mění. Proto má smysl ptát se, zda za 10 roků ke změně populačního průměru došlo a předpokládat, že rozptyl σ^2 je známý.

Zvolme $1-\alpha = 95\ \%$. Z dat uvedených v tabulce 12.1 zjistíme, že je $n = 15$, $\bar{x} = 139{,}133$ cm, takže 95% interval spolehlivosti je

$$\left(139{,}133 - (6{,}4/\sqrt{15})1{,}96, 139{,}133 + (6{,}4/\sqrt{15})1{,}96\right) = (135{,}9, 142{,}4).$$

Populační průměr z roku 1951 je tímto intervalem pokryt, takže nemáme dost vážný důvod tvrdit, že se střední výška desetiletých hochů za uvažované desetiletí změnila. K této úloze se ještě vrátíme v příkladu 12.5. ○

12.2 Odhad parametrů metodou maximální věrohodnosti

V předchozím oddílu jsme naznačili způsob, jakým statistik může odhadovat neznámý parametr. Výběrový průměr $\bar{X}$ a interval (12.4) jsou bodovým a intervalovým odhadem parametru μ. Zdaleka ne pokaždé odhadujeme střední hodnotu. Uveďme proto obecnější postup.

Mějme náhodný výběr rozsahu n (posloupnost n nezávislých stejně rozdělených náhodných veličin) $X_1, X_2, \ldots, X_n$ z rozdělení, které závisí na (obecně vektorovém) parametru $\boldsymbol{\theta}$. Najít **bodový odhad** znamená najít takovou funkci náhodných veličin $X_1, X_2, \ldots, X_n$ (též statistiku, výběrovou charakteristiku), která je v nějakém smyslu blízko skutečné hodnotě $\boldsymbol{\theta}$. Označme tento odhad jako $\boldsymbol{T} = \boldsymbol{T}(X_1, X_2, \ldots, X_n)$. Protože je $\boldsymbol{T}$ funkce náhodných

veličin, je také náhodnou veličinou. Konstanta (vektor konstant) $\mathbf{b} = \mathsf{E}\boldsymbol{T} - \boldsymbol{\theta}$ se nazývá **vychýlení**. Řekneme, že odhad $\boldsymbol{T}$ parametru $\boldsymbol{\theta}$ je **nestranným** (nevychýleným) **odhadem**, jestliže platí $\mathsf{E}\boldsymbol{T} = \boldsymbol{\theta}$ pro každé $\boldsymbol{\theta}$. Nestrannost znamená, že vychýlení je nulové. Požadavek nestrannosti tedy znamená, že „v průměru" odhaduje přímo parametr $\boldsymbol{\theta}$, přičemž výraz „v průměru" se vztahuje k opakování výběru.

Lze si představit, že máme k disposici více nestranných odhadů nějakého parametru. Například k odhadu střední hodnoty lze místo výběrového průměru $\bar{X}$ použít pouze pozorování X_1. Takovéto konkurenční odhady lze porovnat podle velikosti kolísání kolem odhadované hodnoty. Lepší bude takový odhad, který kolísá méně.

Řekneme, že odhad $\boldsymbol{T}_0 \in \mathcal{T}$ je **nejlepší odhad** parametru $\boldsymbol{\theta}$ ve třídě odhadů $\mathcal{T}$, jestliže má mezi všemi odhady z této třídy nejmenší rozptyl, tj. jestliže platí implikace

$$\boldsymbol{T} \in \mathcal{T} \Rightarrow \mathsf{var}\boldsymbol{T}_0 \leq \mathsf{var}\boldsymbol{T}. \tag{12.5}$$

(O nerovnosti mezi maticemi viz A3.)

Někdy vyšetřujeme asymptotické chování odhadu v závislosti na rozsahu výběru n. Říkáme, že odhad $\boldsymbol{T} = \boldsymbol{T}_n$ je **konzistentním odhadem** parametru $\boldsymbol{\theta}$, jestliže pro každé $\varepsilon > 0$ platí

$$\lim_{n\to\infty} \mathsf{P}\left[\|\boldsymbol{T}_n - \boldsymbol{\theta}\| < \varepsilon\right] = 1, \tag{12.6}$$

tedy, jestliže posloupnost $\boldsymbol{T}_n$ konverguje k $\boldsymbol{\theta}$ podle pravděpodobnosti. V jednorozměrném případě k tomu například stačí, aby posloupnost středních hodnot odhadů konvergovala k odhadovanému parametru a posloupnost jejich rozptylů konvergovala k nule, jak uvádí následující věta.

Věta 12.1. Platí-li

$$\begin{aligned} \lim_{n\to\infty} \mathsf{E}T_n &= \theta, \\ \lim_{n\to\infty} \mathsf{var}T_n &= 0, \end{aligned}$$

pak je T_n konzistentním odhadem θ.

D ů k a z: Zvolme libovolně $\varepsilon > 0$. Podle Čebyševovy nerovnosti (věta 9.1) platí pro každé n

$$\mathsf{P}\left[|T_n - \mathsf{E}T_n| < \varepsilon/2\right] \geq 1 - \frac{\mathsf{var}T_n}{(\varepsilon/2)^2}.$$

Současně je pro $n > n_0$ vždy $|\mathsf{E}T_n - \theta| < \varepsilon/2$. Pro $n > n_0$ však platí

$$\begin{aligned} \mathsf{P}\left[|T_n - \theta| < \varepsilon\right] &\geq \mathsf{P}\left[|T_n - \mathsf{E}T_n| < \varepsilon/2, |\mathsf{E}T_n - \theta| < \varepsilon/2\right] \\ &= \mathsf{P}\left[|T_n - \mathsf{E}T_n| < \varepsilon/2\right], \end{aligned}$$

neboť druhý z jevů na pravé straně je jevem jistým. Poslední pravděpodobnost však podle Čebyševovy nerovnosti konverguje k 1. □

Příklad 12.2. Uvažujme náhodný výběr $X_1, \ldots, X_n$ z normálního rozdělení se střední hodnotou μ a rozptylem σ^2. Jako odhad rozptylu σ^2 se někdy používá statistika

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Použijeme-li větu 11.5, zjistíme, že tento odhad není nestranným odhadem rozptylu σ^2:

$$\mathsf{E}\hat{\sigma}^2 = \mathsf{E}\frac{n-1}{n}S^2 = \frac{n-1}{n}\sigma^2.$$

Zřejmě však platí $\lim_{n\to\infty} \hat{\sigma}^2 = \sigma^2$. Použijeme-li výsledku cvičení 11.2, dostaneme podobně

$$\lim_{n\to\infty} \mathsf{var}S^2 = \lim_{n\to\infty} \frac{2\sigma^2}{n-1} = 0,$$

takže podle věty 12.1 je vyšetřovaná statistika konzistentním odhadem rozptylu σ^2. ○

Zaveďme obecně také intervalový odhad. Platí-li pro statistiky $\boldsymbol{T}_L = \boldsymbol{T}_L(X_1, X_2, \ldots, X_n)$, $\boldsymbol{T}_U = \boldsymbol{T}_U(X_1, X_2, \ldots, X_n)$ vztah

$$\mathsf{P}\left[\boldsymbol{T}_L \leq \boldsymbol{\theta} \leq \boldsymbol{T}_U\right] = 1 - \alpha,$$

říkáme, že $(\boldsymbol{T}_L, \boldsymbol{T}_U)$ tvoří **interval spolehlivosti (intervalový odhad)** pro parametr $\boldsymbol{\theta}$ s koeficientem spolehlivosti $1 - \alpha$.

Zbývá uvést dostatečně obecnou metodu k nalezení odhadu. Předpokládejme, že náhodný výběr byl pořízen z rozdělení, které je charakterizováno hustotou $f(x; \boldsymbol{\theta})$. Sdružená hustota náhodného vektoru $(X_1, X_2, \ldots, X_n)^T$ je vzhledem k předpokládané nezávislosti dána předpisem

$$f(x_1, x_2, \ldots, x_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}). \tag{12.7}$$

Když vyšetřujeme sdruženou hustotu $f(\mathbf{x};\boldsymbol{\theta})$ jako funkci $\boldsymbol{\theta}$, nazýváme ji **věrohodnostní funkcí**. Zpravidla je technicky výhodnější pracovat s logaritmem věrohodnostní funkce, tedy s **logaritmickou věrohodnostní funkcí**

$$l(\boldsymbol{\theta}) = \ln f(\mathbf{x};\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(x_i;\boldsymbol{\theta}),$$

přičemž tam, kde je $f(x) = 0$, definujeme $l(\boldsymbol{\theta}) = -\infty$. Podobně v případě diskrétního rozdělení charakterizovaného pravděpodobnostmi $\mathsf{P}[X = x_i;\boldsymbol{\theta}]$, $i = 1, 2, \ldots$, definujeme logaritmickou věrohodnostní funkci jako

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \mathsf{P}[X_i = x_i;\boldsymbol{\theta}]. \tag{12.8}$$

Odhadem parametru $\boldsymbol{\theta}$ metodou **maximální věrohodnosti** je taková hodnota parametru $\widehat{\boldsymbol{\theta}} \in \Theta$, která maximalizuje na množině možných hodnot Θ parametru $\boldsymbol{\theta}$ věrohodnostní funkci (resp. logaritmickou věrohodnostní funkci):

$$\boldsymbol{\theta} \in \Theta \Rightarrow l(\widehat{\boldsymbol{\theta}}) \geq l(\boldsymbol{\theta}).$$

V případě diskrétního rozdělení lze říci, že odhad metodou maximální věrohodnosti vybere takovou hodnotu parametru, při které je pravděpodobnost, že nastane právě ta hodnota náhodného výběru, která se skutečně realizovala, největší možná.

Příklad 12.3. Mějme náhodný výběr $Y_1, Y_2, \ldots, Y_n$ z alternativního rozdělení $\mathsf{bi}(1, p)$, kde $p \in (0,1)$ je neznámý parametr. Sdružené rozdělení náhodného vektoru $(Y_1, Y_2, \ldots, Y_n)^T$ je dáno výrazem

$$\prod_{i=1}^{n} \mathsf{P}[Y_i = y_i] = p^y (1-p)^{n-y},$$

kde $y = \sum_{i=1}^{n} y_i$, takže logaritmickou věrohodnostní funkci můžeme zapsat jako

$$l(p) = y \ln p + (n-y)\ln(1-p).$$

Při hledání maxima řešíme rovnici

$$\frac{\partial l(p)}{\partial p} = \frac{y}{p} - \frac{n-y}{1-p} = 0,$$

která vede k odhadu (přesněji k odhadní funkci)

$$\widehat{p} = \frac{1}{n} Y,$$

kde je $Y = \sum_{i=1}^{n} Y_i$. Snadno se ověří, že jde skutečně o maximum věrohodnostní funkce. ○

Příklad 12.4. Mějme náhodný výběr rozsahu n z normálního rozdělení $\mathsf{N}(\mu, \sigma^2)$, v němž odhadujeme oba parametry. Logaritmická věrohodnostní funkce má pro $\mu \in \mathbb{R}, \sigma^2 > 0$ tvar

$$l(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Derivováním dostaneme

$$\frac{\partial l}{\partial \mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(-2)(x_i - \mu) = 0 \quad \Rightarrow \quad \widehat{\mu} = \bar{X},$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 \quad \Rightarrow \quad \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

(Pozor, odhadujeme veličinu σ^2, tedy derivujeme podle veličiny označené σ^2, nikoliv podle σ.) Dosadíme-li do matice druhých derivací odhady $\widehat{\mu}, \widehat{\sigma^2}$), dostaneme negativně definitní matici

$$\begin{pmatrix} -\frac{n}{\widehat{\sigma^2}} & 0 \\ 0 & -\frac{n}{2\left(\widehat{\sigma^2}\right)^2} \end{pmatrix},$$

což znamená, že nalezená řešení soustavy rovnic jsou opravdu maximálně věrohodnými odhady. Připomínám, že v příkladu 12.2 jsme zjistili, že $\hat{\sigma}^2$ není nestranným odhadem, pouze odhadem konzistentním. ○

12.3 Testování hypotéz

Vyjdeme z náhodného vektoru $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ se sdruženou distribuční funkcí $F_{\boldsymbol{X}}(\mathbf{x})$. Hypotézou rozumíme nějaké tvrzení o rozdělení určeném touto distribuční funkcí. Ve statistice formulujeme **nulovou hypotézu** H_0 a **alternativní hypotézu** H_A, která je zpravidla negací hypotézy nulové. Možnými rozhodnutími jsou nulovou hypotézu **zamítnout** nebo **nezamítnout**. Rozhodujeme na základě realizace náhodného vektoru, takže nemůžeme zaručit bezchybné rozhodnutí. Když hypotézu zamítneme, přestože platí, nastává **chyba prvního druhu**. Když hypotézu nezamítneme

rozhodnutí	skutečnost	
	H_0 platí	H_0 neplatí
zamítnout H_0	chyba prvního druhu	-
nezamítnout H_0	-	chyba druhého druhu

Tabulka 12.2: Možné situace při statistickém rozhodování

v situaci, kdy neplatí, nastává **chyba druhého druhu**. Čtyři možné situace znázorňuje tabulka 12.3.

Statistické rozhodování probíhá tak, že předem určíme **kritický obor** W, což je množina výsledků pokusu, při kterých budeme hypotézu zamítat. *Tvar* kritického oboru stanovíme tak, aby náhodný vektor $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ do kritického oboru padl za platnosti alternativní hypotézy co možná nejčastěji, kdežto za platnosti nulové hypotézy H_0 zřídka. *Velikost* kritického oboru zvolíme tak, abychom platnou hypotézu zamítali nejvýše s pravděpodobností α. Této maximální dovolené pravděpodobnosti chyby prvního druhu se říká **hladina testu**. Zpravidla se volí $\alpha = 0{,}05$ nebo $\alpha = 0{,}01$. S rozvojem výpočetní techniky se stále častěji používá postup, při kterém se určí *nejmenší* hladina, při které bychom ještě hypotézu zamítli. Tato **dosažená hladina testu** se v anglicky psaných výstupech označuje jako **P–value, Sig. level, P** a podobně. Vyjadřuje pravděpodobnost spočítanou za platnosti nulové hypotézy, že dostaneme právě náš vektor $\boldsymbol{X}$ nebo vektor ještě více odporující testované hypotéze.

Pro stanovení kritického oboru můžeme ve většině situací použít věrohodnostní funkci. Představme si situaci, kdy nulová i alternativní hypotéza určují rozdělení jednoznačně. Uvažujme rozdělení dané hustotou závislou na parametru $\boldsymbol{\theta}$. Za nulové hypotézy je rozdělení náhodného vektoru $\boldsymbol{X}$ dáno hustotou $f(\mathbf{x};\boldsymbol{\theta}_0)$, za alternativní hypotézy hustotou $f(\mathbf{x};\boldsymbol{\theta}_1)$. Dosaďme do těchto hustot skutečně realizované hodnoty náhodných veličin $X_1 = x_1, \ldots, X_n = x_n$. Pokud bude $f(\mathbf{x};\boldsymbol{\theta}_0)$ výrazně větší než $f(\mathbf{x};\boldsymbol{\theta}_1)$, svědčí to ve prospěch platnosti H_0, opačná situace svědčí spíše pro zamítnutí nulové hypotézy. Následující tvrzení ukáže, že naznačený postup vede k optimálnímu rozhodování.

Věta 12.2. (Neymanovo-Pearsonovo lemma) Nechť k danému $\alpha \in (0,1)$ existuje takové $c > 0$, že pro množinu

$$W_c = \{\mathbf{x} : f(\mathbf{x};\boldsymbol{\theta}_1) \geq c\, f(\mathbf{x};\boldsymbol{\theta}_0)\} \tag{12.9}$$

platí

$$\int_{W_c} f(\mathbf{x};\boldsymbol{\theta}_0)\mathrm{d}\mathbf{x} = \alpha. \tag{12.10}$$

Potom pro každou měřitelnou množinu W takovou, že je

$$\int_{W} f(\mathbf{x};\boldsymbol{\theta}_0)\mathrm{d}\mathbf{x} = \alpha, \tag{12.11}$$

platí

$$\int_{W_c} f(x;\boldsymbol{\theta}_1)\mathrm{d}\mathbf{x} \geq \int_{W} f(\mathbf{x};\boldsymbol{\theta}_1)\mathrm{d}\mathbf{x}. \tag{12.12}$$

D ů k a z: Množiny W, W_c lze psát jako disjunktní sjednocení $W = (W - W_c) \cup (W \cap W_c)$, $W_c = (W_c - W) \cup (W \cap W_c)$. Proto rozdíl

$$\Delta = \int_{W_c} f(\mathbf{x};\boldsymbol{\theta}_1) - \int_{W} f(\mathbf{x};\boldsymbol{\theta}_1),$$

lze psát ve tvaru

$$\Delta = \int_{W_c - W} f(\mathbf{x};\boldsymbol{\theta}_1) + \int_{W \cap W_c} f(\mathbf{x};\boldsymbol{\theta}_1) - \int_{W \cap W_c} f(\mathbf{x};\boldsymbol{\theta}_1) - \int_{W - W_c} f(\mathbf{x};\boldsymbol{\theta}_1).$$

Dva prostřední integrály se zruší. Integrační obor prvního integrálu je částí množiny W_c, takže vzhledem k definici této množiny můžeme tento integrál odhadnout zdola. Podobně integrační obor posledního integrálu *není* částí množiny W_c, takže odečítaný integrál můžeme ze stejného důvodu odhadnout *shora*. Je tedy

$$\Delta \geq c \int_{W_c - W} f(\mathbf{x};\boldsymbol{\theta}_0) - c \int_{W - W_c} f(\mathbf{x};\boldsymbol{\theta}_0).$$

Když sem znovu přidáme integrály přes $W_c \cap W$, pravá strana se nezmění, takže dohromady dostaneme (s použitím předpokladů (12.10), (12.11))

$$\begin{aligned} \Delta &\geq c \int_{W_c} f(\mathbf{x};\boldsymbol{\theta}_0) - c \int_{W} f(\mathbf{x};\boldsymbol{\theta}_0) \\ &= c\alpha - c\alpha = 0. \end{aligned}$$ □

Předpoklady (12.10) a (12.11) požadují, aby kritické obory W_c a W měly za platnosti nulové hypotézy stejnou pravděpodobnost α, tedy stejnou hladinu významnosti. Tvrzení (12.12) porovnává pro dva uvažované kritické

obory W a W_c pravděpodobnost, s jakou zamítneme nulovou hypotézu, když platí hypotéza alternativní (tzv. **sílu testu**). Pro kritický obor W_c je tato pravděpodobnost přinejmenším stejná nebo větší, než pro obecný kritický obor W. To znamená, že kritický obor W_c dá silnější test. A protože jako W vystupuje jakýkoliv kritický obor s hladinou významnosti α, je W_c mezi kritickými obory s danou hladinou α nejsilnější možný.

12.4 Test hypotézy o střední hodnotě v normálním rozdělení se známým rozptylem

Uvažujme náhodný výběr rozsahu n z rozdělení $\mathsf{N}(\mu, \sigma^2)$. Rozptyl σ^2 považujeme za známý. Testujme nulovou hypotézu $\mathsf{H}_0 : \mu = \mu_0$ proti alternativní hypotéze $\mathsf{H}_A : \mu = \mu_1$, kde je $\mu_0 < \mu_1$. Protože je

$$f(\mathbf{x}; \mu, \sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right),$$

má kritický obor z Neymanova-Pearsonova lemmatu svůj tvar určen nerovností

$$\frac{f(\mathbf{x}; \mu_1, \sigma^2)}{f(\mathbf{x}; \mu_0, \sigma^2)} = \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i - \mu_1)^2 - \sum_{i=1}^{n}(x_i - \mu_0)^2\right)\right) \geq c,$$

tedy po úpravě (nejprve logaritmování)

$$2\bar{x}\,(\mu_1 - \mu_0) - \left(\mu_1^2 - \mu_0^2\right) \geq \frac{2\sigma^2}{n} \ln c.$$

Předpoklad $\mu_1 > \mu_0$ vede k nerovnosti

$$\bar{x} \geq \frac{\mu_1 + \mu_0}{2} + \frac{\sigma^2}{n(\mu_1 - \mu_0)} \ln c = x^*.$$

Vidíme, že *tvar* kritického oboru W_c nezávisí na hodnotě neznámého parametru σ^2.

Velikost tohoto oboru určíme z požadavku na hladinu testu:

$$\alpha = \mathsf{P}_0\left[\bar{X} \geq x^*\right] = \mathsf{P}_0\left[\frac{\bar{X} - \mu_0}{\sigma}\sqrt{n} \geq \frac{x^* - \mu_0}{\sigma}\sqrt{n}\right].$$

Index 0 u symbolu pravděpodobnosti znamená, že pravděpodobnost počítáme za platnosti nulové hypotézy H_0. Protože v tomto případě má statistika

$$\frac{\bar{X} - \mu_0}{\sigma}\sqrt{n}$$

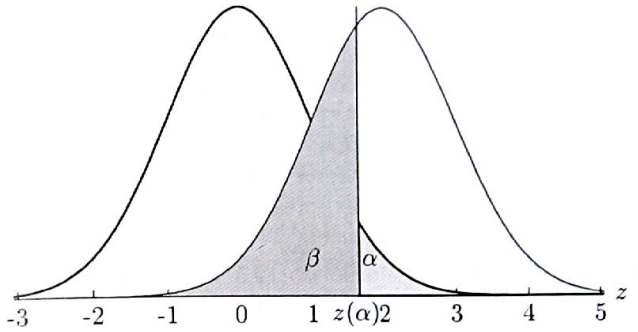
rozdělení $\mathsf{N}(0, 1)$, je třeba zvolit

$$\frac{x^* - \mu_0}{\sigma}\sqrt{n} = z(\alpha).$$

Nejlepší kritický obor je dán požadavkem zamítat nulovou hypotézu H_0 : $\mu = \mu_0$ ve prospěch alternativní hypotézy $\mathsf{H}_A : \mu = \mu_1 (> \mu_0)$, když nastane

$$Z = \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n} \geq z(\alpha). \tag{12.13}$$

Tento kritický obor má velmi důležitou vlastnost – nezávisí na hodnotě μ_1. (Pozor, musí platit $\mu_1 > \mu_0$).



Obrázek 12.1: Znázornění pravděpodobnosti chyby prvního (α) a druhého (β) druhu. Tence je uvedena hustota Z za platnosti alternativní hypotézy: $\mu = \mu_0 + 2\sigma$, tučně za platnosti nulové hypotézy

Příklad 12.5. Vraťme se k příkladu 12.1. Je známo, že každá následující generace je v průměru o něco vyšší než generace předcházející. Můžeme se tedy ptát, zda průměr $\bar{x} = 139{,}133$ cm zjištěný v náhodném výběru rozsahu $n = 15$ znamená, že na 5% hladině máme zamítnout nulovou hypotézu $\mathsf{H}_0 : \mu = 136{,}1$ cm (zjištění z roku 1951) ve prospěch alternativní hypotézy $\mathsf{H}_A : \mu > 136{,}1$ cm. Rozptyl $\sigma^2 = 6{,}4^2$ cm^2, zjištěný v roce 1951 můžeme považovat za známý, neboť variabilita výšek zůstává (na rozdíl od střední výšky) téměř nezměněná. Po dosazení do (12.13) dostaneme

$$z = \frac{139{,}133 - 136{,}1}{6{,}4}\sqrt{15} \doteq 1{,}835,$$

což je hodnota překračující kritickou hodnotu $z(0{,}05) = 1{,}645$. Nulovou hypotézu tedy na 5% hladině zamítneme ve prospěch alternativní hypotézy, že se střední výška desetiletých hochů zvětšila. Dosažená hladina odpovídající testové statistice je rovna 3,33 %, takže například při $\alpha = 2{,}5$ % by dosažený výsledek již nebyl statisticky významný. ○

Pozorný čtenář zjistil, že jsme došli k jinému závěru než v příkladu 12.1. Proč? Ano, kritický obor (12.13) se vztahuje k jednostranné alternativní hypotéze $\mathsf{H}_A : \mu > \mu_0$, kdežto konfidenční interval (12.4) je symetrický (oboustranný). Každý z uvažovaných kritických oborů je založen na *jiných předpokladech*. Skutečnost, že v příkladu 12.5 nepřipouštíme možnost poklesu populačního průměru, umožňuje věnovat celou pravděpodobnost chyby prvního druhu pouze situacím, kdy je $\bar{X} > \mu_0$.

Vyšetřeme nyní případ oboustranné alternativní hypotézy.

Mějme náhodný výběr $X_1, \ldots, X_n$ z rozdělení $\mathsf{N}(\mu, \sigma^2)$ se známým rozptylem σ^2 a na hladině významnosti α testujme nulovou hypotézu $\mathsf{H}_0 : \mu = \mu_0$, kde μ_0 je daná konstanta proti alternativní hypotéze $\mathsf{H}_A : \mu \neq \mu_0$. Spíše ve prospěch alternativní než nulové hypotézy svědčí situace, kdy bude odhad $\bar{X}$ parametru μ od hypotetické hodnoty parametru μ_0 příliš vzdálen, ať už bude větší nebo menší. Proto je vhodné rozhodovat pomocí statistiky $|Z|$. Protože náhodná veličina Z má za hypotézy *symetrické* normální rozdělení, použijeme kritický obor daný nerovností $|Z| \geq z(\alpha/2)$.

Příklad 12.6. Kdybychom v příkladu 12.5 nemohli použít apriorní znalost, že neplatnost hypotézy nutně musí znamenat, že je $\mu > \mu_0$, takže bychom uvažovali oboustrannou alternativní hypotézu $\mathsf{H}_A : \mu \neq 136{,}1$ cm, nezamítali bychom na 5% hladině nulovou hypotézu, protože je $|z| = 1{,}835 < 1{,}960 = u(0{,}025)$. Při oboustranné alternativě bude dosažená hladina dvojnásobná v porovnání s jednostrannou alternativou (v příkladu 12.5), bude tedy rovna hodnotě 6,66 %. ○

13. Lineární model

Následující kapitolka slouží k přípravě obecného schematu, které je základem řady statistických postupů, z nichž některé budou uvedeny v kapitole příští.

13.1 Průmět do podprostoru

Prvky matice $\mathbf{X}$ budeme značit x_{ij}, její sloupce $\mathbf{x}_i$. V této kapitole budeme všude předpokládat, že $\mathbf{X}$ je matice konstant o n řádcích a k *lineárně nezávislých* sloupcích. Jsou tedy $\mathbf{x}_1, \ldots, \mathbf{x}_k$ lineárně nezávislé vektory. **Lineární obal** těchto vektorů budeme značit $\mathcal{M}(\mathbf{X})$, je tedy

$$\mathcal{M}(\mathbf{X}) = \left\{ \mathbf{x} \in \mathbb{R}^n : \mathbf{x} = \mathbf{X}\mathbf{t}, \mathbf{t} \in \mathbb{R}^k \right\}. \tag{13.1}$$

Zřejmě musí platit $k \leq n$, všude dál budeme předpokládat $k < n$. V lineárním podprostoru $\mathcal{M}(\mathbf{X})$ existuje ortonormální báze. Nechť právě sloupce matice $\mathbf{P}$ tvoří ortonormální bázi $\mathcal{M}(\mathbf{X})$. Uvedenou bázi lze doplnit na ortonormální bázi celého prostoru $\mathbb{R}^n$ pomocí sloupců nějaké matice $\mathbf{R}$. Potom je matice $\mathbf{Q} = (\mathbf{P}, \mathbf{R})$ ortonormální, což speciálně znamená, že platí

$$\mathbf{P}^T\mathbf{P} = \mathbf{I}_k, \tag{13.2}$$

$$\mathbf{R}^T\mathbf{R} = \mathbf{I}_{n-k}, \tag{13.3}$$

$$\mathbf{P}^T\mathbf{R} = \mathbf{O}, \tag{13.4}$$

$$\mathbf{P}\mathbf{P}^T + \mathbf{R}\mathbf{R}^T = \mathbf{I}_n. \tag{13.5}$$

Vzhledem k tomu, že je $\mathcal{M}(\mathbf{P}) = \mathcal{M}(\mathbf{X})$, platí nutně

$$\mathbf{R}^T\mathbf{X} = \mathbf{O}. \tag{13.6}$$

Věta 13.1. Nechť $\mathbf{y} \in \mathbb{R}^n$ je pevně zvolený vektor. Označme

$$\widehat{\mathbf{y}} = \mathbf{P}\mathbf{P}^T\mathbf{y}. \tag{13.7}$$

Potom platí

$$\mathbf{z} \in \mathcal{M}(\mathbf{X}) \Rightarrow \|\mathbf{y} - \mathbf{z}\|^2 \geq \|\mathbf{y} - \widehat{\mathbf{y}}\|^2$$

s rovností právě když je $\mathbf{z} = \widehat{\mathbf{y}}$.

D ů k a z: Je-li $\mathbf{z} \in \mathcal{M}(\mathbf{X}) = \mathcal{M}(\mathbf{P})$, pak existuje $\mathbf{t} \in \mathbb{R}^k$ takové, že je $\mathbf{z} = \mathbf{P}\mathbf{t}$. Pro každé $\mathbf{y}$ platí

$$\begin{aligned} \|\mathbf{y} - \mathbf{z}\|^2 &= \|(\mathbf{y} - \mathbf{P}\mathbf{P}^T\mathbf{y}) + (\mathbf{P}\mathbf{P}^T\mathbf{y} - \mathbf{P}\mathbf{t})\|^2 \\ &= \|\mathbf{R}\mathbf{R}^T\mathbf{y}\|^2 + \|\widehat{\mathbf{y}} - \mathbf{P}\mathbf{t})\|^2 \\ &\geq \|\mathbf{R}\mathbf{R}^T\mathbf{y}\|^2 \end{aligned}$$

s rovností právě když platí

$$\widehat{\mathbf{y}} = \mathbf{P}\mathbf{t} = \mathbf{z}. \qquad \square$$

Věta udává návod, jak nalézt v podprostoru $\mathcal{M}(\mathbf{X})$ k danému vektoru $\mathbf{y}$ vektor nejbližší a navíc tvrdí, že tento nejbližší vektor je dán jednoznačně. Matice $\mathbf{P}\mathbf{P}^T$ se někdy nazývá **projekční matice**, vektor $\widehat{\mathbf{y}}$ se pak nazývá ortogonální **průmět vektoru** $\mathbf{y}$ do podprostoru $\mathcal{M}(X)$. Protože vektor $\widehat{\mathbf{y}}$ je určen jednoznačně, stejnou vlastnost má nutně i vektor $\mathbf{y} - \widehat{\mathbf{y}}$.

Věta 13.2. Matice $\mathbf{H} = \mathbf{P}\mathbf{P}^T$ a $\mathbf{M} = \mathbf{R}\mathbf{R}^T$ jsou určeny jednoznačně.

D ů k a z: Nechť $\mathbf{P}_1, \mathbf{P}_2$ jsou dvě ortonormální báze prostoru $\mathcal{M}(\mathbf{X})$. Je-li $\mathbf{Q}_1 = (\mathbf{P}_1, \mathbf{R})$ ortonormální matice, má vzhledem k $\mathcal{M}(\mathbf{P}_1) = \mathcal{M}(\mathbf{P}_2)$ stejnou vlastnost i matice $\mathbf{Q}_2 = (\mathbf{P}_2, \mathbf{R})$. Pak je ovšem

$$\begin{aligned} \mathbf{P}_1\mathbf{P}_1^T &= \mathbf{I} - \mathbf{R}\mathbf{R}^T, \\ \mathbf{P}_2\mathbf{P}_2^T &= \mathbf{I} - \mathbf{R}\mathbf{R}^T, \end{aligned}$$

takže vzhledem ke shodě pravých stran musí být shodné levé strany. $\square$

Věta 13.3. Jediný rozklad vektoru $\mathbf{y}$ tvaru

$$\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2, \qquad \mathbf{y}_1 \in \mathcal{M}(\mathbf{X}), \quad \mathbf{y}_2 \perp \mathcal{M}(\mathbf{X}), \tag{13.8}$$

je dán vztahy

$$\mathbf{y}_1 = \widehat{\mathbf{y}}, \quad \mathbf{y}_2 = \mathbf{y} - \widehat{\mathbf{y}}. \tag{13.9}$$

D ů k a z: Vektory uvedené v (13.8) požadované vztahy splňují, řešení požadavků (13.8) tedy existuje. Nechť $\mathbf{y}_1{}^*, \mathbf{y}_2{}^*$ jsou další takové vektory. Potom je

$$\mathbf{y}_1 - \mathbf{y}_1{}^* \in \mathcal{M}(\mathbf{X}), \tag{13.10}$$

$$\mathbf{y}_2 - \mathbf{y}_2{}^* \perp \mathcal{M}(\mathbf{X}) \tag{13.11}$$

a současně

$$\mathbf{y}_1 - \mathbf{y}_1{}^* = \mathbf{y}_2 - \mathbf{y}_2{}^*.$$

Jediný vektor, který vyhovuje požadavkům (13.10), (13.11), je nulový vektor. $\square$

Z uvedeného je zřejmé, že vektor $\mathbf{y} - \widehat{\mathbf{y}}$ je průmětem na podprostor $\mathcal{M}(\mathbf{R})$, který je ortogonální k $\mathcal{M}(\mathbf{X})$.

Věta 13.4. Platí

$$\widehat{\mathbf{y}} = \mathbf{y} \quad \Leftrightarrow \mathbf{y} \in \mathcal{M}(\mathbf{X}), \tag{13.12}$$

$$\widehat{\mathbf{y}} = \mathbf{0} \quad \Leftrightarrow \mathbf{y} \perp \mathcal{M}(\mathbf{X}). \tag{13.13}$$

D ů k a z: Pro $\mathbf{y} \in \mathcal{M}(\mathbf{X})$ je nejbližším prvkem tohoto prostoru samotný prvek $\mathbf{y}$, proto je také svým průmětem. Rovnost $\widehat{\mathbf{y}} = \mathbf{y}$ nemůže platit pro žádný vektor mimo $\mathcal{M}(\mathbf{X})$, protože tuto rovnost žádný takový vektor splnit nemůže. Druhá dokazovaná ekvivalence je důsledkem pohledu na $\mathbf{y} - \widehat{\mathbf{y}}$ jako na průmět na ortogonální doplněk podprostoru $\mathcal{M}(\mathbf{X})$. $\square$

13.2 Metoda nejmenších čtverců

Mějme náhodný vektor $\boldsymbol{Y}$, o kterém předpokládáme, že platí

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \sigma\boldsymbol{Z}, \tag{13.14}$$

kde $\boldsymbol{\beta}$ je vektor neznámých parametrů, $\sigma > 0$ je neznámý kladný parametr (směrodatná odchylka). Matice $\mathbf{X}$ je stejně jako v celé této kapitole maticí známých konstant o n řádcích a k sloupcích hodnosti k, přičemž platí $k < n$. Náhodný vektor $\boldsymbol{Z}$ je složen z nezávislých náhodných veličin, z nichž každá má normální rozdělení s nulovou střední hodnotou a jednotkový rozptyl. Souhrnně můžeme předpoklady na $\boldsymbol{Z}$ zapsat jako

$$\boldsymbol{Z} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}). \tag{13.15}$$

Uvedené předpoklady budeme označovat jako **lineární model**. Někdy se tento model podrobněji označuje jako normální lineární model s úplnou hodností.

Odhad vektoru $\mathbf{X}\boldsymbol{\beta}$ metodou nejmenších čtverců je takový vektor $\widehat{\boldsymbol{Y}} = \mathbf{X}\mathbf{b} \in \mathcal{M}(X)$, který minimalizuje čtverec délky vektoru $\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}$, tedy funkci

$$S(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{k} x_{ij}\beta_j)^2.$$

Podle věty 13.1 víme, že musí být

$$\begin{aligned} \widehat{\boldsymbol{Y}} &= \mathbf{P}\mathbf{P}^T\boldsymbol{Y} \\ &= \mathbf{P}\mathbf{P}^T(\mathbf{X}\boldsymbol{\beta} + \sigma\boldsymbol{Z}) \\ &= \mathbf{X}\boldsymbol{\beta} + \sigma\mathbf{P}\mathbf{P}^T\boldsymbol{Z} \qquad \text{(viz (13.12) pro sloupce } \mathbf{X}) \\ &= \mathbf{X}\boldsymbol{\beta} + \sigma\mathbf{P}\boldsymbol{V}, \end{aligned} \tag{13.16}$$

kde jsme označili

$$\boldsymbol{V} = \mathbf{P}^T \boldsymbol{Z}. \tag{13.17}$$

Podobně zavedeme

$$\boldsymbol{U} = \mathbf{R}^T \boldsymbol{Z}. \tag{13.18}$$

Protože je $(\boldsymbol{V}^T, \boldsymbol{U}^T)^T = \mathbf{Q}^T \boldsymbol{Z}$ a matice $\mathbf{Q}$ je ortonormální, podle věty 8.3 dostaneme, že složky náhodného vektoru $(\boldsymbol{V}^T, \boldsymbol{U}^T)^T$ jsou nezávislé a mají všechny rozdělení $\mathsf{N}(0,1)$. Speciálně tedy platí $\boldsymbol{U} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}), \boldsymbol{V} \sim \mathsf{N}(\mathbf{0}, \mathbf{I})$. Náhodný vektor $\boldsymbol{Z}$ lze s použitím (13.5) zapsat jako součet nezávislých náhodných vektorů $\mathbf{R}\boldsymbol{U}$ a $\mathbf{P}\boldsymbol{V}$. Náhodný vektor $\boldsymbol{Y}$ lze tedy vyjádřit jako součet

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \sigma \mathbf{P}\boldsymbol{V} + \sigma \mathbf{R}\boldsymbol{U}. \tag{13.19}$$

Protože první dva sčítance na pravé straně jsou z podprostoru $\mathcal{M}(\mathbf{X})$, je jejich součet roven právě $\widehat{\boldsymbol{Y}}$. Vzhledem k definici mnohorozměrného normálního rozdělení tedy platí

Věta 13.5. Platí

$$\widehat{\boldsymbol{Y}} \sim \mathsf{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{P}\mathbf{P}^T\right) \tag{13.20}$$

Vlastnosti reziduí $\boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ udává následující věta.

Věta 13.6. Platí

$$\boldsymbol{Y} - \widehat{\boldsymbol{Y}} \sim \mathsf{N}\left(\mathbf{0}, \sigma^2 \mathbf{R}\mathbf{R}^T\right), \tag{13.21}$$

$$\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|^2/\sigma^2 \sim \chi^2(n-k). \tag{13.22}$$

D ů k a z: Důkaz prvního tvrzení je podobně jako u předchozí věty založen na vyjádření (13.19). Druhá část plyne z toho, že je

$$\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|^2/\sigma^2 = \|\mathbf{R}\boldsymbol{U}\|^2 = \|\boldsymbol{U}\|^2 \sim \chi^2(n-k), \tag{13.23}$$

neboť jde o součet čtverců $n-k$ nezávislých náhodných veličin s rozdělením $\mathsf{N}(0,1)$. □

Někdy je vhodné testovat hypotézu, která tvrdí, že k popisu středních hodnot vystačíme s menším počtem parametrů. Náhodný vektor $\boldsymbol{Y}$ splňuje **podmodel**, když vedle (13.14) platí

$$\boldsymbol{Y} = \mathbf{X}^0 \boldsymbol{\beta}^0 + \sigma \boldsymbol{Z}, \tag{13.24}$$

kde požadavky na $\boldsymbol{Z}$ (13.15) zůstávají v platnosti a existuje matice $\mathbf{T}$ taková, že je

$$\mathbf{X}^0 = \mathbf{X}\mathbf{T},$$

(což implikuje $\mathcal{M}(\mathbf{X}^0) \subset \mathcal{M}(\mathbf{X})$), a když současně platí

$$0 < q < k, \tag{13.25}$$

kde q je hodnost matice $\mathbf{X}^0$. V tomto případě můžeme najít ortonormální bázi podprostoru $\mathcal{M}(\mathbf{X}^0)$ tvořenou sloupci matice $\mathbf{P}^0$, kterou lze rozšířit pomocí $k-q$ sloupců matice $\mathbf{P}^1$ na ortonormální bázi podprostoru $\mathcal{M}(\mathbf{X})$ tvořenou sloupci matice $\mathbf{P} = (\mathbf{P}^0, \mathbf{P}^1)$. Vektor $\boldsymbol{V}$ můžeme rozdělit na dva podvektory

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}^0 \\ \boldsymbol{V}^1 \end{pmatrix} = \begin{pmatrix} \mathbf{P}^{0T} \boldsymbol{Z} \\ \mathbf{P}^{1T} \boldsymbol{Z} \end{pmatrix}. \tag{13.26}$$

Platí-li podmodel, pak platí

$$\begin{aligned} \widehat{\boldsymbol{Y}}^0 &= \mathbf{P}^0 \mathbf{P}^{0T} \boldsymbol{Y} \\ &= \mathbf{X}^0 \boldsymbol{\beta}^0 + \sigma \mathbf{P}^0 \boldsymbol{V}^0, \\ \boldsymbol{Y} - \widehat{\boldsymbol{Y}}^0 &= \sigma \mathbf{P}^1 \boldsymbol{V}^1 + \sigma \mathbf{R}\boldsymbol{U}, \qquad (13.27) \\ \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}^0\|^2 &= \sigma^2 \|\boldsymbol{V}^1\|^2 + \sigma^2 \|\boldsymbol{U}\|^2. \qquad (13.28) \end{aligned}$$

Zejména poslední tvrzení bude užitečné při testování hypotézy, že platí podmodel.

Vraťme se nyní k obecnému modelu a zabývejme se odhadem vektoru $\boldsymbol{\beta}$. Hledáme takovou lineární kombinaci $\mathbf{X}\mathbf{b}$ sloupců matice $\mathbf{X}$, která dá odhad $\widehat{\boldsymbol{Y}}$ vektoru $\mathbf{X}\boldsymbol{\beta}$. Protože je nutně $\mathbf{X}^T(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}) = \mathbf{0}$ (ortogonalita sloupců $\mathbf{X}$ a $\mathbf{R}$), dostaneme $\mathbf{b}$ posloupností implikací:

$$\mathbf{X}\mathbf{b} = \widehat{\boldsymbol{Y}} \Rightarrow \mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\widehat{\boldsymbol{Y}} \Rightarrow \mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\boldsymbol{Y},$$

tedy

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}. \tag{13.29}$$

Vzhledem k tomu, že matice $\mathbf{X}$ má lineárně nezávislé sloupce, existuje *regulární* matice $\mathbf{T}$, pro kterou je $\mathbf{X} = \mathbf{P}\mathbf{T}$. Potom lze vektor $\mathbf{b}$ psát jako

$$\begin{aligned} \mathbf{b} &= (\mathbf{T}^T\mathbf{P}^T\mathbf{P}\mathbf{T})^{-1}\mathbf{T}^T\mathbf{P}^T\boldsymbol{Y} \\ &= \mathbf{T}^{-1}\mathbf{T}^{T^{-1}}\mathbf{T}^T\mathbf{P}^T(\mathbf{P}\mathbf{T}\boldsymbol{\beta} + \sigma\boldsymbol{Z}) \\ &= \boldsymbol{\beta} + \sigma\mathbf{T}^{-1}\boldsymbol{V}. \qquad (13.30) \end{aligned}$$

Hodnotu $\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|^2$ nazveme **reziduální součet čtverců** a označíme RSS. Jako odhad rozptylu σ^2 použijeme **reziduální rozptyl** definovaný jako

$$S^2 = \|\boldsymbol{Y} - \mathbf{X}\mathbf{b}\|^2/(n-k). \tag{13.31}$$

Věta 13.7. V lineárním modelu platí

$$\mathbf{b} \sim \mathsf{N}\left(\beta, \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right), \tag{13.32}$$

$$(n-k)S^2/\sigma^2 \sim \chi^2(n-k), \tag{13.33}$$

$$\mathsf{E}S^2 = \sigma^2, \tag{13.34}$$

přičemž náhodné veličiny $\mathbf{b}$ a S^2 jsou nezávislé.

D ů k a z: Varianční matici odhadu $\mathbf{b}$ dostaneme z (13.30), když použijeme vztah

$$(\mathbf{X}^T\mathbf{X})^{-1} = \left((\mathbf{PT})^T(\mathbf{PT})\right)^{-1} = \mathbf{T}^{-1}\mathbf{T}^{T^{-1}}.$$

Tvrzení (13.33) je jen jiným vyjádřením (13.22). Poslední tvrzení plyne z toho, že střední hodnota χ^2 rozdělení je rovna počtu stupňů volnosti. Nezávislost plyne ze skutečnosti, že $\mathbf{b}$ je funkcí $\boldsymbol{V}$, kdežto odhad rozptylu je funkcí reziduálního součtu čtverců a tedy vektoru $\boldsymbol{U}$. □

Platí-li podmodel (13.24), pak je

$$(RSS^0 - RSS)/\sigma^2 = \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}^0\|^2/\sigma^2 - \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|^2/\sigma^2 = \|\boldsymbol{V}^1\|^2 \sim \chi^2(k-q).$$

Proto platí následující tvrzení.

Věta 13.8. Platí-li podmodel (13.24) lineárního modelu (13.14), pak je

$$F = \frac{(RSS^0 - RSS)/(k-q)}{RSS/(n-k)} \sim \mathsf{F}(k-q, n-k). \tag{13.35}$$

V dalším výkladu bude užitečné použít označení

$$\mathbf{C} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}. \tag{13.36}$$

Pro jednotlivé složky náhodného vektoru $\mathbf{b}$ pak lze psát

$$b_j \sim \mathsf{N}\left(\beta_j, \sigma^2 c_{jj}\right). \tag{13.37}$$

Použijeme-li místo neznámého parametru σ^2 jeho odhad S^2, dostaneme odtud (viz příklad 8.8)

$$T_j = \frac{b_j - \beta_j}{S\sqrt{c_{jj}}} \sim \mathsf{t}(n-k), \tag{13.38}$$

což lze použít v řadě speciálních případů. Náhodné veličiny $T_1, \ldots, T_k$ samozřejmě nejsou nezávislé.

14. Speciální případy lineárního modelu

14.1 Jeden výběr

Nechť má matice $\mathbf{X} = \mathbf{1}$ pouze jediný sloupec obsahující samé jedničky, takže je $k = 1$, nechť mají nezávislé náhodné veličiny $Y_1, \ldots, Y_n$ normální rozdělení $\mathsf{N}(\beta, \sigma^2)$. Potom po dosazení do (13.29) a (13.31) zřejmě dostaneme

$$b = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2, \tag{14.1}$$

což jsou odhady známé z oddílu 11.3. Statistika (viz (13.38))

$$T = \frac{\bar{Y} - \beta_0}{S}\sqrt{n} \tag{14.2}$$

umožňuje testovat nulovou hypotézu $\mathsf{H}_0 : \beta = \beta_0$ proti alternativní hypotéze $\mathsf{H}_1 : \beta \neq \beta_0$ (tzv. **jednovýběrový t test**). Protože má statistika T za platnosti nulové hypotézy rozdělení $\mathsf{t}(n-1)$, budeme nulovou hypotézu zamítat na hladině α v případě, že je $|T| \geq t_{n-1}(\alpha)$.

Analogií intervalu (12.4) je interval spolehlivosti pro střední hodnotu β se spolehlivostí $100(1-\alpha)$ tvaru

$$\left(\bar{Y} - \frac{S}{\sqrt{n}}t_{n-1}(\alpha), \bar{Y} + \frac{S}{\sqrt{n}}t_{n-1}(\alpha)\right). \tag{14.3}$$

Nulovou hypotézu $\mathsf{H}_0 : \beta = \beta_0$ zamítneme na hladině α, právě když hodnota β_0 *není* v intervalu spolehlivosti (14.3).

Příklad 14.1. Každý z dobrovolníků, kteří pobývali několik dní bez zpráv o vnějším světě v jeskyni, měl za úkol bez technických pomůcek odhadnout časový interval dvou hodin. Pokus byl organizován tak, že se ve svých odhadech nemohli jednotliví dobrovolníci ovlivňovat. Skutečná trvání údajně dvouhodinového intervalu v minutách jsou uvedena v tabulce 14.1. Na 5% hladině testujme hypotézu, že odhad délky časového intervalu je v těchto podmínkách nestranný. Standardní výpočty dají hodnoty $n = 14$, $\bar{x} = 138{,}143$, $s = 22{,}364$, takže vyjde $t = (138{,}143 - 120)\sqrt{14}/22{,}364 = 3{,}035$, což po porovnání s tabulkovou kritickou hodnotou $t_{13}(0{,}05) = 2{,}160$ vede k zamítnutí nulové hypotézy na 5% hladině. Tomu odpovídá také dosažená hladina rovná 0,96 %. ○

135	108	152	139	145	126	95
176	138	149	110	166	152	143

Tabulka 14.1: Skutečné délky časových intervalů

Jiným použitím našeho modelu je tak zvaný **párový t–test**. Předpokládá se, že máme *nezávislé* dvojice náhodných veličin $(W_{11}, W_{21}), \ldots, (W_{1n}, W_{2n})$, jejichž *rozdíly* $Y_i = W_{1i} - W_{2i}, i = 1, \ldots, n$ mají rozdělení $\mathsf{N}(\beta, \sigma^2)$. Testujeme hypotézu, že pro každé i platí $\mathsf{E}W_{1i} = \mathsf{E}W_{2i}$. Pro náhodné veličiny Y_i to znamená požadavek $\mathsf{H}_0 : \beta = 0$. Protože je $\bar{Y} = \bar{W}_1 - \bar{W}_2$, použijeme k rozhodování statistiku

$$T = \frac{\bar{W}_1 - \bar{W}_2}{S}\sqrt{n}, \tag{14.4}$$

kde je S^2 určeno vztahem (14.1).

Doporučujeme, aby si čtenář pečlivě rozmyslel rozdíl mezi formulací *nezávislé dvojice náhodných veličin* a formulací *dvojice nezávislých náhodných veličin*. V případě párových testů je totiž výhodné, když náhodné veličiny W_{1i}, W_{2i} jsou korelované.

Příklad 14.2. Na 9 pokusných polích byly zjišťovány výnosy nové odrůdy pšenice. Označíme-li jako y_i zjištěné rozdíly, snadno zjistíme, že je

sezóna	místo								
	A	B	C	D	E	F	G	H	I
1983	4,23	5,09	4,55	5,31	5,04	5,45	4,76	5,57	4,97
1985	5,83	7,05	6,00	4,86	6,31	5,22	6,44	4,71	6,09
rozdíl	1,60	1,96	1,45	-0,45	1,27	-0,23	1,68	-0,86	1,12

Tabulka 14.2: Výnosy pšenice ve dvou sezónách

$\bar{y} = 0{,}833$, $s = 1{,}053$, $t = 0{,}833\sqrt{9}/1{,}053 = 2{,}387$, což je hodnota o málo větší než 5% kritická hodnota $t_8(0{,}05) = 2{,}306$. Dosažená hladina významnosti je rovna 4,4 %. Na 5% hladině se dosažené výnosy v obou sezónách významně liší. ○

14.2 Dva výběry

Nechť náhodný vektor $\boldsymbol{Y}$ je složen z podvektorů $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})^T$, i=1, 2. Předpokládáme, že všechny složky náhodného vektoru $\boldsymbol{Y}$ jsou nezávislé a mají všechny stejný rozptyl σ^2, dále že náhodné veličiny z jednoho

podvektoru mají stejnou střední hodnotu μ_1 resp. μ_2. Označme $\beta_1 = \mu_1$, $\beta_2 = \mu_2 - \mu_1$, takže vztah (13.14) lze zapsat pomocí vektoru $\boldsymbol{Y}$ a matice $\mathbf{X}$ o dvou sloupcích jako

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \sigma \boldsymbol{Z}.$$

Označme

$$\bar{Y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{1i}, \qquad \bar{Y}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_{2i}.$$

Protože je (podle (13.29))

$$\begin{aligned} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} &= \begin{pmatrix} n_1 + n_2 & n_2 \\ n_2 & n_2 \end{pmatrix}^{-1} \begin{pmatrix} n_1\bar{Y}_1 + n_2\bar{Y}_2 \\ n_2\bar{Y}_2 \end{pmatrix} \\ &= \frac{1}{n_1 n_2}\begin{pmatrix} n_2 & -n_2 \\ -n_2 & n_1 + n_2 \end{pmatrix}\begin{pmatrix} n_1\bar{Y}_1 + n_2\bar{Y}_2 \\ n_2\bar{Y}_2 \end{pmatrix} \\ &= \begin{pmatrix} \bar{Y}_1 \\ \bar{Y}_2 - \bar{Y}_1 \end{pmatrix} \end{aligned}$$

a

$$\mathbf{C} = \begin{pmatrix} 1/n_1 & -1/n_1 \\ -1/n_1 & 1/n_1 + 1/n_2 \end{pmatrix},$$

k testování hypotézy $\mathsf{H}_0 : \mu_1 = \mu_2$ (tj. $\beta_2 = 0$) lze použít testovou statistiku

$$T = \frac{\bar{Y}_2 - \bar{Y}_1}{S}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}, \tag{14.5}$$

kde je tentokrát

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left(\sum_{i=1}^{n_1}(Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2}(Y_{2i} - \bar{Y}_2)^2\right),$$

neboť vyjde zřejmě $\hat{Y}_{1i} = \bar{Y}_1$ a $\hat{Y}_{2i} = \bar{Y}_2$. Nulovou hypotézu $\mathsf{H}_0 : \mu_1 = \mu_2$ na hladině α zamítáme ve prospěch oboustranné alternativy $\mathsf{H}_1 : \mu_1 \neq \mu_2$, jestliže platí

$$|T| \geq t_{n_1+n_2-2}(\alpha). \tag{14.6}$$

Uvedená metoda je v často uváděna jako **dvouvýběrový t test**. Zejména je nutné zdůraznit požadavek *nezávislosti* dvou vyšetřovaných výběrů.

Ke stejnému kritickému oboru bychom dospěli volbou $\beta_1 = \mu_1$, $\beta_2 = \mu_2$ a matice $\mathbf{X}$, jejíž první sloupec obsahuje nejprve n_1 jedniček a pak n_2 nul a druhý sloupec n_1 nul a n_2 jedniček. Všimněte si, že tato i původně zvolená matice $\mathbf{X}$ mají stejný lineární obal $\mathcal{M}(\mathbf{X})$.

Příklad 14.3. Použijme ještě jednou data o výškách 15 náhodně vybraných desetiletých chlapců z příkladu 12.1. Máme k disposici také údaje o výškách dvanácti náhodně vybraných desetiletých dívek (tabulka 14.3). Z příkladu 12.1 známe hodnoty $n_1 = 15$, $\bar{y}_1 = 139{,}133$, vypočteme dále

135	141	143	132	146	146
151	141	140	131	142	141

Tabulka 14.3: Výšky dvanácti desetiletých dívek

$\sum(y_{1i}-\bar{y}_1)^2 = 601{,}733$. Podobně pro dívky dostaneme $n_2 = 12$, $\bar{y}_2 = 140{,}75$, $\sum(y_{2i} - \bar{y}_2)^2 = 372{,}25$. Proto je

$$t = \frac{139{,}133 - 140{,}75}{\sqrt{(601{,}733 + 372{,}25)/(15 + 12 - 2)}}\sqrt{\frac{15 \cdot 12}{15 + 12}} = -0{,}669,$$

což je hodnota v absolutní hodnotě nepochybně menší než $t_{25}(0{,}05) = 2{,}060$, takže na 5% hladině nulovou hypotézu nezamítáme. Pro zajímavost, dosažená hladina testu je v tomto případě 50,96 %. ○

14.3 Několik výběrů

Mějme k disposici $k \geq 2$ *nezávislých* výběrů z normálních rozdělení se stejným rozptylem, tj. nechť

$$\begin{aligned} Y_{11}, \ldots, Y_{1n_1} &\sim \mathsf{N}(\mu_1, \sigma^2), \\ Y_{21}, \ldots, Y_{2n_2} &\sim \mathsf{N}(\mu_2, \sigma^2), \\ &\ldots \\ Y_{k1}, \ldots, Y_{kn_k} &\sim \mathsf{N}(\mu_k, \sigma^2). \end{aligned}$$

Uvedený model se zpravidla označuje jako model **analýzy rozptylu jednoduchého třídění**. Budeme testovat nulovou hypotézu $\mathsf{H}_0 : \mu_1 = \cdots = \mu_k$ (společnou hodnotu označíme μ) proti alternativní hypotéze, že aspoň dva výběry mají různé střední hodnoty.

Když uspořádáme hodnoty $Y_{11}, \ldots, Y_{1n_1}, Y_{21}, \ldots, Y_{2n_2}, \ldots, Y_{k1}, \ldots, Y_{kn_k}$ do vektoru, můžeme úlohu zapsat jako lineární model

$$\begin{pmatrix} Y_{11} \\ \cdot \\ Y_{1n_1} \\ Y_{21} \\ \cdot \\ \cdot \\ Y_{k1} \\ \cdot \\ Y_{kn_k} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & 1 \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \cdot \\ \mu_k \end{pmatrix}. \tag{14.7}$$

Jako odhady středních hodnot μ_i dostaneme průměry (ověřte dosazením do (13.29))

$$\bar{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}.$$

Zřejmě potom je $\hat{Y}_{ij} = \bar{Y}_{i\cdot}$, takže reziduální součet čtverců je roven součtu součtů čtverců odchylek od průměrů v jednotlivých výběrech:

$$RSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2.$$

Platí-li testovaná hypotéza (je-li ve všech výběrech stejná střední hodnota), platí vlastně podmodel, ve kterém mají všechny složky náhodného vektoru $\boldsymbol{Y}$ stejnou střední hodnotu

$$Y_{ij} \sim \mathsf{N}(\mu, \sigma^2), \qquad j = 1, \ldots, n_i, \quad i = 1, \ldots, k.$$

Odhadem této společné střední hodnoty je celkový průměr

$$\bar{Y}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{Y}_{i\cdot}$$

a reziduálním součtem čtverců (za hypotézy) je výraz

$$RSS^0 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2.$$

Protože v modelu popisuje střední hodnoty k nezávislých parametrů a v podmodelu je pouze jediný ($q = 1$), má testová statistika (13.35) tvar

$$F = \frac{(RSS^0 - RSS)/(k-1)}{RSS/(n-k)}. \tag{14.8}$$

Výpočty se zapisují do tabulky, která má zpravidla tvar jako tabulka 14.4. V tabulce lze zřetelně číst rozklad

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_{i.} - \bar{Y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2,$$

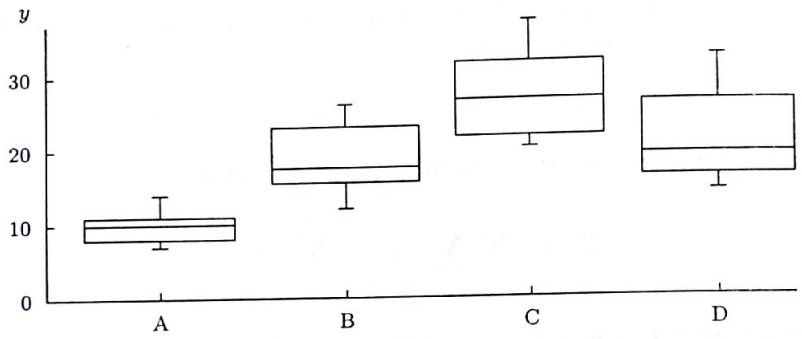
tj.

$$RSS^0 = (RSS^0 - RSS) + RSS.$$

Celkovou variabilitu jsme vyjádřili jako součet variability vysvětlené modelem ($RSS^0 - RSS$) a variability modelem nevysvětlené (reziduální, RSS).

variabilita	součet čtverců	st. vol.	podíl	F	P
ošetření	$RSS^0 - RSS$	$k-1$	$\frac{RSS^0 - RSS}{k-1}$	F	$P[F_{k-1,n-k} \geq F]$
reziduální	RSS	$n-k$	S^2		
celková	RSS^0	$n-1$			

Tabulka 14.4: Tabulka analýzy rozptylu jednoduchého třídění



Obrázek 14.1: Krabicové diagramy hmotností kořenů

Příklad 14.4. Máme čtyři různé živné roztoky, v nichž pěstujeme pšenici. Z každého živného roztoku pořídíme několik vzorků, u nichž zjistíme hmotnost kořenové části rostlin. Zajímá nás, zda je střední hmotnost ve všech živných roztocích stejná. Data jsou uvedena v tabulce 14.5. Z obrázku

roztok	hmotnost								$\bar{Y}_{i.}$
A	8	11	7	11	10	14	10		0,997
B	26	23	16	18	12	15	23	17	1,260
C	31	21	20	32	22	37	29	24	1,422
D	26	19	32	30	14	16	16		1,319

Tabulka 14.5: Hmotnosti kořenové části v různých živných roztocích

14.1 je patrné, že variabilita dat závisí na poloze a že při větším mediánu je směrodatná odchylka větší. Takovou závislost se zpravidla podaří odstranit logaritmováním jednotlivých (kladných) měření, u našich dat je tento postup úspěšný. Proto místo původních hmotností budeme pro jednotlivé živné roztoky porovnávat logaritmy hmotností. Součty čtverců a výsledná testová statistika jsou uvedeny v tabulce analýzy rozptylu tab. 14.6. ○

variabilita	součet čtverců	st. vol.	podíl	F	P
roztoky	0,7176	3	0,2392	18,2532	$< 0{,}001$
reziduální	0,3407	26	0,0131		
celková	1,0583	29			

Tabulka 14.6: Tabulka analýzy rozptylu jednoduchého třídění pro hmotnost kořenů

14.4 Regresní přímka

Klasickým speciálním případem lineárního modelu je **jednoduchá lineární regrese**, kdy předpokládáme, že nezávislé náhodné veličiny $Y_1, \ldots, Y_n$ mají rozdělení $\mathsf{N}(\beta_0 + \beta_1 x_i, \sigma^2)$, kde x_i, $i = 1, \ldots, n$, jsou *dané* konstanty, které nejsou všechny stejné. Rozptyly Y_i jsou stejné, kdežto střední hodnoty lze vyjádřit jako lineární funkci známých konstant x_i pomocí *neznámých* parametrů β_0, β_1. Odhady těchto parametrů dostaneme ze vztahu (speciální případ (13.29))

$$\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^{n} x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^{n} x_i Y_i \end{pmatrix}$$

$$= \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}\begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^{n}x_iY_i \end{pmatrix}$$

$$= \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\begin{pmatrix} n\bar{Y}\sum_{i=1}^{n}x_i^2 - \bar{x}\sum_{i=1}^{n}x_iY_i + n\bar{x}^2\bar{Y} - n\bar{x}^2\bar{Y} \\ \sum_{i=1}^{n}x_iY_i - n\bar{x}\bar{Y} \end{pmatrix}$$

$$= \begin{pmatrix} \bar{Y} - \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\bar{x} \\ \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \end{pmatrix}. \tag{14.9}$$

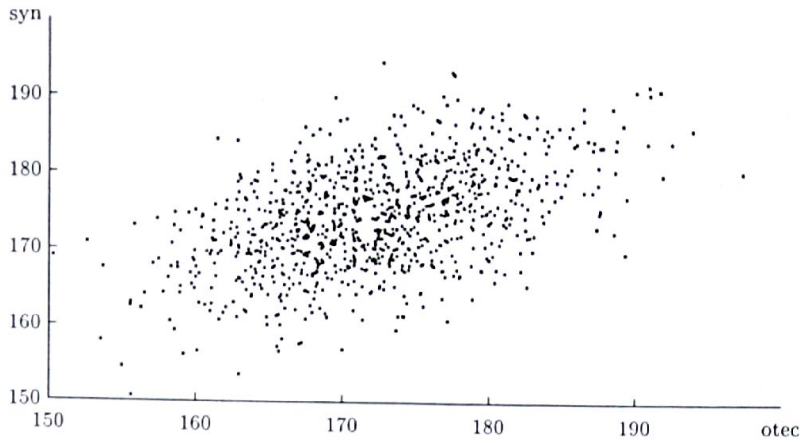
Odhad b_0 je výhodné počítat ze vztahu $b_0 = \bar{Y} - b_1\bar{x}$. Z výpočtu (14.9) je zřejmé, že je $\mathsf{var}b_1 = \sigma^2/\sum_{i=1}^{n}(x_i-\bar{x})^2$. Proto k testování hypotézy, že střední hodnota náhodné veličiny Y nezávisí na x, tj. hypotézy $\mathsf{H}_0 : \beta_1 = 0$, lze použít testovou statistiku (speciální případ (13.38))

$$T = \frac{b_1}{S}\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}. \tag{14.10}$$

Pro výpočet reziduálního součtu čtverců lze využít vztah

$$RSS = \sum_{i=1}^{n}(Y_i-\bar{Y})^2 - b_1\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y}). \tag{14.11}$$

Příklad 14.5. Klasickým příkladem je sledování závislosti výšky syna na výšce jeho otce. Na obrázku 14.2 je znázorněno asi tisíc dvojic takových údajů. Je snad zřejmé, že pro každou hodnotu x (výška otce) kolísají hodnoty Y (výška syna) kolem střední hodnoty, která na x závisí přibližně lineárně. Rozdíly $Y_i - \beta_0 - \beta_1 x_i$ jsou náhodné veličiny, o kterých lze předpokládat normální rozdělení $\mathsf{N}(0,\sigma^2)$, kde σ^2 je neznámým parametrem. Od F. Galtona,



Obrázek 14.2: Závislost výšky syna na výšce jeho otce

který se touto úlohou zabýval, pochází i pojmenování – regrese. Odhad parametru β_1 je roven přibližně jedné polovině, takže například u otce, který je o 10 cm vyšší, než je průměrná výška otců, očekáváme, že výška syna je jen o 5 cm větší, než je průměrná výška synů. Odchylky od průměru se tedy nereprodukují úplně, je tu zřetelná tendence návratu zpět (regrese) k celkovému průměru. ○

14.5 Mnohonásobná lineární regrese

Předpokládejme nyní, že nezávislé náhodné veličiny $Y_1, \ldots, Y_n$ mají vesměs normální rozdělení s rozptylem σ^2 a že střední hodnoty lze vyjádřit pomocí k neznámých parametrů jako

$$\mathsf{E}Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{k-1}x_{i,k-1}.$$

Speciálně pro k=1 bychom dostali regresní přímku z předchozí kapitoly. O matici $\mathbf{X}$ stejně jako dříve předpokládáme, že má lineárně nezávislé sloupce, navíc její první sloupec obsahuje pouze jedničky. Výpočet odhadu $\mathbf{b}$ vektoru parametrů $\boldsymbol{\beta}$ a dalších statistik ponecháme zpravidla na vhodném programovém vybavení. Výsledkem bývá tabulka, která vedle odhadu b_j uvádí odhad jeho směrodatné chyby $S\sqrt{c_{jj}}$ a t statistiku pro test hypotézy, že odhadovaný parametr β_j je nulový (viz (13.38). Pokud je k této statistice uvedena dosažená hladina, usnadňuje to rozhodování.

Při interpretaci nulové hypotézy tvrdící, že regresní koeficient $\beta_j=0$, je třeba vzít vždy v úvahu, že všechny ostatní veličiny (tzv. regresory) ve vyjádření střední hodnoty zůstávají zachovány. Nejde tedy o testování prostého tvrzení, že na j-té veličině střední hodnota náhodné veličiny Y nezávisí.

Dostupné statistické programy udávají ještě například odhad S směrodatné odchylky σ nebo **koeficient determinace** definovaný vztahem

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} = 1 - \frac{RSS}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}. \tag{14.12}$$

Koeficient determinace ukazuje, jakou část variability závisle proměnné se pomocí uvažované závislosti podařilo vysvětlit variabilitou nezávisle proměnných. Proto se často hodnota koeficientu determinace udává v procentech.

Příklad 14.6. Použijeme klasická data z literatury [8]. U 50 mladých mužů, kteří se ucházeli o službu u policie, bylo mimo jiné zjištěno procento tuku, dále jejich hmotnost a výška. Když se ze známé výšky pokusíme odhadnout (předpovědět) procento tuku, dostaneme tabulku 14.7. Máme-li předpovědět rozdíl v procentu tuku u dvou mužů, jejichž výšky se liší o jednotku výšky, pak použijeme odhad regresního koeficientu 0,391. Při zkoumání závislosti procenta tuku na hmotnosti (tabulka 14.8) zjistíme mimo jiné, že

parametr	odhad	směr. chyba	t	P
abs. člen	-55,905	24,959	-2,240	0,030
výška	0,391	0,140	2,794	0,007
S=6,445		R^2=13,990 %		

Tabulka 14.7: Závislost procenta tuku na výšce

parametr	odhad	směr. chyba	t	P
abs. člen	-26,854	4,298	-6,247	<0,001
hmotnost	0,519	0,054	9,553	<0,001
S=4,359		R^2=65,530 %		

Tabulka 14.8: Závislost procenta tuku na hmotnosti

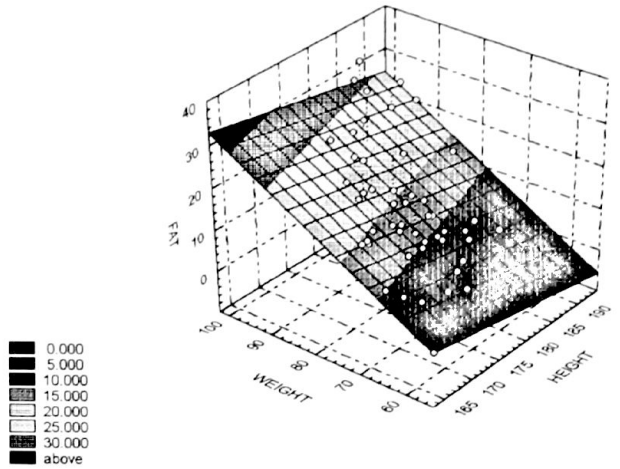
podobně jako u výšky s rostoucí hmotností roste naše předpověď procenta tuku. V obou případech na 5% hladině zamítáme nulovou hypotézu tvrdící, že procento tuku nezávisí na výšce resp. na váze.

Nyní vyšetříme závislost procenta tuku na dvojici veličin výška, hmotnost. V tabulce 14.9 uvedené odhady jsou zajímavé tím, že odhad regresního

parametr	odhad	směr. chyba	t	P
abs. člen	-13,292	16,787	0,792	0,432
výška	-0,273	0,111	-2,466	0,017
hmotnost	0,627	0,068	9,244	<0,001
S=4,145		R^2=69,478 %		

Tabulka 14.9: Závislost procenta tuku na výšce a hmotnosti

koeficientu u výšky změnil v porovnání s tabulkou 14.7 své znaménko. Tentokrát odhadujeme, že vyšší ze dvou mužů, kteří však mají **stejnou hmotnost** a liší se výškou o 1 cm, má tuku **méně** o 0,273 jednotek. Podobně, jako se změnila interpretace odhadu regresního koeficientu, mění se také interpretace nulové hypotézy, tvrdící, že koeficient u výšky je roven nule. Nyní tato hypotéza tvrdí, že výška již nedává o procentu tuku další informaci nad tu, kterou dává hmotnost. ○



Obrázek 14.3: Grafické znázornění závislosti procenta tuku na výšce a hmotnosti

15. Testy dobré shody

15.1 Multinomické rozdělení

Uvažujme situaci podobnou jako při zavedení binomického rozdělení, tedy posloupnost n nezávislých pokusů. Místo dvou možných výsledků pokusu budeme však připouštět celkem m možných výsledků pokusu označených $A_1, \ldots, A_m$, po řadě s pravděpodobnostmi $p_1, \ldots, p_m$. Jako X_j označíme počet dílčích pokusů, v nichž nastal výsledek A_j $(1 \le j \le m)$. Zaveďme (podobně jako v příkladu 6.4) náhodné vektory $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ tak, že

$$\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im})^T,$$

kde $Y_{ij} = 1$ právě když v i-tém pokusu nastal jev A_j, jinak je $Y_{ij} = 0$. Pro každé i, j je tedy

$$\mathsf{P}[Y_{ij} = 1] = p_j, \qquad \mathsf{P}[Y_{ij} = 0] = 1 - p_j,$$

přičemž pro $i \neq k$ jsou náhodné vektory $\boldsymbol{Y}_i, \boldsymbol{Y}_k$ nezávislé. Platí tedy $X_j = \sum_{i=1}^n Y_{ij}$. Proto je

$$\begin{aligned}
\mathsf{E}Y_{ij} &= p_j, & j &= 1, \ldots, m, \quad i = 1, \ldots, n,\\
\mathsf{E}Y_{ij}^2 &= p_j, & j &= 1, \ldots, m, \quad i = 1, \ldots, n,\\
\mathsf{E}Y_{ij}Y_{iq} &= 0, & j, q &= 1, \ldots, m, \quad j \neq q,\\
\mathsf{cov}(Y_{ij}, Y_{iq}) &= p_j(\delta_{jq} - p_q), & j, q &= 1, \ldots, m,
\end{aligned}$$

kde δ_{jq} je známé Kroneckerovo δ, tedy prvek jednotkové matice. S využitím nezávislosti vektorů $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ odtud dostaneme

$$\begin{aligned}
\mathsf{E}X_j &= np_j, \quad j = 1, \ldots, m\\
\mathsf{cov}(X_j, X_q) &= \mathsf{cov}\Big(\sum_{i=1}^n Y_{ij}, \sum_{k=1}^n Y_{kq}\Big)\\
&= \sum_{i=1}^n \sum_{k=1}^n \mathsf{cov}(Y_{ij}, Y_{kq})\\
&= \sum_{i=1}^n \mathsf{cov}(Y_{ij}, Y_{iq})\\
&= np_j(\delta_{jq} - p_q).
\end{aligned}$$

Zavedeme-li vektor $\mathbf{p}$ a diagonální matici $\mathbf{D}$ vztahy

$$\mathbf{p} = (\sqrt{p_1}, \ldots, \sqrt{p_m})^T, \qquad \mathbf{D} = \mathrm{diag}\{\sqrt{np_1}, \ldots, \sqrt{np_m}\},$$

můžeme střední hodnotu a varianční matici náhodného vektoru $\boldsymbol{X}$ zapsat

$$\mathsf{E}\boldsymbol{X} = \mathbf{D}^2\mathbf{1}, \tag{15.1}$$

$$\mathsf{var}\boldsymbol{X} = \mathbf{D}\,(\mathbf{I} - \mathbf{p}\mathbf{p}^T)\,\mathbf{D}. \tag{15.2}$$

Říkáme, že náhodný vektor $\boldsymbol{X} = (X_1, \ldots, X_m)^T$ má **multinomické rozdělení** s parametry $n, p_1, \ldots, p_m$. Parametry musí být nezáporné a splňovat požadavek $\sum_{j=1}^m p_j = 1$. Jednotlivé náhodné veličiny $X_1, \ldots, X_m$ nejsou nezávislé, protože platí $\sum_{i=1}^m X_i = n$. Binomické rozdělení je vlastně speciálním případem multinomického rozdělení pro $m = 2$, přičemž se sleduje pouze jedna ze dvou složek dvourozměrného rozdělení.

Podobně jako u binomického rozdělení (viz (4.7)) zjistíme, že multinomické rozdělení lze popsat pomocí

$$\mathsf{P}(X_1 = x_1, \ldots, X_m = x_m) = \frac{n!}{x_1! \cdots x_m!} p_1^{x_1} \cdots p_m^{x_m}, \tag{15.3}$$

pro libovolnou m-tici $x_1, \ldots, x_m$ splňující $\sum_{j=1}^m x_j, x_j \ge 0, j = 1, \ldots, m$.

Pro nás nejdůležitější bude jistá asymptotická vlastnost multinomického rozdělení, kterou si nyní odvodíme. Uvažujme náhodný vektor

$$\boldsymbol{W} = \mathbf{D}^{-1}(\boldsymbol{X} - \mathbf{D}^2\mathbf{1}),$$

který dostaneme lineární transformací náhodného vektoru $\boldsymbol{X}$ s multinomickým rozdělením. Vzhledem k (15.1) a (15.2) je

$$\mathsf{E}\boldsymbol{W} = \mathbf{0}, \tag{15.4}$$

$$\mathsf{var}\boldsymbol{W} = \mathbf{I} - \mathbf{p}\mathbf{p}^T. \tag{15.5}$$

Doplňme vektor $\mathbf{p}$ s jednotkovou délkou pomocí matice $\mathbf{R}$ na ortonormální matici $\mathbf{Q} = (\mathbf{p}, \mathbf{R})$ a podobně jako v (13.17) a (13.18) zaveďme náhodnou veličinu a náhodný vektor

$$V = \mathbf{p}^T\boldsymbol{W}, \qquad \boldsymbol{U} = \mathbf{R}^T\boldsymbol{W}.$$

Snadno zjistíme, že je

$$V = \mathbf{p}^T\mathbf{D}^{-1}(\boldsymbol{X} - \mathbf{D}^2\mathbf{1}) = \frac{1}{\sqrt{n}}\mathbf{1}^T(\boldsymbol{X} - \mathbf{D}^2\mathbf{1}) = \frac{1}{\sqrt{n}}(n - n) = 0,$$

a dále

$$\mathsf{E}\boldsymbol{U} = \mathbf{0}, \tag{15.6}$$

$$\mathsf{var}\boldsymbol{U} = \mathbf{R}^T(\mathbf{I} - \mathbf{p}\mathbf{p}^T)\mathbf{R}$$

$$= \mathbf{I}_{m-1}. \tag{15.7}$$

Náhodný vektor $\sqrt{n}\boldsymbol{U}$ můžeme zapsat ve tvaru

$$\sqrt{n}\boldsymbol{U} = \sqrt{n}\mathbf{R}^T\mathbf{D}^{-1}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \frac{1}{n}\mathbf{D}^2\mathbf{1}).$$

Každá složka tohoto vektoru je tedy součtem n náhodných veličin, které mají všechny stejné rozdělení s nulovou střední hodnotou a jednotkovým rozptylem (máme n nezávislých sčítanců a $\mathsf{var}\sqrt{n}U_j = n$). Proto podle centrální limitní věty (věta 9.6) má každá složka vektoru

$$\boldsymbol{U} = \frac{1}{\sqrt{n}}\sqrt{n}\boldsymbol{U}$$

asymptoticky rozdělení $\mathsf{N}(0,1)$. Proto asymptoticky má čtverec délky náhodného vektoru $\boldsymbol{U}$ rozdělení $\chi^2(m-1)$. Odtud z definice rozdělení χ^2 plyne následující věta.

Věta 15.1. Nechť náhodný vektor $\boldsymbol{X}$ má multinomické rozdělení s parametry $n, p_1, \ldots, p_m$. Potom pro velká $n = \sum_{j=1}^{m} X_j$ platí

$$\sum_{j=1}^{m}\frac{(X_i - np_j)^2}{np_j} \sim \chi^2(m-1). \tag{15.8}$$

D ů k a z: Důkaz plyne z toho, že $V = 0$ a z ortonormality matice $\mathbf{Q}$. Platí

$$\sum_{j=1}^{m}\frac{(X_j - np_j)^2}{np_j} = \|\boldsymbol{W}\|^2 = \|\boldsymbol{U}\|^2 = \sum_{j=1}^{m-1}U_j^2$$

□

Ve vztahu (15.8) aproximujeme *diskrétní* rozdělení statistiky na levé straně rozdělením spojitým. Tato aproximace se považuje za použitelnou, když je pro každé j splněna nerovnost $np_j > 5$.

15.2 χ^2 test dobré shody

Vztah (15.8) je základem řady testů, v nichž konfrontujeme skutečně zjištěné četnosti $X_1, \ldots, X_m$ jevů $A_1, \ldots, A_m$ se středními hodnotami těchto četností $np_1^0, \ldots, np_m^0$, v nichž jsou pravděpodobnosti $p_1^0, \ldots, p_m^0$ určeny z platnosti nějakého pravděpodobnostního modelu. Nulová hypotéza tvrdí, že pravděpodobnosti jevů $A_1, \ldots, A_m$ jsou po řadě rovny $p_1^0, \ldots, p_m^0$. Testová statistika má tvar

$$X^2 = \sum_{j=1}^{m}\frac{(X_j - np_j^0)^2}{np_j^0} = \sum_{j=1}^{m}\frac{X_j^2}{np_j^0} - n. \tag{15.9}$$

Nulovou hypotézu na hladině významnosti α zamítáme v případě, že je $X^2 \geq \chi^2_{m-1}(\alpha)$. Z tvaru testové statistiky je zřejmé, že neshodě skutečnosti s hypotézou odpovídají právě velké hodnoty této statistiky.

Příklad 15.1. Při 600 hodech hrací kostkou byly zjištěny následující četnosti jednotlivých stran: 85, 99, 91, 108, 119, 98. Lze na 5% hladině považovat tuto hrací kostku za symetrickou? Nulová hypotéza tvrdí, že $p_j^0 = 1/6, i = 1, \ldots, 6$. Odtud je

$$\begin{aligned}\chi^2 &= \frac{(85-100)^2}{100} + \frac{(99-100)^2}{100} + \frac{(91-100)^2}{100} + \\ &\quad + \frac{(108-100)^2}{100} + \frac{(119-100)^2}{100} + \frac{(98-100)^2}{100} \\ &= 7{,}36,\end{aligned}$$

což je hodnota menší než $\chi^2_5(0{,}05) = 11{,}070$, takže hypotézu nemůžeme na 5% hladině zamítnout. Dosažená hladina testu je rovna 19,5 %. ○

Poněkud komplikovanější situace nastává, když hypotéza neurčuje jedinou m-tici pravděpodobností. Zpravidla jde o případ, kdy za platnosti hypotézy jsou tyto pravděpodobnosti známými funkcemi malého počtu parametrů. Označme takový vektor (nezávislých) parametrů jako $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^T$. Abychom mohli použít statistiku X^2 a porovnat zjištěné četnosti s jejich za nulové hypotézy očekávanými hodnotami, musíme nejprve neznámé parametry odhadnout. Metoda maximální věrohodnosti (viz oddíl 12.2) použitá na (15.3) vede na řešení soustavy rovnic

$$\sum_{j=1}^{m}\frac{x_j}{p_j^0(\boldsymbol{\theta})}\frac{\partial p_j^0(\boldsymbol{\theta})}{\partial\theta_j} = 0, \qquad j = 1, \ldots, q. \tag{15.10}$$

Označme řešení této soustavy symbolem $\hat{\boldsymbol{\theta}}$. Dosadíme-li do statistiky X^2 tento odhad, pak asymptoticky (pro dostatečně velká n) platí

$$\sum_{j=1}^{m}\frac{(X_j - np_j^0(\hat{\boldsymbol{\theta}}))^2}{np_j^0(\hat{\boldsymbol{\theta}})} \sim \chi^2(m-q-1). \tag{15.11}$$

Příklad 15.2. Uvažujeme genetický model s fenotypy AA, aA, aa. Pokud se genotypy A, a od obou rodičů kombinují náhodně, pak mají jednotlivé fenotypy po řadě pravděpodobnosti $\theta^2, 2\theta(1-\theta), (1-\theta)^2$. Místo (15.10) dostaneme rovnici

$$\frac{x_1}{\theta^2}2\theta + \frac{x_2}{2\theta(1-\theta)}(2-4\theta) + \frac{x_3}{(1-\theta)^2}(-2(1-\theta)) = 0,$$

jejímž řešením je po úpravě

$$\hat{\theta} = \frac{1}{2} + \frac{X_1 - X_3}{2n}.$$

Jestliže jsme zjistili četnosti po řadě 18, 7, 25, dostaneme odhad

$$\hat{\theta} = \frac{1}{2} + \frac{18 - 25}{100} = 0{,}43,$$

takže testová statistika pro test dobré shody je rovna

$$\chi^2 = \frac{(18 - 9{,}245)^2}{9{,}245} + \frac{(7 - 24{,}510)^2}{24{,}510} + \frac{(25 - 16{,}245)^2}{16{,}245} = 25{,}524,$$

což je podstatně více, než činí kritická hodnota $\chi^2_1(0{,}05) = 3{,}84$. Dosažená hladina je v tomto případě menší než 0,1 %. ○

Místo soustavy rovnic (15.10) by stačilo hledat takovou hodnotu $\hat{\boldsymbol{\theta}}$, pro kterou nabude své minimální hodnoty funkce

$$X^2(\boldsymbol{\theta}) = \sum_{j=1}^{k} \frac{(X_j - np_j^0(\boldsymbol{\theta}))^2}{np_j^0(\boldsymbol{\theta})}. \tag{15.12}$$

Takový výpočet je technicky velmi náročný, ale když pro nějaký „rozumný" odhad parametru $\boldsymbol{\theta}$ dosazený do vztahu (15.12) vyjde hodnota menší než odpovídající kritická hodnota $\chi^2_{m-q-1}(\alpha)$, minimalizující odhad by byl tím spíše nevýznamný. V takovém případě lze i při nedokonalém odhadu parametru $\boldsymbol{\theta}$ rozhodnout, jak také učiníme v následujícím příkladu.

Příklad 15.3. Vraťme se k údajům o počtech zásahů raketami za 2. světové války v Londýně, uvedeným v příkladu 4.3 v tabulce 4.1. Abychom provedli test dobré shody (že data pocházejí z Poissonova rozdělení), použijeme „naivní" odhad parametru λ uvedený v citovaném příkladu: $\hat{\lambda} = 537/576$. V tabulce 4.1 jsou uvedeny očekávané četnosti, spočítané z tohoto odhadu. Protože poslední očekávaná četnost je příliš malá (menší než 5), sloučíme poslední dvě třídy do nové třídy s alespoň 4 zásahy. Zjištěné četnosti 7+1=8 pak odpovídá očekávaná četnost 7,1+1,6=8,7. Výsledná hodnota statistiky

$$\chi^2 = \frac{(229 - 226{,}7)^2}{226{,}7} + \frac{(211 - 211{,}4)^2}{244{,}4} + \frac{(93 - 98{,}5)^2}{98{,}5} + \frac{(35 - 30{,}6)^2}{30{,}6} + \frac{(8 - 8{,}7)^2}{8{,}7} = 1{,}569$$

je nepochybně menší, než kritická hodnota $\chi^2_{5-1-1}(0{,}05) = 7{,}81$, takže není důvod na 5% hladině zamítat nulovou hypotézu tvrdící, že zjištěné četnosti odpovídají Poissonovu rozdělení. ○

15.3 Nezávislost nominálních veličin

Uvažujme náhodné veličiny U, Y, které nabývají hodnot $1, \ldots, r, 1, \ldots, c$ s pravděpodobnostmi $\mathsf{P}[U = i, Y = j]$. Uvažované celočíselné hodnoty jsou zpravidla zástupnými hodnotami za hodnoty v nominálním měřítku. Náhodné veličiny U, Y jsou nezávislé (viz (5.6)), právě když platí

$$\mathsf{P}[U = i, Y = j] = \mathsf{P}[U = i] \cdot \mathsf{P}[Y = j], \quad 1 \le i \le r, 1 \le j \le c. \tag{15.13}$$

Místo celkového počtu $rc - 1$ nezávislých parametrů (pravděpodobností) vystačíme v případě nezávislosti náhodných veličin U, Y s menším počtem nezávislých parametrů. Označíme-li $\beta_i = \mathsf{P}[U = i], i = 1, \ldots, r-1$, $\gamma_j = \mathsf{P}[Y = j], j = 1, \ldots, c-1$, a dále zavedeme $\beta_r = 1 - \sum_{i=1}^{r-1} \beta_i, \gamma_c = 1 - \sum_{j=1}^{c-1} \gamma_j$, je zřejmé, že vystačíme s parametry $\beta_1, \ldots, \beta_{r-1}, \gamma_1, \ldots, \gamma_{c-1}$. Označme symbolem N_{ij} celkovou četnost jevu $[U = i, Y = j]$, pro marginální četnosti zaveďme označení

$$N_{i\cdot} = \sum_{j=1}^{c} N_{ij}, \qquad N_{\cdot j} = \sum_{i=1}^{r} N_{ij}.$$

Neznámé parametry $\beta_1, \ldots, \beta_{r-1}, \gamma_1, \ldots, \gamma_{c-1}$ odhadneme pomocí

$$\hat{\beta}_i = \frac{N_{i\cdot}}{n}, \qquad \hat{\gamma}_j = \frac{N_{\cdot j}}{n}, \quad i = 1, \ldots, r, \quad j = 1, \ldots, c.$$

Dosadíme-li do funkcí $p_{ij} = \beta_i \gamma_j$ uvedené odhady, dostaneme výslednou testovou statistiku

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(N_{ij} - N_{i\cdot} N_{\cdot j}/n)^2}{N_{i\cdot} N_{\cdot j}/n}. \tag{15.14}$$

Za platnosti hypotézy nezávislosti veličin U, Y má statistika X^2 z (15.14) asymptoticky χ^2 rozdělení, jehož počet stupňů volnosti je dán

$$rc - 1 - (r - 1) - (c - 1) = (r - 1)(c - 1).$$

Hypotézu nezávislosti tedy zamítáme, když je $X^2 \ge \chi^2_{(r-1)(c-1)}(\alpha)$.

Podobnou úlohu dostaneme, když máme k disposici nezávislé realizace multinomického rozdělení $(N_{1j}, \ldots, N_{rj})$ z každé c populací a testujeme hypotézu, že ve všech vyšetřovaných populacích mají multinomická rozdělení stejné pravděpodobnosti. I když v tomto případě jsou součty $n_{\cdot j} = \sum_{i=1}^{r} N_{ij}$ pevné (jsou dány předem, nejsou výsledkem náhodného pokusu), za platnosti testované nulové hypotézy má statistika (15.14) opět χ^2 rozdělení

s $(r-1)(c-1)$ stupni volnosti. V tomto případě hovoříme o **testu homogenity**.

Příklad 15.4. Porovnejme podíl plánovaných těhotenství v Praze a mimo Prahu. K tomu účelu jsme vyzpovídali 70 nastávajících maminek v jedné pražské porodnici, z nichž 43 se hlásilo k plánovanému těhotenství. Mezi 29 maminkami v jisté mimopražské porodnici se jich k plánovanému těhotenství hlásilo celkem 15. Liší se podíly plánovaných těhotenství v Praze a mimo Prahu?

zjištěné četnosti

	porodnice		
plán	Praha	okres	celkem
ano	43	15	58
ne	27	14	41
celkem	70	29	99

očekávané četnosti

	porodnice		
plán	Praha	okres	celkem
ano	41,81	16,99	58
ne	28,99	12,01	41
celkem	70	29	99

Tabulka 15.1: Zjištěné a očekávané četnosti při porovnávání podílu plánovaných těhotenství v Praze a mimo Prahu

Podle vzorce (15.14) dostaneme $\chi^2 = 0{,}7959 < 3{,}84 = \chi^2(0{,}05)$, což znamená, že jsme nezamítli nulovou hypotézu tvrdící, že podíl plánovaných těhotenství je v obou populacích stejný. ($P = 0{,}3723$.) ○

Příklad 15.5. V parlamentu se projednává zajímavý zákon (důchody, školné, spotřební daň malých pivovarů a podobně) a nás zajímá, zda spolu souvisí souhlas s projednávaným zákonem a postoj voličů k vládní koalici. Proto u namátkou vybraných voličů byly zjištěny údaje uvedené v tabulce 15.2. Jednoduchým výpočtem dostaneme $\chi^2 = 1{,}924 < 3{,}84 = \chi^2_1(0{,}05)$, což znamená, že na 5% hladině jsme neprokázali závislost dvou sledovaných nominálních znaků. (Všimněte si, že všechny očekávané četnosti jsou dostatečně velké.) ○

zjištěné četnosti

	zákon		
koalice	ano	ne	celkem
ano	9	5	14
ne	4	7	11
celkem	13	12	25

očekávané četnosti

	zákon		
koalice	ano	ne	celkem
ano	7,28	6,72	14
ne	5,72	5,28	11
celkem	13	12	25

Tabulka 15.2: Zjištěné a očekávané četnosti při šetření závislosti volebních preferencí a postoje k zákonu

A Dodatky

A1 Kombinatorika pro klasický pravděpodobnostní prostor

Při konstrukci klasického pravděpodobnostního prostoru jsou užitečné mnohé pojmy kombinatoriky.

Neprázdná množina A má **mohutnost** n (píšeme $|A| = n$), když existuje bijekce (prosté zobrazení) mezi A a množinou přirozených čísel $(n) = \{1, 2, \ldots, n\}$. Mohutnost prázdné množiny je nula. Mohutnost zjišťujeme tak, že neprázdnou množinu A zobrazíme pomocí bijekce na takovou množinu B, jejíž mohutnost n známe nebo umíme stanovit. Činíme pak závěr, že množina A má n prvků, tj. $|A| = n$. Bez újmy na obecnosti můžeme každou množinu mohutnosti n reprezentovat pomocí úseku přirozených čísel $(n) = \{1, 2, \ldots, n\}$. Užitečné jsou zejména následující kombinatorické objekty (n a r jsou přirozená čísla):

Posloupnosti. Posloupnost prvků množiny (n) délky r je zobrazení množiny (r) do množiny (n). Množina všech posloupností délky r prvků množiny (n) je kartézský součin r kopií množiny (n)

$$(n)^r = \{(x_1, x_2, \ldots, x_r) : x_j \in (n), \quad 1 \le j \le r\},$$

tj. r-rozměrná krychle se stranou (n). Počet posloupností prvků (n) délky r je tedy roven číslu n^r, $|(n)^r| = n^r$. Dvě posloupnosti jsou totožné, když mají stejnou délku r a stejné prvky na prvém, druhém,..., r-tém místě.

V pravděpodobnosti pracujeme s posloupnostmi tehdy, když je výsledek pokusu popsán r-člennou uspořádanou skupinou (třeba opakujících se) charakteristik z množiny (n).

Variace a permutace. Pokud je $r \le n$, nalezneme v množině posloupností $(n)^r$ posloupnosti $(x_1, x_2, \ldots, x_n)$, ve kterých se prvky množiny (n) neopakují, tj. $|\{x_1, x_2, \ldots, x_r\}| = r$. Tyto posloupnosti reprezentují právě všechna prostá zobrazení z (r) do (n). Množinu všech takových posloupností značíme $V(n, r)$ a nazýváme je posloupnosti prvků množiny (n) délky r bez opakování nebo také variace r-té třídy z n prvků bez opakování. Zřejmě platí

$$|V(n,r)| = n(n-1)\cdots(n-r+1) = \frac{n!}{r!} = r!\binom{n}{r} \tag{A1}$$

(Sestrojte bijekci mezi $V(n, r)$ a $(n) \times (n-1) \times \cdots \times (n-r+1)$).

Posloupnosti v množině $P_n = V(n, n)$ jsou tedy právě všechny bijekce množiny (n) na sebe, nazývají se **permutace** množiny (n), nebo také permutace n-té třídy. Zřejmě je $|P_n| = n!$.

Variace a permutace používáme tehdy, je-li výsledek pokusu popsán r-člennou uspořádanou skupinou neopakujících se charakteristik z množiny (n).

Podmnožiny, rostoucí posloupnosti, kombinace. Nechť je $r \leq n$. Označme

$\exp_r(n)$ množinu všech podmnožin $A \subset (n)$ mohutnosti r,
$\exp(n)$ množinu všech podmnožin $A \subset (n)$ včetně prázdné množiny,
$C(n,r)$ množinu všech rostoucích posloupností z $V(n,r)$ (posloupnosti prvků množiny (n) délky r),
$(r \times 1, (n-r) \times 0)$ množinu všech posloupností nul a jedniček délky n, ve kterých je právě r jedniček.

Prvky množiny $C(n,r)$ se také nazývají **kombinace** r-té třídy z n prvků bez opakování.

Platí

$$|C(n,r)| = |(r \times 1, (n-1)^{r} \times 0)| = |\exp_r(n)| = \binom{n}{r}. \tag{A2}$$

Prvá rovnost plyne snadno, uvážíme-li bijekci, která přiřazuje rostoucí posloupnosti $(k_1, k_2, \ldots, k_r) \in C(n,r)$ posloupnost nul a jedniček délky n, která má jedničky právě na místech $k_1, k_2, \ldots, k_r$. Například posloupnosti $(1,3,5) \in C(5,3)$ přiřadíme posloupnost $(1,0,1,0,1) \in (3 \times 1, 2 \times 0)$. Druhá rovnost je zřejmá, třetí plyne z rovnosti (A1).

Snadno také ověříme rovnosti

$$|\exp(n)| = |(2)^n| = |\{0,1\}^n| = 2^n.$$

Kombinace z $C(n,r)$ používáme tehdy, je-li výsledek pokusu popsán r-člennou neuspořádanou skupinou neopakujících se charakteristik z množiny (n) nebo rostoucí posloupností délky r prvků této množiny nebo podmnožinou $s \subset (n)$ o r prvcích.

Neklesající posloupnosti, kombinace s opakováním. Nechť $C'(n,r) \subset (n)^r$ je množina všech neklesajících posloupností prvků množiny (n) délky r. Posloupnost v $C'(n,r)$ se také nazývá **kombinace** r-té třídy z n prvků **s opakováním**. Důkaz identity

$$|C'(n,r)| = |(r \times 1, (n-1) \times 0)| = \binom{n-1+r}{r} \tag{A3}$$

je kombinatoricky poněkud obtížnější než naše předchozí úvahy. Druhá rovnost je samozřejmě důsledkem rovnosti (A2). Abychom odvodili rovnost prvou, „zašifrujeme" neklesající posloupnost $k_1 \leq k_2 \leq \ldots \leq k_r$ z $C'(n,r)$ do posloupnosti nul a jedniček z $(r \times 1, (n-1) \times 0)$ takto:

Napíšeme vedle sebe $n+1$ nul, mezi těmito nulami je právě n mezer, do j-té z nich napíšeme tolik jedniček, kolikrát se číslo j v naší neklesající posloupnosti opakuje (pokud se číslo j nevyskytuje vůbec, necháme j-tou mezeru prázdnou). Z takto vytvořené posloupnosti vymažeme první a poslední nulu a dostaneme prvek množiny $(r \times 1, (n-1) \times 0)$.

Pro $r = 7$ a $n = 6$ šifrujeme tedy posloupnost 2, 3, 3, 5, 5, 5, 6 jako 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1.

Toto šifrování samozřejmě definuje bijekci mezi množinami $C'(n,r)$ a $(r \times 1, (n-1) \times 0)$.

Kombinace s opakováním $C'(n,r)$ používáme tehdy, je-li výsledek pokusu popsán r-člennou neuspořádanou skupinou nikoliv nutně různých charakteristik z množiny (n). Ekvivalentně použijeme kombinace s opakováním tehdy, je-li výsledek pokusu popsán neklesající posloupností číselných charakteristik z množiny (n). Dvě kombinace s opakováním jsou totožné, vyskytují-li se v nich stejné prvky množiny (n) se stejnou násobností 0, 1, 2, ..., r.

Tabulka A1: Počty variací a kombinací r-té třídy z n prvků

	bez opakování	s opakováním
variace	$\frac{n!}{(n-r)!}$	n^r
kombinace	$\binom{n}{r}$	$\binom{n+r-1}{r}$

A2 Γ a B funkce

Gamma funkce $\Gamma(a)$ se zavádí vztahem

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx, a > 0,$$

beta funkce $B(a,b)$ vztahem

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}, a > 0, b > 0.$$

Souvislost udává vztah

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Při výpočtu Γ-funkce často vystačíme se znalostí

$$\Gamma(1/2) = \sqrt{\pi},\ \Gamma(1) = 1,$$

což spolu s vlastností

$$\Gamma(a+1) = a\Gamma(a)$$

znamená, že Γ-funkci můžeme chápat jako zobecnění pojmu faktoriál:

$$\Gamma(n) = (n-1)!,\ n \in \mathbb{N}.$$

A3 Maticové značení

V textu skript jsou všechny vektory chápány jako *sloupcové* vektory. Speciálně nulový vektor značíme symbolem $\mathbf{0}$, vektor ze samých jedniček $\mathbf{1}$. Výraz $\mathbf{1}\mathbf{1}^T$ tedy značí čtvercovou matici ze samých jedniček. Podobně nulovou matici značíme symbolem $\mathbf{O}$, jednotkovou matici symbolem $\mathbf{I}$. Pokud je třeba zdůraznit rozměr jednotkové matice, píšeme $\mathbf{I}_k$ pro jednotkovou matici řádu k. Jsou-li $\mathbf{A}$ a $\mathbf{B}$ dvě symetrické matice stejného rozměru, pak nerovnost $\mathbf{A} \geq \mathbf{B}$ znamená, že matice $\mathbf{A} - \mathbf{B}$ je pozitivně semidefinitní, to znamená, že pro všechny vektory $\mathbf{x}$ odpovídajícího rozměru platí $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq \mathbf{x}^T\mathbf{B}\mathbf{x}$.

A4 Poznámky o historii pravděpodobnosti a statistiky

Teorie pravděpodobnosti vznikla jako věda v polovině 17. století zásluhou velmi významných matematiků této doby: **B. Pascal** (1623–1662), **P. de Fermat** (1601–1655) a **Ch. Huygens** (1629–1695). I když některé zajímavé výpočty pravděpodobností byly presentovány (v souvislosti s hazardními hrami) již v 15. a 16. století (Cardano, Pacioli, Tartaglia), prvé systematické pokusy vytvořit obecné metody byly diskutovány a navrženy v knize P. DeFermat, B. Pascal: *De Ratiociniis in Aleae Ludo* (O počítání při náhodných hrách), kterou v roce 1657 vydal Ch. Huygens. Zde se poprvé objevují pravidla počítání s pravděpodobnostmi a pojem střední hodnoty.

Skutečná historie pravděpodobnosti však začíná v roce 1713, kdy vychází traktát *Ars Conjectandi* (Umění předvídat), ve kterém autor **J. Bernoulli** (1654–1705) dokázal, a to správně, zákon velkých čísel. V roce 1730 následuje další významná monografie *Miscellanea Analytica Supplementum* (Rozličné analytické metody), kde **A. DeMoivre** (1667–1754) dokazuje centrální limitní větu pro symetrickou posloupnost Bernoulliových pokusů.

P. S. Laplace (1749–1827) v *Theorie Analytique des Probabilitiès* (1812) velmi podstatně zobecňuje výsledky svých předchůdců. **S. D. Poisson** (1781–1840), **K. F. Gauss** (1777–1855) **P. L. Čebyšev** (1821–1894) a konečně

A. A. Markov (1856–1922) jsou pak ti, kteří se nejvíce zasloužili o rozvoj pravděpodobnosti jako matematické disciplíny, o vytvoření téměř úplné teorie pro součty nezávislých náhodných veličin.

Počátkem 20. století je stále více pociťována potřeba vytvořit řádnou matematickou axiomatiku teorie pravděpodobnosti. První práce v tomto směru napsali **S. N. Bernstein** (1880–1968), **R. von Mises** (1883–1953) a **E. Borel** (1871–1956). Byl to však **A. N. Kolmogorov** (1903–), který v knize *Grundbegriffe der Wahrscheinlichkeitsrechnung* z roku 1933 využil již tehdy dobře známého aparátu obecné teorie míry, aby axiomatizoval pravděpodobnost způsobem, který ji dobře slouží dodnes.

Počátky historie statistiky někteří nacházejí už ve starém Egyptě, kdy byly pravidelně evidovány například dosažené výnosy. Počátky modernější statistiky zpracovávající naměřená data se vztahují k astronomii (K. F. Gauss) nebo například ke genetice. **F. Galton** (1822-1911), kterého ke genetice přivedla kniha jeho strýce Ch. Darwina *O původu druhů*, vytvořil pojem regrese. Za zakladatele průmyslové statistiky bychom mohli považovat **W. Gosseta** (1876–1937), který pracoval pro dodnes známý pivovar Guinness. Při analýze výběrových průměrů spočítaných z malých výběrů objevil t rozdělení. Práci publikoval pod pseudonymem **Student**, což vysvětluje druhý název tohoto rozdělení. Za řadu základních statistických pojmů (např. variační koeficient, směrodatná odchylka, modus) vděčíme **K. Pearsonovi** (1857–1936). Navrhl χ^2 test dobré shody nebo koeficienty mnohonásobné a parciální korelace, používané k popisu závislosti několika náhodných veličin. Patřil k zakladatelům předního statistického časopisu *Biometrika*, v letech 1901–1936 jej také řídil. Mnozí pokládají Pearsona za skutečného zakladatele moderní matematické statistiky.

Sir **R. A. Fisher** (1890–1962) vystudoval vystudoval sice astronomii, ale brzy se stal vedoucím statistikem ve slavné Rothamsted Agricultural Station, kde pro svou práci vytvořil analýzu rozptylu a položil základy plánování pokusů. Od Fishera pochází pojem statistiky jako funkce počítané z výběru nebo metoda maximální věrohodnosti, jeho kniha *Statistical Methods for Research Workers* vychází opakovaně do dnešní doby, i když první vydání se objevilo již v roce 1925.

Teprve po 2. světové válce se vůdčí osobnosti rozvoje statistiky objevují také mimo Velkou Británii. **F. Wilcoxon** (1892–1965) vystudoval fyzikální chemii. V textu uvedené dva testy pocházejí z roku 1945. K rozvoji teorie neparametrických metod, kam tyto dva testy patří, přispěl významně **J. Hájek** (1926–1974), zakladatel české statistické školy.

B Statistické tabulky

B1 Kritické hodnoty rozdělení N(0, 1)

α	0,5	0,10	0,05	0,025	0,01	0,005	0,001
$z(\alpha)$	0	1,282	1,645	1,960	2,326	2,576	3,090

B2 Kritické hodnoty rozdělení $\chi^2(f)$

	α						
f	0,99	0,975	0,90	0,10	0,05	0,025	0,01
1	0,000	0,001	0,004	2,706	3,842	5,024	6,635
2	0,020	0,051	0,103	4,605	5,992	7,378	9,211
3	0,115	0,216	0,352	6,252	7,815	9,349	11,346
4	0,297	0,484	0,711	7,780	9,488	11,144	13,278
5	0,554	0,831	1,145	9,237	11,071	12,834	15,088
6	0,872	1,237	1,635	10,645	12,593	14,451	16,814
7	1,239	1,690	2,167	12,018	14,068	16,015	18,478
8	1,646	2,180	2,733	13,363	15,509	17,537	20,093
9	2,088	2,700	3,325	14,685	16,921	19,025	21,669
10	2,558	3,247	3,940	15,988	18,309	20,486	23,213
11	3,053	3,816	4,575	17,276	19,677	21,923	24,729
12	3,570	4,404	5,226	18,551	21,028	23,340	26,221
13	4,107	5,008	5,892	19,814	22,365	24,739	27,693
14	4,660	5,628	6,570	21,066	23,688	26,123	29,146
15	5,229	6,262	7,261	22,309	24,999	27,492	30,583
16	5,812	6,907	7,961	23,544	26,299	28,850	32,006
17	6,407	7,564	8,671	24,771	27,591	30,196	33,415
18	7,014	8,230	9,390	25,992	28,873	31,531	34,812
19	7,632	8,906	10,116	27,206	30,147	32,858	36,198
20	8,260	9,590	10,850	28,415	31,415	34,175	37,574
25	11,523	13,118	14,610	34,386	37,658	40,654	44,324
30	14,952	16,789	18,491	40,261	43,780	46,988	50,904
35	18,506	20,567	22,463	46,065	49,810	53,214	57,356
40	22,161	24,430	26,506	51,812	55,768	59,354	63,707
45	25,897	28,362	30,608	57,514	61,668	65,425	69,976
50	29,701	32,352	34,760	63,177	67,518	71,437	76,175
60	37,477	40,475	43,182	74,409	79,099	83,319	88,406
70	45,431	48,748	51,731	85,542	90,552	95,049	100,458
80	53,527	57,141	60,381	96,596	101,904	106,659	112,367
90	61,738	65,632	69,113	107,586	113,174	118,171	124,161
100	70,045	74,204	77,914	118,522	124,375	129,602	135,858

B3 Kritické hodnoty rozdělení t(f)

	α					
f	0,10	0,05	0,02	0,01	0,005	0,001
1	6,314	12,706	31,821	63,656	127,324	636,611
2	2,920	4,303	6,965	9,925	14,089	31,602
3	2,353	3,182	4,541	5,841	7,453	12,923
4	2,132	2,776	3,747	4,604	5,598	8,610
5	2,015	2,571	3,365	4,032	4,773	6,869
6	1,943	2,447	3,143	3,707	4,317	5,959
7	1,895	2,365	2,998	3,499	4,029	5,408
8	1,860	2,306	2,896	3,355	3,833	5,041
9	1,833	2,262	2,821	3,250	3,690	4,781
10	1,812	2,228	2,764	3,169	3,581	4,587
11	1,796	2,201	2,718	3,106	3,497	4,437
12	1,782	2,179	2,681	3,055	3,428	4,318
13	1,771	2,160	2,650	3,012	3,372	4,221
14	1,761	2,145	2,624	2,977	3,326	4,140
15	1,753	2,131	2,602	2,947	3,286	4,073
16	1,746	2,120	2,583	2,921	3,252	4,015
17	1,740	2,110	2,567	2,898	3,222	3,965
18	1,734	2,101	2,552	2,878	3,197	3,922
19	1,729	2,093	2,539	2,861	3,174	3,883
20	1,725	2,086	2,528	2,845	3,153	3,850
25	1,708	2,060	2,485	2,787	3,078	3,725
30	1,697	2,042	2,457	2,750	3,030	3,646
35	1,690	2,030	2,438	2,724	2,996	3,591
40	1,684	2,021	2,423	2,704	2,971	3,551
45	1,679	2,014	2,412	2,690	2,952	3,520
50	1,676	2,009	2,403	2,678	2,937	3,496
60	1,671	2,000	2,390	2,660	2,915	3,460
70	1,667	1,994	2,381	2,648	2,899	3,435
80	1,664	1,990	2,374	2,639	2,887	3,416
90	1,662	1,987	2,368	2,632	2,878	3,402
100	1,660	1,984	2,364	2,626	2,871	3,390

B4 Kritické hodnoty rozdělení $\mathsf{F}(m, f)$ pro $\alpha = 0{,}10$

f	m = 1	2	3	4	5	6	8	10
1	39,86	49,50	53,59	55,83	57,24	58,20	59,44	60,20
2	8,53	9,00	9,16	9,24	9,29	9,33	9,37	9,39
3	5,54	5,46	5,39	5,34	5,31	5,28	5,25	5,23
4	4,54	4,32	4,19	4,11	4,05	4,01	3,95	3,92
5	4,06	3,78	3,62	3,52	3,45	3,40	3,34	3,30
6	3,78	3,46	3,29	3,18	3,11	3,05	2,98	2,94
7	3,59	3,26	3,07	2,96	2,88	2,83	2,75	2,70
8	3,46	3,11	2,92	2,81	2,73	2,67	2,59	2,54
9	3,36	3,01	2,81	2,69	2,61	2,55	2,47	2,42
10	3,29	2,92	2,73	2,61	2,52	2,46	2,38	2,32
11	3,23	2,86	2,66	2,54	2,45	2,39	2,30	2,25
12	3,18	2,81	2,61	2,48	2,39	2,33	2,24	2,19
13	3,14	2,76	2,56	2,43	2,35	2,28	2,20	2,14
14	3,10	2,73	2,52	2,39	2,31	2,24	2,15	2,10
15	3,07	2,70	2,49	2,36	2,27	2,21	2,12	2,06
16	3,05	2,67	2,46	2,33	2,24	2,18	2,09	2,03
17	3,03	2,64	2,44	2,31	2,22	2,15	2,06	2,00
18	3,01	2,62	2,42	2,29	2,20	2,13	2,04	1,98
19	2,99	2,61	2,40	2,27	2,18	2,11	2,02	1,96
20	2,97	2,59	2,38	2,25	2,16	2,09	2,00	1,94
25	2,92	2,53	2,32	2,18	2,09	2,02	1,93	1,87
30	2,88	2,49	2,28	2,14	2,05	1,98	1,88	1,82
35	2,85	2,46	2,25	2,11	2,02	1,95	1,85	1,79
40	2,84	2,44	2,23	2,09	2,00	1,93	1,83	1,76
45	2,82	2,42	2,21	2,07	1,98	1,91	1,81	1,74
50	2,81	2,41	2,20	2,06	1,97	1,90	1,80	1,73
60	2,79	2,39	2,18	2,04	1,95	1,87	1,77	1,71
70	2,78	2,38	2,16	2,03	1,93	1,86	1,76	1,69
80	2,77	2,37	2,15	2,02	1,92	1,85	1,75	1,68
90	2,76	2,36	2,15	2,01	1,91	1,84	1,74	1,67
100	2,76	2,36	2,14	2,00	1,91	1,83	1,73	1,66

B5 Kritické hodnoty rozdělení $\mathsf{F}(m, f)$ pro $\alpha = 0{,}05$

f	m = 1	2	3	4	5	6	8	10
1	161,45	199,50	215,71	224,58	230,16	233,99	238,88	241,88
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,85	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,85	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,77	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,35
25	4,24	3,39	2,99	2,76	2,60	2,49	2,34	2,24
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,16
35	4,12	3,27	2,87	2,64	2,49	2,37	2,22	2,11
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,08
45	4,06	3,20	2,81	2,58	2,42	2,31	2,15	2,05
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	2,03
60	4,00	3,15	2,76	2,53	2,37	2,25	2,10	1,99
70	3,98	3,13	2,74	2,50	2,35	2,23	2,07	1,97
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,95
90	3,95	3,10	2,71	2,47	2,32	2,20	2,04	1,94
100	3,94	3,09	2,70	2,46	2,31	2,19	2,03	1,93

B6 Kritické hodnoty rozdělení $\mathsf{F}(m, f)$ pro $\alpha = 0{,}01$

f \ m	1	2	3	4	5	6	8	10
1	4052,1	4999,5	5403,3	5624,6	5763,6	5859,0	5981,2	6055,8
2	98,50	99,00	99,17	99,25	99,30	99,33	99,37	99,40
3	34,12	30,82	29,46	28,71	28,24	27,91	27,49	27,23
4	21,20	18,00	16,69	15,98	15,52	15,21	14,80	14,55
5	16,26	13,27	12,06	11,39	10,97	10,67	10,29	10,05
6	13,75	10,92	9,78	9,15	8,75	8,47	8,10	7,87
7	12,25	9,55	8,45	7,85	7,46	7,19	6,84	6,62
8	11,26	8,65	7,59	7,01	6,63	6,37	6,03	5,81
9	10,56	8,02	6,99	6,42	6,06	5,80	5,47	5,26
10	10,04	7,56	6,55	5,99	5,64	5,39	5,06	4,85
11	9,65	7,21	6,22	5,67	5,32	5,07	4,74	4,54
12	9,33	6,93	5,95	5,41	5,06	4,82	4,50	4,30
13	9,07	6,70	5,74	5,21	4,86	4,62	4,30	4,10
14	8,86	6,51	5,56	5,04	4,69	4,46	4,14	3,94
15	8,68	6,36	5,42	4,89	4,56	4,32	4,00	3,80
16	8,53	6,23	5,29	4,77	4,44	4,20	3,89	3,69
17	8,40	6,11	5,18	4,67	4,34	4,10	3,79	3,59
18	8,29	6,01	5,09	4,58	4,25	4,01	3,71	3,51
19	8,18	5,93	5,01	4,50	4,17	3,94	3,63	3,43
20	8,10	5,85	4,94	4,43	4,10	3,87	3,56	3,37
25	7,77	5,57	4,68	4,18	3,85	3,63	3,32	3,13
30	7,56	5,39	4,51	4,02	3,70	3,47	3,17	2,98
35	7,42	5,27	4,40	3,91	3,59	3,37	3,07	2,88
40	7,31	5,18	4,31	3,83	3,51	3,29	2,99	2,80
45	7,23	5,11	4,25	3,77	3,45	3,23	2,94	2,74
50	7,17	5,06	4,20	3,72	3,41	3,19	2,89	2,70
60	7,08	4,98	4,13	3,65	3,34	3,12	2,82	2,63
70	7,01	4,92	4,07	3,60	3,29	3,07	2,78	2,59
80	6,96	4,88	4,04	3,56	3,26	3,04	2,74	2,55
90	6,93	4,85	4,01	3,53	3,23	3,01	2,72	2,52
100	6,90	4,82	3,98	3,51	3,21	2,99	2,69	2,50

Odkazy

[1] J. Anděl. *Matematická statistika*. SNTL, Praha 1978.

[2] J. Anděl. *Statistické metody*. Matfyzpress, Praha 1993.

[3] H. Cramér. *Mathematical Methods of Statistics*. Stockholm 1946.

[4] V. Čermák, I. Vodrážková. Chování čtvrtého normovaného momentu u standardních rozdělení. *Ekonomicko-matematický obzor* **27** (1991), 86–110.

[5] V. Dupač. *Teorie pravděpodobnosti a matematická statistika*. Skripta, SPN, Praha 1977.

[6] W. Feller. *An Introduction to Probability Theory and its Applications. Vol. 1*. Wiley, New York 1966.

[7] G. M. Fichtengolc. *Kurs differencialnogo isčislenija. Tom II*. Nauka, Moskva 1966.

[8] R. L. Gunst, R. L. Mason. *Regression Analysis and its Applicatons*. Marcel Dekker, New York 1980.

[9] J. Hájek, D. Vorlíčková. *Matematická statistika*. Skripta, SPN, Praha 1977.

[10] V. Jarník. *Diferenciální počet II*. NČSAV, Praha 1956.

[11] M. Josífko. *Pravděpodobnost a matematická statistika pro biology*. Skripta, SPN, Praha 1969.

[12] J. Jurečková. *Úvod do teorie pravděpodobnosti*. Skripta, SPN, Praha 1978.

[13] S. Komenda. *Biometrie*. Skripta, Vydavatelství Univerzity Palackého v Olomouci, Olomouc 1994.

[14] R. Potocký a kol. *Zbierka úloh z pravdepodobnosti a matematickej štatistiky*. ALFA Bratislava, SNTL Praha 1986.

[15] C. R. Rao. *Lineární metody statistické indukce a jejich aplikace*. Academia, Praha 1978 (překlad z angličtiny).

[16] A. Rényi. *Teorie pravděpodobnosti*. NČSAV, Praha 1972 (překlad z němčiny).

[17] A. A. Svěšnikov. *Sbírka úloh z teorie pravděpodobnosti, matematické statistiky a teorie náhodných funkcí*. SNTL, Praha 1971 (překlad z ruštiny).

[18] J. Štěpán. *Teorie pravděpodobnosti. Matematické základy*. Academia, Praha 1987.

[19] J. Štěpán, J. Machek. *Pravděpodobnost a statistika pro učitelské studium*. Skripta, SPN, Praha 1985.

[20] I. Štěpánová, J. Štěpán. Osm úloh o kombinatorické pravděpodobnosti. *Matematika – fyzika – informatika* III (1993–94), 113–119.

[21] K. Zvára. *Regresní analýza*. Academia, Praha 1989.

[22] K. Zvára, Z. Prášková. *Pravděpodobnost a matematická statistika*. Skripta, SPN, Praha 1986.

Rejstřík